B.Sc. in Computer Science and Engineering Thesis

Prediction of Biophysical Properties of Therapeutic Antibodies from Protein Sequences

Submitted by

Md. Rajibul Islam
201505001

Irtesam Mahmud Khan
201505019

Supervised by

Dr. M Sohel Rahman

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

February 2021

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis, titled, "Prediction of Biophysical Properties of Therapeutic Antibodies from Protein Sequences", is the outcome of the investigation and research carried out by us under the supervision of Dr. M Sohel Rahman.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md. Rajibul Islam
201505001

Irtesam Mahmud Khan
201505019

CERTIFICATION

This thesis titled, **"Prediction of Biophysical Properties of Therapeutic Antibodies from Protein Sequences"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in February 2021.

Group Members:

Md. Rajibul Islam

Irtesam Mahmud Khan

Supervisor:

Dr. M Sohel Rahman
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

ACKNOWLEDGEMENT

We are indebted to Professor Dr. M Sohel Rahman for guiding us throughout this work and sharing his invaluable insights with us.

Dhaka
February 2021

Md. Rajibul Islam

Irtesam Mahmud Khan

Contents

CANDIDATES' DECLARATION i

CERTIFICATION ii

ACKNOWLEDGEMENT iii

List of Figures vi

List of Tables vii

ABSTRACT viii

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Our Work . . . 2
- 1.3 Roadmap of the Thesis . . . 3

2 Background 4
- 2.1 Antibody . . . 4
  - 2.1.1 Structure of an Antibody . . . 4
  - 2.1.2 Types of Antibodies . . . 5
  - 2.1.3 Therapeutic and Monoclonal Antibody . . . 6
- 2.2 Bio-physical Properties of Antibody . . . 6
  - 2.2.1 Affinity Capture Self-interaction Nanoparticle Spectroscopy (AC-SINS) 7
  - 2.2.2 Hydrophobic Interaction Chromatography (HIC) Retention Time . . . 7
  - 2.2.3 Poly-Specificity Reagent (PSR) Binding Assay . . . 7
- 2.3 Machine Learning Algorithms . . . 7
  - 2.3.1 Linear Regression . . . 8
  - 2.3.2 Lasso Regression . . . 8
  - 2.3.3 Random Forest Algorithm . . . 9
  - 2.3.4 Support Vector Machine(SVM) . . . 9
- 2.4 Literature Review . . . 9

3 Materials and Methods 11

3.1 Materials 11
3.2 Workflow 13
3.3 Feature Extraction and Selection 13
3.3.1 Feature Extraction 13
3.3.2 Feature Selection and Ranking 14
3.4 Prediction 15
3.5 Evaluation Metrics 16

4 Results **18**
4.1 Experimental Results 18
4.2 Discussion 20

5 Conclusion **22**

References **23**

A Figures **27**

B Codes **29**

List of Figures

2.1 The Four Chain Structure of an antibody [1] 5
2.2 The Five Isotypes of antibodies [1] 5

3.1 A complete Workflow Diagram 12

4.1 Effects of Feature Size on different performance metrics for AC-SINS 19
4.2 Effects of Feature Size on different performance metrics for HIC Retention Time 19
4.3 Effects of Feature Size on different performance metrics for PSR SMP Score . . 20

A.1 Actual vs Predicted results for AC-SINS using SVM 27
A.2 Actual vs Predicted results for HIC Retention Time using SVM 28
A.3 Actual vs Predicted results for PSR SMP Score using SVM 28

List of Tables

3.1 Reduced Feature Space for AC-SINS 15
3.2 Reduced Feature Space for HIC 15
3.3 Reduced Feature Space for PSR 15

4.1 Best Results for AC-SINS 18
4.2 Best Results for HIC Retention Time 19
4.3 Best Results for PSR SMP Score 20

ABSTRACT

Antibodies have become one of the most predominant class of drugs in modern days. Around the world, at least 570 therapeutic monoclonal antibodies have been studied in clinical trials by commercial companies, and 79 therapeutic monoclonal antibodies have been approved by the United States Food and Drug Administration (US FDA). But a lot of the candidate drugs fail at different stages of development. Some biophysical assays have been proposed for screening candidates at the early stage of development. But this wet lab experimnets are costly and time consuming. Instead several computational/theoretical tools have been developed in the recent years. In this study, we have proposed a machine learning based computational method to predict three important biophysical assays(AC-SINS,HIC retention time and PSR) from antibody sequences only. We have used only the sequence order information as features and used several machine learning techniques to reduce the feature space and predict the target biophysical assay. Our model can predict this biophysical assays with surprisingly high degree of accuracy. The low computational expense and a high accuracy makes our method very feasible for reducing cost of monoclonal antibodies development.

Chapter 1

Introduction

Antibody is one of the most important part of human immune system. Antibody(also known as immunoglobulin) is the search and destroy unit of our immune system i.e. it detects any potentially harmful foreign body(Antigen) and neutralize them. Many scientists have been trying to produce antibodies that will attack specific antigens for some time. With the discovery of monoclonal antibodies by Köhler and Milstein around three decades ago, this became a reality [2]. Monoclonal antibodies(mAbs) are a class of human synthesized antibodies that can target a specific antigen. Over the last three decades, monoclonal antibodies have made a major transformation from scientific tools to powerful human therapeutics [3]. The first antibody approved by FDA was *Muromonab* in 1986 [4]. After that there was no looking back. Antibodies have become one of the most predominant class of drugs in modern days. Recently developed antibodies have a very few adverse side effect because of high degree of specificity. To overcome immunogenicity risk, new technologies for the generation of predominately or entirely human origin mAbs were developed [3]. Humira (Adalimumab) is the first fully human antibody approved in 2004, for the treatment of rheumatoid arthritis [5]. Currently, most mAbs developed are humanized or fully human [6]. Around the world, at least 570 therapeutic mAbs have been studied in clinical trials by commercial companies [7], and 79 therapeutic mAbs have been approved by the United States Food and Drug Administration (US FDA) and are currently on the market [8]. The global therapeutic monoclonal antibody market was valued at approximately US$115.2 billion in 2018 and was expected to generate revenue of $150 billion by the end of 2019 and $300 billion by 2025 [9]. Although antibody has become one of the most important drug, there is still some issues related to the production of antibodies. 90% of candidate drugs fail in different stage of clinical development [10]. But the underlying reasons of why these candidates fail remains a mystery to some extent.

1.1 Motivation

Developability is an umbrella term that encompasses various drug-like properties, manufacturability and safety profiles of therapeutic antibodies. Immunogenicity, polyspecificity, high viscosity, instability, self-association or poor expression are some of the many reasons that prevents a candidate antibody to become a fully developed drug [11]. As many candidate drugs fail in several phases of the development, the cost of monoclonal antibody production rises substantially. Identifying candidate drugs that may not work out at the early stage of the development can reduce the production cost significantly. Many methods have been proposed to identify the developability issues at the early stage of clinical trials. Many high throughput developability assays have been proposed for screening antibodies. But this methods need large number of candidate samples to perform well. Generating candidates remain costly, so is performing these biophysical assays.

Many computational/theoretical methods have been developed in recent times to solve the above mentioned problems and reduce the overall cost of the development [12–15]. This methods have used several techniques for predicting different developability risks. For example, *Obrezanova et al.* has developed a tool to predict aggregation risk propensity from antibody sequences. But very few works has been done on the prediction of biophysical assays(e.g. AC-SINS,HIC,PSR etc) from the sequences. After the release of antibody dataset by *Jain et al.*, many recent works have been done on this sector [16]. *Hebditch et al.* has predicted biophysical properties from sequences directly [17], but the result is not very accurate. *Dzisoo et al.* also developed an online tool for predicting a subset of biophysical assays [18]. This work does not predict exact values of the related assays rather try to classify if an antibody is developable in terms of that assay. Another work by *Jain et al.* has predicted HIC from antibody sequences. But in this work, they have leveraged 3-dimensional homology structure of the protein as well [19].

1.2 Our Work

The objective of our study is to predict values of some biophysical assays of monoclonal antibodies directly from antibody sequences. An auxiliary objective is to analyse the impact of the sequence order information to predict these biophysical assays.

In this study, we have predicted three biophysical assays(AC-SINS,HIC and PSR) directly from heavy and light chain sequences of monoclonal antibodies. We have used the dataset provided in [16] that includes information about 137 antibodies in later stage of clinical developments.

1.3 Roadmap of the Thesis

In this chapter, we have provided a basic overview of the problem we are working on. Chapter 2 discusses the required preliminary information that is needed to understand this work. In Chapter 3, we have discussed the data source and methods we have used for analysis and prediction. Chapter 4 discusses the experimental results obtained from this study. Chapter 5 includes some conclusive remarks about this study and future direction of this research,

Chapter 2

Background

2.1 Antibody

Antibody, also called *immunoglobulin*,is a protective protein produced by the immune system in response to the presence of a foreign substance, called an antigen [1]. An antigen can be any foreign body(virus,bacteria,other disease causing organisms and toxic substances) that is harmful and alien to our body. If an antigen enters our body,human immune system can recognize it as an alien substance. In response, immune system produces antibodies to neutralize the threat. Some specialized white blood cells called *B lymphocytes (or B cells)* produce antibodies.Antibodies attack antigens by binding to them.Some antigens(usually toxins) are neutralized by only binding of the antibody and changing the chemical composition of the antigen.In some other scenarios,the antibody binds the antigen and invites other antibodies to make the antigen immobile.In another case,after the antibody gets attached to the antigen,a substance called complement uses chemical reaction to destroy the antigen.

2.1.1 Structure of an Antibody

Although basic structure of all antibodies are similar,they can very a great deal in the region that binds with the antigen.As shown in Figure 2.1,four polypeptide chains(two heavy and two light) make a Y-shaped form to create an antibody. The structure of the tip of the ”Y” varies greatly among different antibodies so that they can bind to different antigens.This region is called *Variable Region*. This variable region is composed of 110-130 amino acid.The sequence of this amino acids can potentially determine the functionalities of that particular antibody.On the contrary,the base of the Y is called *Constant region* as they do not vary significantly among different antibodies.This constant region determines how the antigen is neutralized.

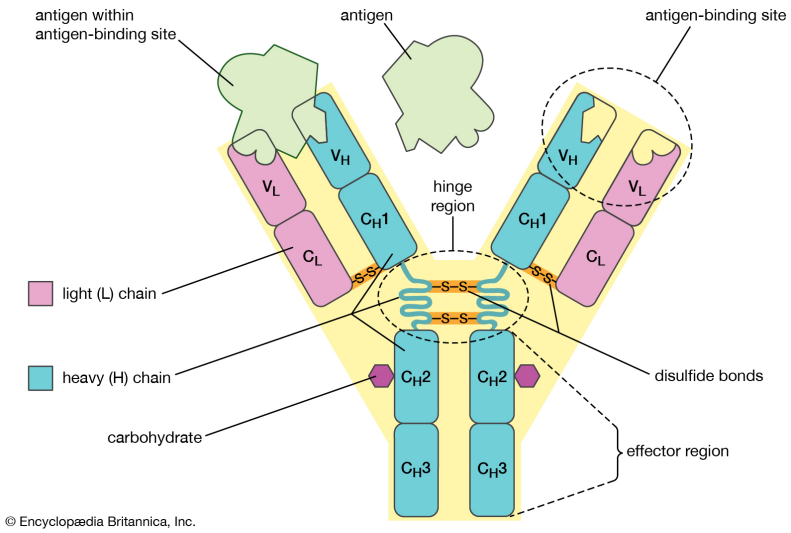

Figure 2.1: The Four Chain Structure of an antibody [1]

2.1.2 Types of Antibodies

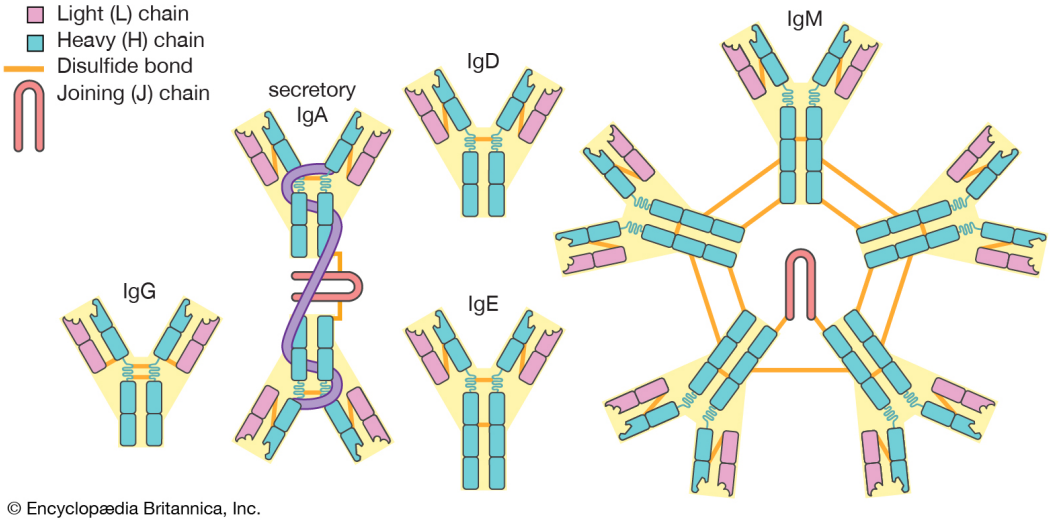

Figure 2.2: The Five Isotypes of antibodies [1]

Human antibodies can be divided into five isotypes,namely IgM,IgD,IgG,IgA and IgE,based on their constant region.Each isotype has distinct shapes,characteristics and functionalities.Each of the isotype play a different role in our immune system.

IgG

IgG is the most abundant antibodies in human serum.IgG comprises of almost 70-75% of all the antibodies in human body.IgG immunoglobulins are usual Y-shaped antibodies.They are divided into four subclasses(IgG1, 2, 3, and 4).Though this subtypes have almost 95% similarity in the constant region,they vary in the hinge regions.

IgM

IgM is the largest antibody and it arrives first after the initial recognition of the antigen [20].Spleen is mainly responsible for the production of IgM antibodies in most mammals including humans.IgM comprises of the 10% of human antibodies.It has a pentametric structure and for this reason it has hogher avidity.

IgA

IgA accounts for 10-15% of human antibodies.This is found in nasal fluids,serum,saliva etc.This antibody forms dimer after secretion.

IgE

IgE is found in a very small amount in human. They account for only 0.001% of the human antibodies.

IgD

IgD comprises of 1% of the human antibodies.The functionality of IgD is a mystery from the time of its discovery.

2.1.3 Therapeutic and Monoclonal Antibody

Therapeutic antibodies are specifically targeted antibodies that binds to the cell surface of the targeted antigen.Therapeutic antibodies are used in fight for cancers and other diseases.Therapeutic antibodies are most often *Monoclonal Antibodies*.Monoclonal Antibodies(*mAbs*) are human synthesized antibodies that are produced by cloning a unique white blood cell.Monoclonal antibodies usually have monovalent affinity, that is it binds to the same region(*Epitope*) of an antigen.On the contrary,*polyclonal antibodies* can bind to multiple epitopes.*Polyclonal antibodies* are also produced from different B-cells.Monoclonal antibodies are widely used as therapeutic antibodies and they are very useful for biopharmaceuticals.

2.2 Bio-physical Properties of Antibody

Monoclonal antibodies are used in treatment of different diseases nowadays.However, the development of mAbs with desirable properties remains quite challenging. Some desirable properties can be (1)express well,(2)elicit a desirable biological effect upon binding and (3)remain soluble and display low viscosity at high concentrations [21].While developing therapeutic antibodies screening antibodies with only the desirable properties is a costly and complicated task.There are some biophysical assays that can help in early stage screening of the antibodies.Some related assays will be discussed in the current chapter.

2.2.1 Affinity Capture Self-interaction Nanoparticle Spectroscopy (AC-SINS)

This assay measures the degree of self-interaction of an antibody.A common challenge of producing mAbs is preventing self-association.So a high throughput screening assay estimating self-association early in the development process can save both money and effort.AC-SINS is such an assay that can work with dilute solutions of mAbs available at the early stage of development and assess self-association and aggregation.This assay uses gold nanoparticles coated with anti-Fc antibodies. When a dilute solution of antibodies is added, they rapidly become immobilised on the gold beads. If these antibodies subsequently attract one another, it leads to shorter interatomic distances and longer absorption wavelengths that can be detected by spectroscopy [21].Consequently this assay can assess the degree of self-association of the monoclonal antibodies.

2.2.2 Hydrophobic Interaction Chromatography (HIC) Retention Time

The hydrophobicity of mAbs is another important biophysical property for their developability into therapeutic antibodies. HIC retention time is used to measure the hydrophobicity of the mAbs.This assay can also measure the heterogeneity of the mAbs.Monoclonal antibodies are mixed with a polar phase and then washed over a hydrophobic column.Subsequently,UV aborption and other techniques can be used to determine the degree of adhesion [22, 23].

2.2.3 Poly-Specificity Reagent (PSR) Binding Assay

Polyspecifity is a significant obstacle in the development of monoclonal antibodies.Poly-Specificity Reagent (PSR) Binding Assay is a high-throughput method for examining the polyspecificity of mAbs.This assay uses fluorescence-activated cell sorting (FACS), a type of flow cytometry.This method tries to determine median fluorescence intensity - higher median intensity meaning greater chance of more specific binding [24].

2.3 Machine Learning Algorithms

Machine learning (ML) is the study of computer algorithms that improve automatically through experience [25].It is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.ML algorithms build a model based on sample data, known as **training data**, in order to make predictions or decisions without

being explicitly programmed to do so [26].There exist different types of ML approaches and they are usually divided into three broad categories named as supervised learning, unsupervised learning and reinforcement learning.In **supervised learning**, algorithms build a mathematical model of a set of labeled data.It also analyzes the training data and learns a function which will be used for new examples.**Unsupervised learning** algorithms take a set of data containing only inputs, and find structure in the data.The structure of the data might be grouping,clustering or something else.**Reinforcement learning** is concerned with how software agents take actions in an environment so as to maximize desired reward.Also ML algorithms can be classified as regression and classification. In this project,we have used some regression models for prediction of biophysical properties of therapeutic antibodies.These algorithms are briefly described as follows.

2.3.1 Linear Regression

Linear regression is a linear approach for modelling the relationship between dependent and independent variables.If there is a single independent variable is called simple linear regression and if more than one dependent variables, then it is called multiple linear regression.Actually, simple linear regression is a special case of multiple linear regression.The basic model for multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n \tag{2.1}$$

where n is the number of features used to learn the hypothesis. Linear regression assumes that the relationship between input and output is linear.It does not support anything else.It is a parametric approach where initial assumption of the form of function is mostly unchangeable.It also assumes that data are noiseless and it creates overfitting problem when highly correlated input variables are used.So removing the most correlated variables leads to better result using this model.

2.3.2 Lasso Regression

Lasso stands for *Least Absolute Shrinkage and Selection Operator*.It is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.In machine learning, Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.It is a regularization technique and is used over regression methods for a more accurate prediction.The lasso procedure encourages simple, sparse models.As it performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude

of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.As Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.

2.3.3 Random Forest Algorithm

Random forest(RF) is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees [27].Ensemble learning is a machine learning algorithm that combines multiple base models in order to produce a powerful model .RF consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For classification, it is typically either Gini impurity or information gain/entropy and for regression trees it is variance.For regression this impurity is termed as node impurity.In this project, we try to reduce the feature space by selecting potentially more important features.Features having a positive node impurity were selected.

2.3.4 Support Vector Machine(SVM)

Support-Vector Machine (SVM) [28] is supervised learning model which is developed at AT&T Bell Laboratories by Vladimir Vapnik. It is one of the most robust prediction method, being based on statistical learning frameworks.The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points where N is the number of features.Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features equals to/exceeds 3, then the hyperplane becomes a two/more dimensional plane.

2.4 Literature Review

Therapeutic monoclonal antibodies have become an emerging part of the pharmaceutical industries.For the successful production of a therapeutic mAb,it is not enough that it only binds to the

specific antigen.Rather it should also not suffer from developability issues - poor solubility,high levels of self-aggregation etc [29, 30]. Screening problematic antibodies in the early stage of development remains one of the important part of mAbs production. Several wet lab experiments have been proposed to identify antibodies with poor developablity. For example,ACSINS [21],HIC [22] and PSR [24] are some high throughput assays to determine the developabilty. But large number of candidates,which is both expensive and time consuming,is needed to perform this experiments. Conducting the assay is also costly and the results need to be interpreted as well. Several techniques have been proposed to reduce the number of samples required and increasing the througput [31, 32]. Several computational tools have also been developed to predict protein aggregation [12–15]. This *in silico* tools have utilized several methods for prediction from semi-empirical methods to sequence based approaches. *Obrezanova et al.* has used statistical modeling and machine learning approaches linking the experimental aggregation data with physico-chemical parameters describing the amino acid sequences of antibodies. The resulting model provides qualitative prediction of aggregation risk for antibodies (High or Low risk of aggregation) using the primary sequence of antibodies as input [13]. *Agrawal et al.* has used homology modeling from sequences to determine 3-dimensional structure and predict spatial charge map that can predict high viscosity [33]. *Lauer et al.* has used a newly defined term Developability index to predict aggregation propensity [12]. *Raybould et al.* has proposed five computational developability guidelines for therapeutic antibody profiling [11].

Another computational approach is predicting biophysical assays quantitatively or qualitatively. Several computational methods have been to predict hydrophobicity of antibodies [34]. *Hanke et al.* has analysed the correlation between HIC retention time and surface properties [35]. A recent work of *Jain et al.* predicts HIC retention time directly from sequences [19]. The recent release of biophysical properties of antibody dataset by *Jain et al.* has a been very important milestone for developing further theoretical tools [16]. This dataset is an excellent resource as they have analysed 137 antibodies in advanced clinical stage of the development. Further computational tools have already been developed based on this dataset. *Hebditch et al.* has predicted biophysical properties from sequences directly [17]. *Dzisoo et al.* also developed an online tool for predicting a subset of biophysical assays [18]. This work does not predict exact values of the related assays rather try to classify if an antibody is developable in terms of that assay.

Chapter 3

Materials and Methods

3.1 Materials

In order to create a robust machine learning model,a reliable training dataset is needed. In our study, we have used the dataset collected by *Jain et al.* [16]. This is an excellent source of information for therapeutic antibodies at later stage of clinical trials. This dataset includes 137 antibodies that have reached at least phase-2 of the clinical trials and had USAN or WHO International Nonproprietary Names (INN) designations. 48 antibodies were built from variable region sequences found in clinically approved antibodies (two of them approved so far only outside the United States), 42 are in the phase-3 or phase-2/3 stage, and the remaining 47 are in phase 2. A total of 124 have kappa light chains, and 13 are lambdas. 58 are classified as "fully human" (with -UMAB suffix) and 67 as "humanized" (with –ZUMAB suffix), and 12 have at least one "fully" nonhuman variable region(–XIMAB, –XIZUMAB, or –MONAB suffix).To compare antibody variable domain properties within a common context the chosen set of 137 antibodies was expressed as human IgG1 isotype (allele *01) with standard constant regions for kappa and lambda (alleles IGKC*01 and IGLC2*01,respectively) as appropriate.Each antibody was then subjected to a battery of 12 different biophysical assays in common use for developability assessment. In summary, this dataset includes heavy and light chain variable region sequences and 12 biophysical assay measurements of 137 antibodies. Among these 12 biophysical assays we have tried to predict 3 of them – **AC-SINS**(affinity-capture self-interaction nanoparticle spectroscopy) [21],**PSR** (poly-specificity reagent) [24] and **Hydrophobic Interaction Chromatography**(HIC) [22].

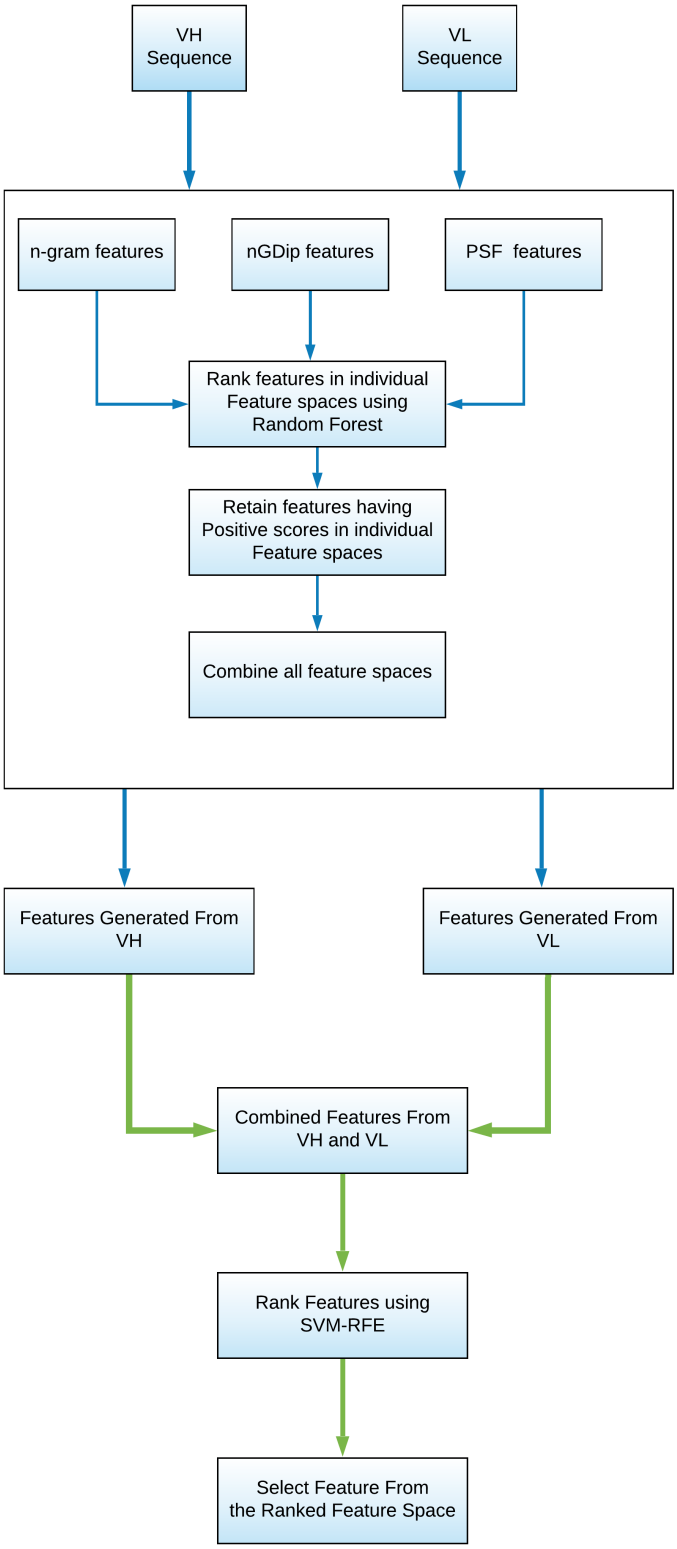

Figure 3.1: A complete Workflow Diagram

3.2 Workflow

A overview of the complete workflow is presented in Figure 3.1. Initially we had variable region sequences of heavy chain(VH) and light chain(VL) of each antibodies. We used three feature extraction techniques, namely nGram,nGDip and PSF(details will be discussed in the subsequent sections), to generate features separately from VH and VL. Next, we used random forest algorithm to rank features in each of these individual feature space. This step is used to reduce the feature space. Then we combined all of the individual feature space from VH and VL into a combined feature space. Subsequently, this combined feature space was ranked with SVM-RFE [36]. Finally, we used several machine learning techniques using several subsets of the ranked feature space to predict the biophysical property. The same procedure is applied to all three biophysical assays.

3.3 Feature Extraction and Selection

3.3.1 Feature Extraction

The input to our algorithms are heavy(VH) and light(VL) weight chain sequence of an antibody. To capture the sequence order information, we have extracted position independent as well as position specific features. Among the position independent features are dipeptides (Dip), tripeptides,tetrapeptides and n-gappeddipeptides (nGDip). These features do not depend on any specific position in the amino acid sequence.We describe each of these feature construction techniques briefly in the following.

Amino Acid Composition

Amino acid composition (AAC) of a protein sequence refers to the normalized frequencies of the 20 native amino acids. The frequencies are normalized by dividing each of these by the protein sequence length.

Dipeptides (Dip)

Dipeptides (Dip) or dipeptide composition (DPC) features are the normalized frequency of adjacent amino acids within the protein sequence. These features provide some sequence-order information.

Tripeptides

Similarly, the normalized frequency of three consecutive amino acids can be used as features. This is called tripeptides composition feature type.

Tetrapeptides

Again, the normalized frequency of four consecutive amino acids can be used as fea-

tures. This is called tetrapeptides composition feature type. AAC, dipeptides and tripeptides,tetrapeptides – all these feature types can be generalized under the umbrella of *n-grams* feature type, where frequencies of n-length peptides are used as feature vectors. In our study we have extracted a total of 7003 n-gram features from VH sequence and 5146 n-gram features from VL sequence,for n=1,2,3,4.

N Gapped Dipeptides (nGDip)

The n-gapped-dipeptides (nGDip) features are extracted by counting the frequency of amino acid dipeptides such that the amino acids are separated by n positions. The frequency is normalized, dividing it by the total number of nGDip (i.e., L-n-1 for a sequence of length L). The nGDip feature extraction technique is motivated by the belief that the gap between any two amino acids may carry significant information about the protein [37]. We have considered up to 25 position gaps. In total,we have generated 9335 and 8773 nGDip features from VH and VL sequence respectively.

Position Specific N-grams (PSN)

The position specific n-grams (PSN) represent whether specific *n-grams* occur in specific positions in the protein sequence. The value of each such feature in any sequence will therefore be either 0 or 1 (on or off). We have considered n-grams for n = 1, 2 and 3 for PSN.Total number of PSN features are 210 and 291 from VH and VL respectively.

3.3.2 Feature Selection and Ranking

The generated number of features from both VH and VL sequences is huge. It is infeasible to train a machine learning model on such a huge number of features but with so few(137) data points. So we needed to reduce the size of feature space. We have done it in two steps. In first step we have got rid of the features with low importance. In the next step, we ranked the remaining features.

Feature Reduction

To reduce the huge quantity of features we have leveraged a special property of *Random Forest(RF)* algorithm. We have previously discussed about this algorithm in Chapter 2. *Random Forest* can both be used as a *classifier* and *regressor*. As our problem is defined as a regression problem, we have used *Random Forest Regression* in this study. RF regressor is basically the ensemble of several decision trees. Every node in each decision tree divides the dataset based on a certain feature. Which feature to use depends on a locally optimal condition node impurity. In the case of regression, as in our case, variance is used as node impurity. Feature importance is a term denoting how much each feature decreases the weighted impurity in a tree. For a

forest, the mean decrease in impurity for each feature can be calculated and used as feature importance. We calculated feature importance for each feature in an individual feature space(for example n-Grams for AC-SINS, nGDip for HIc etc). Only features with a positive importance scores were kept and other features were discarded.

After applying the above algorithm,feature space is reduced significantly.The number of features in various feature space is shown in Tables 3.1,3.2 and 3.3.

	nGram	**nGDip**	**PSN**
VH	613	844	29
VL	479	701	55

Table 3.1: Reduced Feature Space for AC-SINS

	nGram	**nGDip**	**PSN**
VH	748	2418	68
VL	703	2171	115

Table 3.2: Reduced Feature Space for HIC

	nGram	**nGDip**	**PSN**
VH	592	1719	84
VL	556	1539	113

Table 3.3: Reduced Feature Space for PSR

Feature Ranking

This reduced feature spaces were combined for each target variable(i.e AC-SINS,HIC and PSR). All of this features were ranked using SVM-RFE [36]. This is a special feature ranking technique based on another algorithm Support Vector Machine [38]. SVM was first run on the entire dataset using the procedure decribe in [36] and the 25 least ranked features were eliminated. Then same procedure was repeated again recursively. This recursion was repeated again until all the features are eliminated. Thus we can obtain a final ranking of the features that can be used for further analysis.

3.4 Prediction

Now that we have reduced the feature space and ranked the remaining features we can use this features to train machine learning models. We used several machine learning models such as SVM, Random Forest, Linear Regression and Lasso Regression to predict the target variables. To determine the required number of features needed to predict the target variable, we trained

each of the model with different number of features(starting from very small number of features and gradually increasing the number features based on the ranked feature space). As a testing method we have used *Jack-knife cross validation*. Jack-knife cross validation is a method where the prediction model is trained with N-1 training samples(total N training samples) and tested on the left out sample. This procedure is repeated by leaving out each of the training samples. The testing sample is different in each iteration, since a specific sample is left out each time. This partitioning can be done in only one way. That is why the result of Jack-knife cross validation is always unique. This is a strong advantage of this testing method compared to other *k-fold cross validation* methods. The disadvantage of this technique is that it is comparatively slower. This was not a big problem, as our dataset is not that large.

3.5 Evaluation Metrics

After we have performed Jack-knife cross validation, we need some metrics to compare different models. We have used the following metrics for evaluating different machine learning models.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is the average over absolute differences between prediction and actual observations. This metric gives equal weight to all the individual points. This metric does not consider the direction of the errors.

$$MAE = 1/n \sum_{i=1}^{n} |x_i - \bar{x}| \tag{3.1}$$

Root Mean Square Error(RMSE)

Root Mean Square Error(RMSE) is the square root of the average of squared differences between prediction and actual observation. RMSE also measures the average magnitude of the error. RMSE gives relatively greater weight to the larger errrors. So if large errors are undesirable RMSE can be a better measure compared to MAE.

$$RMSE = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.2}$$

R^2 Value

R^2 Value is the proportion of variance in the dependant variable that is predictable from the independent variable. In layman terms, this is a statistical measure of how closely the fitted line follows the trend of the observations. The value of R^2 can be determined from

the following equation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} |x_i - \hat{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|} \tag{3.3}$$

R^2 has a value between 0 and 1, where 1 means the fitted line perfectly matches the observations and 0 means there is no correlation between the fitted line and the observations.

Adjusted R^2 Value

R^2 always increases with the increase of the independent variables, which can cause overfitting. This issue is resolved in Adjusted R^2. Adjusted R^2 also determines how much of the correlation is determined by the addition of another independent variable. In this study, almost all of the R^2 and Adjusted R^2 values were same.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{3.4}$$

where,
R^2 = sample R^2
N = number of samples
p = number of predictors

Chapter 4

Results

4.1 Experimental Results

We used Linear Regression, Support Vector Machine(SVM),Lasso Regression and Random Forest as regression algorithms.And Mean Accuracy Error(MAE),Root Mean Square Error(RMSE),R-Squared and Adjusted R-Squared were used as evaluation metrics.

Regressor	**Feature Size**	**MAE**	**RMSE**	R^2	**Adjusted** R^2
SVM	110	1.432	1.864	0.975	0.974
Random Forest	20	4.268	6.256	0.713	0.711
Linear Regression	60	1.952	2.507	0.947	0.941
Lasso Regression	50	1.884	2.344	0.949	0.948

Table 4.1: Best Results for AC-SINS

For AC-SINS, SVM provides best results for feature size 110. Similarly, the feature sizes are 20,60 and 50 for random forest,linear regression and lasso regression respectively. In Table-4.1, we have shown only best results for these four classifiers. Clearly, SVM has outperformed others. In Figure-4.1,we observe - with the increase of feature size, the performance metrics(MAE,RMSE) decrease for SVM and lasso regression.Again, R^2 and adjusted R^2 increase with increasing of feature size. But random forest and linear regression perform poorly for AC-SINS.

For HIC Retention time, Table-4.2 shows the best results for different performance metrics.SVM outperforms others as usual but linear regression and random forest provide worse performance than AC-SINS.

Here in Figure-4.2,we observe - with the increase of feature size, the performance metrics(MAE,RMSE) are not decreasing smoothly for SVM and lasso regression. Also random forest and linear regression perform poor as before.

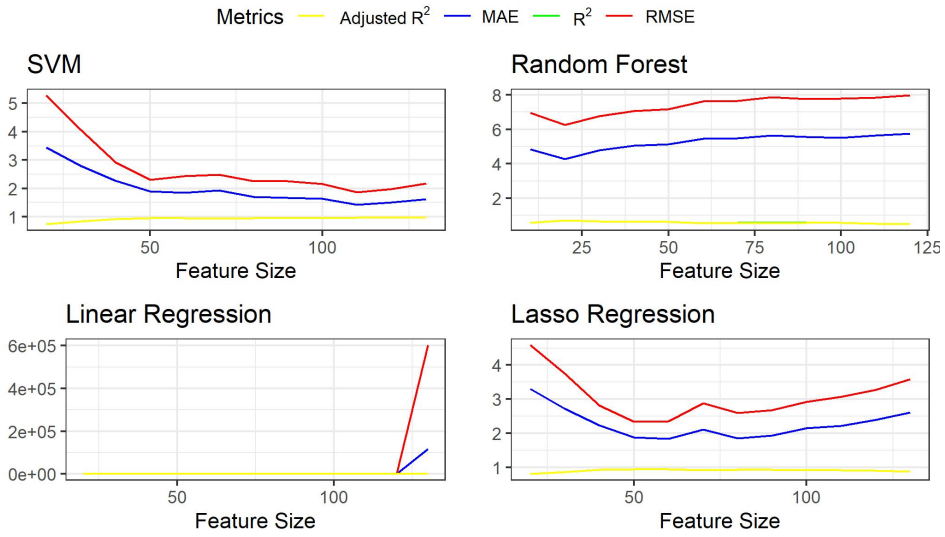

Figure 4.1: Effects of Feature Size on different performance metrics for AC-SINS

Regressor	**Feature Size**	**MAE**	**RMSE**	R^2	**Adjusted** R^2
SVM	90	0.306	0.439	0.957	0.957
Random Forest	110	0.886	1.881	0.194	0.188
Linear Regression	30	252.7	2088.6	0.732	0.730
Lasso Regression	30	0.539	1.0326	0.820	0.819

Table 4.2: Best Results for HIC Retention Time

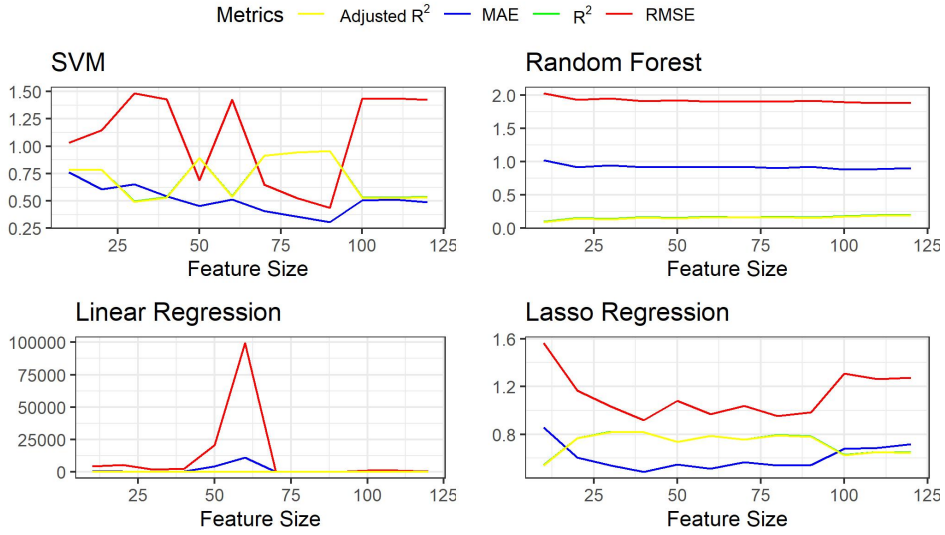

Figure 4.2: Effects of Feature Size on different performance metrics for HIC Retention Time

For PSR SMP Score, Table-4.3 shows the best results for different performance metrics as before. SVM outperforms others as usual and linear regression performs better than before. In Figure-4.3,we observe that MAE and RMSE are decreasing smoothly and R^2 and adjusted R^2 are increasing for SVM and random forest with the increasing of feature size. Finally, we decide that the overall performance of SVM is more promising for sequence based feature extraction

Regressor	Feature Size	MAE	RMSE	R^2	Adjusted R^2
SVM	120	0.030	0.037	0.974	0.974
Random Forest	80	0.116	0.154	0.505	0.502
Linear Regression	70	0.041	0.056	0.925	0.924
Lasso Regression	30	0.073	0.098	0.802	0.801

Table 4.3: Best Results for PSR SMP Score

method.

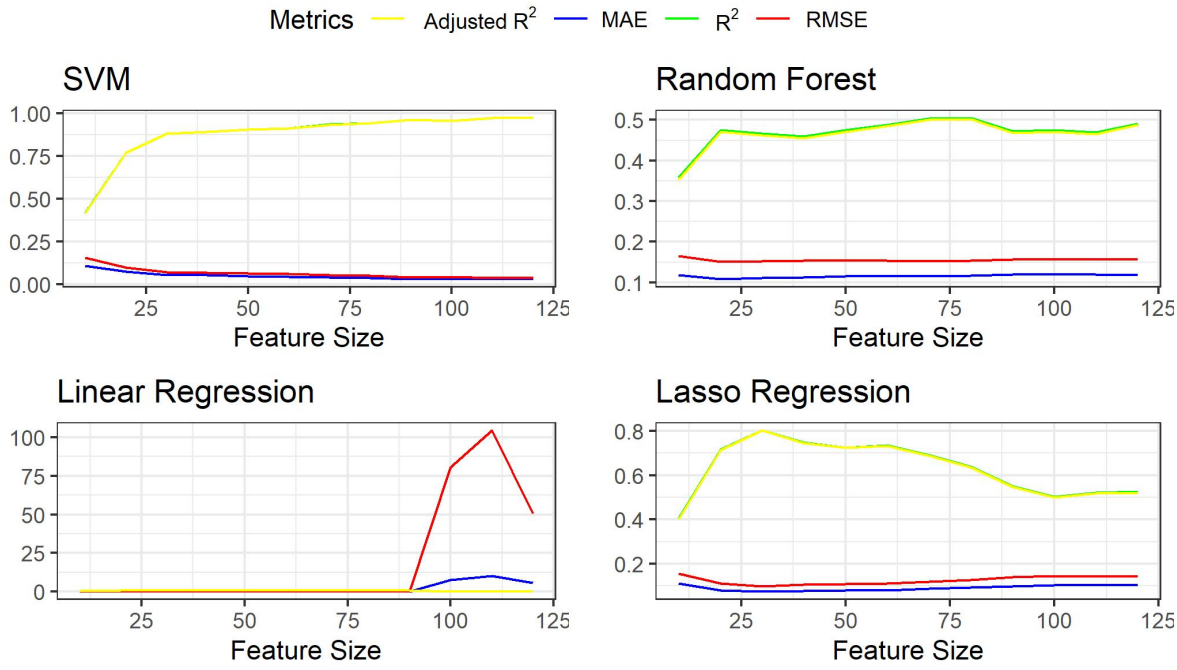

Figure 4.3: Effects of Feature Size on different performance metrics for PSR SMP Score

4.2 Discussion

We have seen SVM outperforming all other regressors for every target variable. A point to be noted is that best results of SVM were obtained by using *linear* kernel. We have also experimented with *radial* kernel, which have produced poor results. The best results were also obtained on and around a feature size of 100(110,90 and 120). Overall our prediction models predict pretty accurately. All the models trained with SVM with optimal feature size have a low MAE and RMSE, on the other hand, have a very high R^2 value(0.95-0.97). This indicates a very good fit of the trained model.

Another important point to be mentioned, a general thought is that prediction accuracy should increase as we include more features to the training model because the features are ranked according to their importance. This is not true in general. More features does not always

guarantee improvement in terms of accuracy. When AC-SINS and PSR were used as target variable, other than linear regression, we have seen a general trend of reduction of RMSE as feature size increases to a certain point(usually the optimal feature size). After that with the increase of feature size RMSE has increased. This can be attributed to the fact that after a certain point the increase of feature size tends to overfit the data. So the validation RMSE and other metrics deteriorate.

A very crucial observation is the poor performance of linear regression. Also, as the feature size increases upto a certain point linear regression seems to perform very poorly - RMSE seems to explode. The underlying reason of this behaviour can be a small size of the dataset. As the feature size increases, given the number of data points are fixed, linear regression starts to perform very poorly. Toverfihat is why lasso regression is also used. Lasso regression uses regularization for not overfitting. This alleviates the problem we were facing with linear regression. That is why we see a better performance from lasso regression as the feature size increases.

Random Forest(RF) has not performed well in any case. A notable issue of RF is poor performance as regressor. Because RF does not predict precise continuous data prediction as is required with regression. Also RF does not predict beyond the range of training data. This may overfit the dataset. For all of these reasons RF seems not to perform at the level of SVM or other regressors here.

There is an apparent anomaly in HIC retention time prediction with SVM. Because RMSE and other metric does not seem to have a monotonic relationship with feature size. Rather the correlation is a bit zigzagged unlike other target variables. This apparent anomaly can be attributed to the fact that only sequence order information may not be enough for predicting HIC retention time. Previous works have showed that this biophysical property can be correlated to amino-acid or atomic propensities weighted by the surface areas obtained from protein 3-dimensional structures [19]. As we have only used sequence order information as features the model may have shown this non-monotonic relationship. However, our model with optimal feature size still performs very well in this regard.

Chapter 5

Conclusion

Predicting developability of monoclonal antibodies have been a important issue in recent years. Our proposed models has successfully predicted three biophysical assays from heavy and light chain sequences of antibody. We have predicted specific values of each biophysical assays rather than classifying which antibody may have a better developability in terms of these assays(as was done in [18]). On the other hand, all of our models with optimal feature size has outperformed the work done in [17]. The low computational expense and a high accuracy makes our method very feasible for reducing cost of monoclonal antibodies development. There is also some limitations to our work because we have only used sequence order information of the antibodies. We have not considered the 3-dimensional structure of the protein, properties of the solution etc. Also according to *Hebditch et al.*, charge and hydrophobicity, calculated from amino acid propensity, are very important in predictive models. As we have used these features, our result can suffer from a certain degree of sensitivity. But we believe, overall the performance of our model to be satisfactory. There is still plenty of improvement that can be done on this topic in the future. A simple improvement can be taking the 3-dimensional structure of the protein in account, We also want to predict the remaining other biophysical assays in the future as an extension to the current research.

References

[1] T. E. o. E. Britannica, "Antibody | Definition, Structure, Function, & Types," 2020. [Online; accessed 15-February-2021].

[2] G. Köhler and C. Milstein, "Continuous cultures of fused cells secreting antibody of predefined specificity," *nature*, vol. 256, no. 5517, pp. 495–497, 1975.

[3] S. Singh, N. K. Tank, P. Dwiwedi, J. Charan, R. Kaur, P. Sidhu, and V. K. Chugh, "Monoclonal antibodies: a review," *Current clinical pharmacology*, vol. 13, no. 2, pp. 85–99, 2018.

[4] M. A. Hooks, C. S. Wade, and W. J. Millikan Jr, "Muromonab cd-3: a review of its pharmacology, pharmacokinetics, and clinical use in transplantation," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 11, no. 1, pp. 26–37, 1991.

[5] V. F. Azevedo, L. D. C. Troiano, N. B. Galli, A. Kleinfelder, N. M. Catolino, and P. C. U. Martins, "Adalimumab: A review of the reference product and biosimilars," *Biosimilars*, vol. 6, pp. 29–44, 2016.

[6] M. L. d. Santos, W. Quintilio, T. M. Manieri, L. R. Tsuruta, and A. M. Moro, "Advances and challenges in therapeutic monoclonal antibodies drug development," *Brazilian Journal of Pharmaceutical Sciences*, vol. 54, no. SPE, 2018.

[7] H. Kaplon, M. Muralidharan, Z. Schneider, and J. M. Reichert, "Antibodies to watch in 2020," in *MAbs*, vol. 12, p. 1703531, Taylor & Francis, 2020.

[8] H. Kaplon and J. M. Reichert, "Antibodies to watch in 2019," in *MAbs*, vol. 11, pp. 219–238, Taylor & Francis, 2019.

[9] R.-M. Lu, Y.-C. Hwang, I.-J. Liu, C.-C. Lee, H.-Z. Tsai, H.-J. Li, and H.-C. Wu, "Development of therapeutic antibodies for the treatment of diseases," *Journal of biomedical science*, vol. 27, no. 1, pp. 1–30, 2020.

[10] D. B. Fogel, "Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review," *Contemporary clinical trials communications*, vol. 11, pp. 156–164, 2018.

[11] M. I. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A. P. Lewis, A. Bujotzek, J. Shi, and C. M. Deane, "Five computational developability guidelines for therapeutic antibody profiling," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4025–4030, 2019.

[12] T. M. Lauer, N. J. Agrawal, N. Chennamsetty, K. Egodage, B. Helk, and B. L. Trout, "Developability index: a rapid in silico tool for the screening of antibody aggregation propensity," *Journal of pharmaceutical sciences*, vol. 101, no. 1, pp. 102–115, 2012.

[13] O. Obrezanova, A. Arnell, R. G. de la Cuesta, M. E. Berthelot, T. R. Gallagher, J. Zurdo, and Y. Stallwood, "Aggregation risk prediction for antibodies and its application to biotherapeutic development," in *MAbs*, vol. 7, pp. 352–363, Taylor & Francis, 2015.

[14] Q. Hou, R. Bourgeas, F. Pucci, and M. Rooman, "Computational analysis of the amino acid interactions that promote or decrease protein solubility," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[15] K. Sankar, S. R. Krystek Jr, S. M. Carl, T. Day, and J. K. Maier, "Aggscore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 11, pp. 1147–1156, 2018.

[16] T. Jain, T. Sun, S. Durand, A. Hall, N. R. Houston, J. H. Nett, B. Sharkey, B. Bobrowicz, I. Caffry, Y. Yu, *et al.*, "Biophysical properties of the clinical-stage antibody landscape," *Proceedings of the National Academy of Sciences*, vol. 114, no. 5, pp. 944–949, 2017.

[17] M. Hebditch and J. Warwicker, "Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies," *PeerJ*, vol. 7, p. e8199, 2019.

[18] A. M. Dzisoo, J. Kang, P. Yao, B. Klugah-Brown, B. A. Mengesha, and J. Huang, "Ssh: A tool for predicting hydrophobic interaction of monoclonal antibodies using sequences," *BioMed Research International*, vol. 2020, 2020.

[19] T. Jain, T. Boland, A. Lilov, I. Burnina, M. Brown, Y. Xu, and M. Vásquez, "Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning," *Bioinformatics*, vol. 33, no. 23, pp. 3758–3766, 2017.

[20] *The American Heritage dictionary of the English language*. Boston: Houghton Mifflin, 4th ed. [updated] ed., 2006.

[21] Y. Liu, I. Caffry, J. Wu, S. B. Geng, T. Jain, T. Sun, F. Reid, Y. Cao, P. Estep, Y. Yu, M. Vásquez, P. M. Tessier, and Y. Xu, "High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy," *mAbs*, vol. 6, pp. 483–492, Mar. 2014.

[22] T. Steimer, G. Theintz, P. Sizonenko, and W. Herrmann, "Hydrophobic interaction chromatography (HIC) for the separation of protein-bound and free steroids. Application to binding protein and receptor assays," *Journal of Steroid Biochemistry*, vol. 23, pp. 955–965, Dec. 1985.

[23] T. Jain, T. Boland, A. Lilov, I. Burnina, M. Brown, Y. Xu, and M. Vásquez, "Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning," *Bioinformatics*, vol. 33, pp. 3758–3766, 08 2017.

[24] Y. Xu, W. Roach, T. Sun, T. Jain, B. Prinz, T.-Y. Yu, J. Torrey, J. Thomas, P. Bobrowicz, M. Vasquez, K. D. Wittrup, and E. Krauland, "Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool," *Protein Engineering Design and Selection*, vol. 26, pp. 663–670, Oct. 2013.

[25] T. M. Mitchell, *Machine Learning*. 1997. OCLC: 36417892.

[26] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design '96* (J. S. Gero and F. Sudweeks, eds.), pp. 151–170, Dordrecht: Springer Netherlands, 1996.

[27] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832–844, Aug. 1998.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 1995.

[29] A. Jarasch, H. Koll, J. T. Regula, M. Bader, A. Papadimitriou, and H. Kettenberger, "Developability assessment during the selection of novel therapeutic antibodies," *Journal of Pharmaceutical Sciences*, vol. 104, no. 6, pp. 1885–1898, 2015.

[30] N. Kohli, N. Jain, M. L. Geddie, M. Razlog, L. Xu, and A. A. Lugovskoy, "A novel screening method to assess developability of antibody-like molecules," in *MAbs*, vol. 7, pp. 752–758, Taylor & Francis, 2015.

[31] V. I Razinkov, M. J Treuheit, and G. W Becker, "Methods of high throughput biophysical characterization in biopharmaceutical development," *Current drug discovery technologies*, vol. 10, no. 1, pp. 59–70, 2013.

[32] A. Man, H. Luo, S. V. Levitskaya, N. Macapagal, and K. J. Newell, "Optimization of a platform process operating space for a monoclonal antibody susceptible to reversible and irreversible aggregation using a solution stability screening approach," *Journal of Chromatography A*, vol. 1597, pp. 100–108, 2019.

[33] N. J. Agrawal, B. Helk, S. Kumar, N. Mody, H. A. Sathish, H. S. Samra, P. M. Buck, L. Li, and B. L. Trout, "Computational tool for the early screening of monoclonal antibodies for their viscosities," in *MAbs*, vol. 8, pp. 43–48, Taylor & Francis, 2016.

[34] A. Mahn, M. E. Lienqueo, and J. C. Salgado, "Methods of calculating protein hydrophobicity and their application in developing correlations to predict hydrophobic interaction chromatography retention," *Journal of chromatography A*, vol. 1216, no. 10, pp. 1838–1844, 2009.

[35] A. T. Hanke, M. E. Klijn, P. D. Verhaert, L. A. van der Wielen, M. Ottens, M. H. Eppink, and E. J. van de Sandt, "Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties," *Biotechnology progress*, vol. 32, no. 2, pp. 372–381, 2016.

[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[37] J.-M. Chang, E. C.-Y. Su, A. Lo, H.-S. Chiu, T.-Y. Sung, and W.-L. Hsu, "PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, pp. 693–710, Feb. 2008.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

Appendix A

Figures

As we saw earlier, among the four regressors- SVM preforms better than others.Here we include result graphs of the comparison of actual and predicted data for SVM only.All these graphs show that actual data and predicted results almost converge.

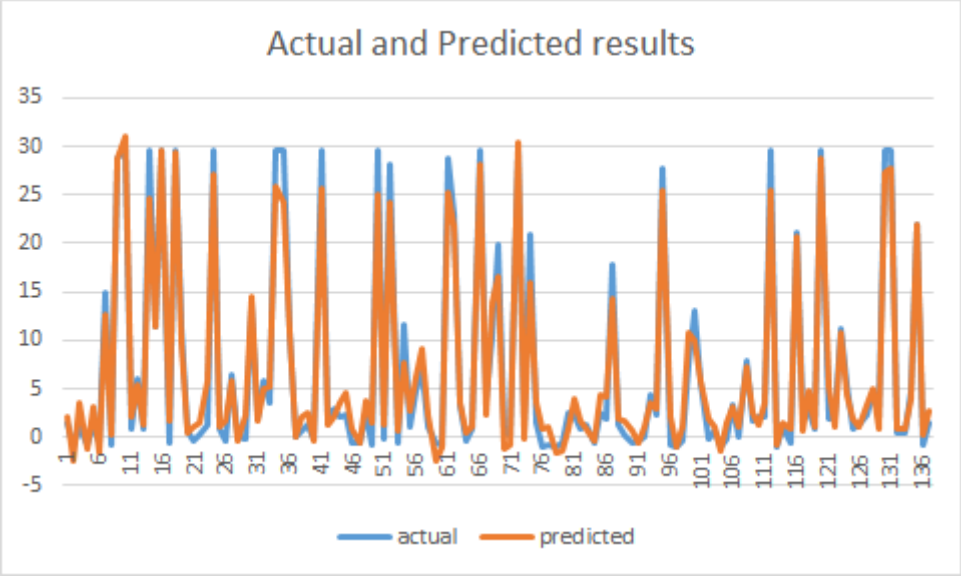

Figure A.1: Actual vs Predicted results for AC-SINS using SVM

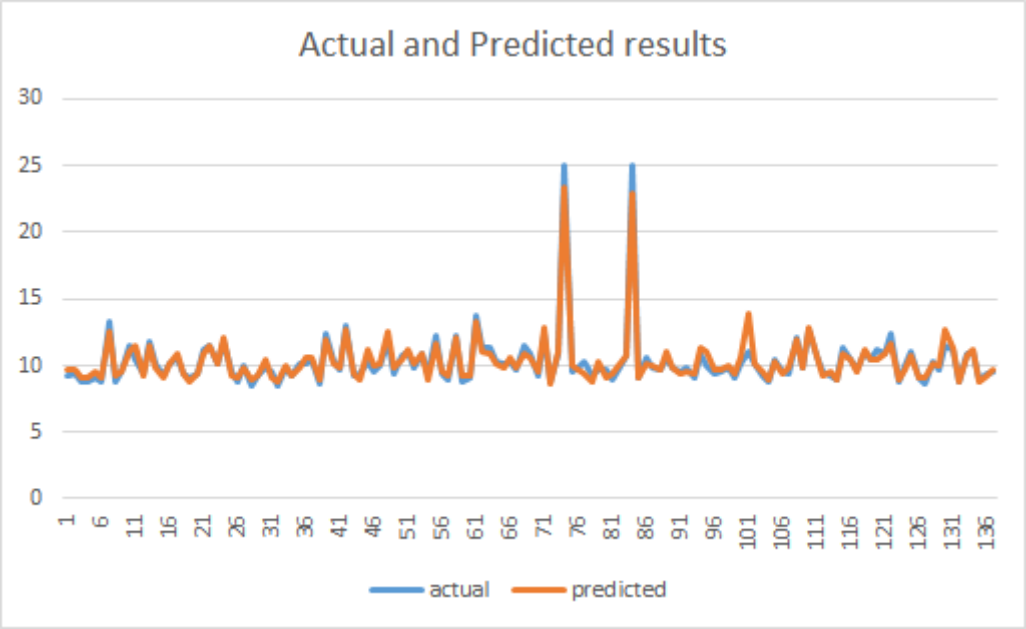

Figure A.2: Actual vs Predicted results for HIC Retention Time using SVM

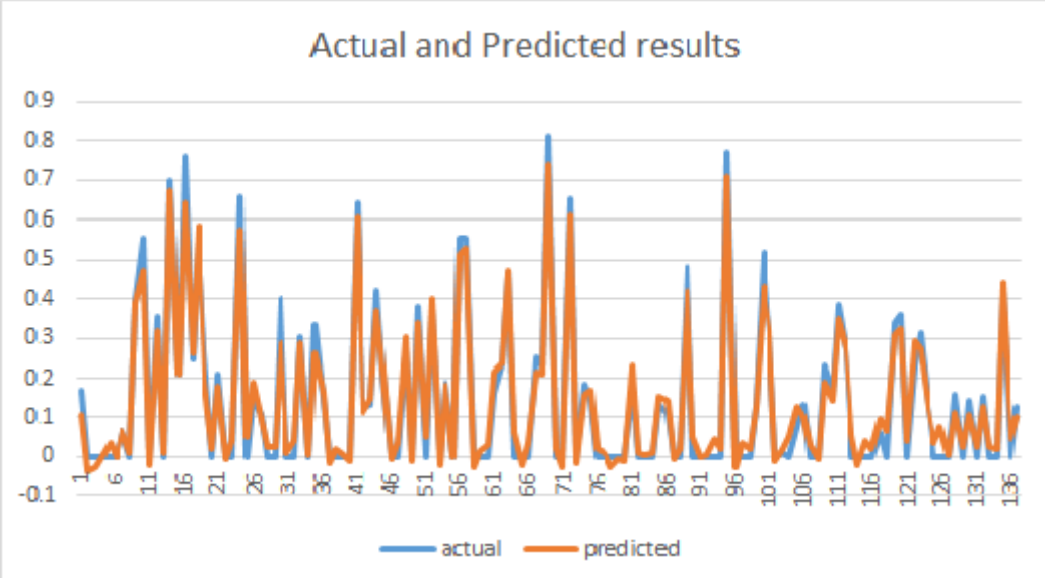

Figure A.3: Actual vs Predicted results for PSR SMP Score using SVM

Appendix B

Codes

All the analysis was done using R version 3.6.1. We have leveraged several R packages for this study. We have used two Windows 10 machine with Intel Core-i5 processor.

All of the codes used in this study can be found on the following github link :
https://github.com/rizvi23061998/cse400Thesis/

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Friday 19th February, 2021 at 11:03am.