BAD BLOOD: COMBINING DATA ANALYTICS AND CHEMICAL KINETICS TO STUDY HUMAN BLOOD COAGULATION IN CERTAIN DISEASES

A Dissertation

by

JAYAVEL ARUMUGAM

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Arun Srinivasa
Committee Members,	J.N. Reddy
	Krishna Narayanan
	Alan Freed
Head of Department,	Andreas Polycarpou

May 2017

Major Subject: Mechanical Engineering

ABSTRACT

Complete description of blood coagulation pathways with respect to patient-specific characterization presents a major challenge. Characteristics of blood coagulation vary drastically between patients. It is essential to characterize abnormalities in blood coagulation to diagnose and treat cardiovascular diseases better.

Given the paucity of patient-specific data to characterize and model the system, there is a greater need to regularize patient-specific models and methods effectively. In this dissertation, we formulate actionable questions and describe our methodology and results.

First, we explore a practical application for using models to classify acute coronary syndrome and coronary artery disease. The classification models were built based on a chemical kinetics model reported in the literature. In a diagnostic setting, the classification models could be employed to screen thousands of patients with greater certainty every year.

Second, we propose a simplified model for a key part of the blood coagulation cascade that demonstrates robust predictive capabilities. The model predicts prolonged activity of thrombin, an important enzyme in the clotting process, in certain plasma factor compositions. The activity sustains beyond the time which is conventionally considered to be the end of clotting. This observation along with the simplified model is a necessary step towards effectively studying clotting in realistic geometries.

DEDICATION

To my parents who taught me to put education before everything else.

ACKNOWLEDGMENTS

I am deeply humbled by the privilege of graduate student life at Texas A&M University. I am grateful for the encouragement, support, and friendship of a large number of people who have helped me grow as a human being.

I am thankful to my dissertation advisor Dr. Arun Srinivasa for helping me grow as a stronger researcher. I am also thankful to Dr. J. N. Reddy for guiding me during my graduate studies. His monumental work is truly inspiring. I extend my deep sense of gratitude for Dr. Krishna Narayanan whose bayesian perspective has certainly affected my thinking. I would like to thank Dr. Alan Freed for making me appreciate the boundaries of scientific methods better. His smile has always been contagious. I also thank Dr. K. R. Rajagopal whose mere presence is enlightening. His love for philosophy and continuum mechanics has certainly left a mark on me as an engineer.

I am thankful to the very knowledgeable faculty members at TAMU. I particularly thank Dr. N. Sivakumar (Mathematics), Dr. Shankar Bhattacharyya (Electrical Engineering), and Dr. Satish Bukkapatnam. Dr. Satish Bukkapatnam has been a tremendous source of inspiration as a teacher.

I thank the many current and past students who have made me feel incredibly fortunate. Srikrishna, Pritha, Naveen, Nazanin, Wang, Ashish, Balaji Ganesan, Shreyas, Shriram, Atul, Hoang, Zimo, Ashif, Afreen, Alagappan, Bharathwaj, and Priya have helped me grow as a better engineer. I certainly appreciate the comic relief that Giridhar provided. Thanks to Dan Kiniry for proofreading parts of this dissertation. Mukundan, Kaarthik, Kabali, Sudarshan, Atul, and Mahesh have made me feel part of a family and formed a second home here in the US. Kaarthik Sundar

was my roommate for quite longer than I had anticipated. His influence on my coding and his inspiring diligence will go a long way. Harsha has been a source of artistic inspiration. Living is made more worthwhile by musicians such as him. Mukundan and Sudarshan have taught me valuable life lessons. I am also grateful to have friends such as Poornima, Vivek, Nirmal, and Hareesh.

I thank my family members who have stood behind my idiosyncrasies and nitty-gritties: my brother Senthil, sister-in-law Geetha, nephew Gurubaran, and parents Srikanthy and Arumugam. Life is a blessing with their unconditional love and support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Arun Srinivasa, Professor J.N. Reddy, and Professor Alan Freed of the Department of Mechanical Engineering and Professor Krishna Narayanan of Electrical and Computer Engineering.

All work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by Graduate Teaching Academy awards from Texas A&M University and a research assistantship from National Science Foundation under grant no. 1028894.

NOMENCLATURE

ACS	Acute coronary syndrome
CAD	Coronary artery disease
GMM	Gaussian mixture model
MDGini	Mean decrease in gini index
OOB	Out-of-bag
factor I	Fibrinogen
factor Ia	Fibrin
factor II	Prothrombin
IIa	Thrombin
AT or ATIII	Antithrombin
TFPI	Tissue factor pathway inhibitor
Tf	Tissue factor
APC	Activated protein C

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION*

"One is struck by the complexity of this figure that I am not even attempting to draw."

– Henri Poincare, *New Methods of Celestial Mechanics*

1.1 Chapter Outline

We motivate the problem by highlighting its socioeconomic burden. Then we introduce the blood coagulation system, briefly review the corresponding literature, and also elaborate the challenges faced while modeling and solving the chemical reaction kinetics of blood coagulation. We describe the deficiencies in the area currently and discuss alternative approaches that address these deficiencies. We also describe the objective and scope.

1.2 Motivation

In the United States, heart diseases were the leading cause of death in the past two centuries [1, 2]. Identifying patients at risk of acute coronary syndromes (ACS) [3] and predicting progress of disease could help provide timely medical intervention; understanding the physiology of the diseases in patient-specific terms could also help design better drugs and monitor treatment more effectively.

ACS refer to a set of diseases[1] that results in a sudden failure of proper functioning

[1]In this dissertation, we do not distinguish between different types of ACS. The reader could find relevant information here [3].

*Part of this chapter is reprinted with permission from "Random Forests Are Able to Identify Differences in Clotting Dynamics from Kinetic Models of Thrombin Generation" by Jayavel Arumugam, Satish T. S. Bukkapatnam, Krishna R. Narayanan, and Arun R. Srinivasa. PloS one, e0153776, Copyright [2016] by Arumugam et al.

of the heart. It is caused due to decreased or blocked blood flow in the arteries of the heart [4]. Infarction can be avoided if flow to the affected artery is restored within 30 minutes but there is no salvage after 6 hours [5]. Timely intervention is critical in reducing costs and saving lives [6].

The cost of ACS is more compared to other health conditions. Annual direct cost in ACS is $\sim$ \$44,023 which is higher compared to $\sim$ \$9,955 in hypertension or $\sim$ \$13,858 in diabetes [7]. Advanced diagnostic modalities are expected to play a major role in reducing unnecessary hospitalizations and hence the associated costs [3]. In addition to diagnosing ACS properly, we would also like to monitor treatment in patient-specific terms. For example, we would like to prognose the course of disease and predict mortality [8]. Treatment is known to cause excessive bleeding and patients continue to be at risk of recurrence of heart diseases [9]. In certain ACS patients, there is a strong association between bleeding and death [10]. Therefore thorough bleeding assessments are recommended before administration of antithrombotic drugs [11].

However, bleeding is a complicated phenomenon and there is a drastic variation in characteristics from one patient to another. Current methods to probe the coagulation system need improvement [12]. Better methods to model, evaluate and characterize the system are sought.

1.3 Blood Coagulation Mechanism

Blood coagulation is a process which stops blood loss upon injury. Anand et al. [13] comprehensively reviewed various mechanical and biochemical factors involved in blood coagulation. In addition, models considering genetic, biochemical, and mechanical factors in blood coagulation have been previously discussed [14, 15].

There are two ways to assess risk:

1. Holistic approaches including considering the effect of factors like smoking, age, air travel, lack of exercise, etc., on risk assessment.

2. Mechanism based approaches that study physiological changes in clotting and associated biomechanical properties.

We will focus on mechanism based approaches where diagnosis and design of cure happens. Different aspects of the physiology that are of interest towards disease diagnosis, monitoring and treatment include:

- Mechanical properties of the artery [16].
- Non-newtonian fluid flow aspects of blood coagulation [17].
- Enzyme kinetics underlying coagulation [18, 19].
- Convection-reaction-diffusion models describing transport of platelets and the protein factors that couple blood chemistry with fluid flow [13].

We will be concerned with the enzyme kinetics involved in blood coagulation which act as a bottleneck for progress. Three major events occur during blood coagulation: clot initiation, clot propagation, and clot disruption and dissolution. Clot initiation is modeled using three pathways: i) the extrinsic pathway which is initiated by tissue-factor, ii) the intrinsic pathway initiated by contact with a negative surface like glass, and iii) initiation due to platelets.

Clotting dynamics is primarily studied based on the dynamics of a key enzyme thrombin (IIa) [18, 20]. Once initiated, thrombin activates factor VIII and factor V. This results in the formation of intrinsic tenase (IXa-VIIIa) and prothrombinase (Xa-Va) complexes which further activate thrombin. Thrombin catalyses the formation of Fibrin (Ia). Fibrin is stabilized into stable clot by activated XIIIa. Further aspects

of clotting include clot dissolution and disruption. Most of these chemical reactions occur on the surface of activated platelets.

1.4 A Brief History of Blood Coagulation

An excellent historical context and understanding of blood coagulation and its alterations in disorders could be found in [21] and [22]. We briefly introduce the 'classical' and the 'modern' theory of blood coagulation based on these two sources.

1.4.1 Classical Theory

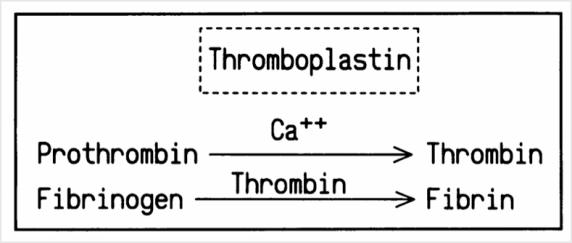

Figure 1.1: Classical theory of clotting. Clotting was explained based on four factors (figure from [22]).

- Blood coagulation was thought of as physical changes in blood [21]. Hippocrates and Aristotle tried to explain coagulation based on cooling. This ancient theory has been invoked many times in the 17th century. William Hewson showed blood could be thawed and that it liquifies before coagulation. This disproved the cooling theory. Another physical explanation was cessation of natural flow of blood[2]. This notion has stood the test of time.

[2]although controversial, the significance of this has been recognized by many eminent hematologists such as Virchow [23, 24]

- Paul Morawitz, and independently Fuld and Spiro, gave the classical theory of blood coagulation (see Figure 1.1) based on 19th century experiments [22]. Three factors - prothrombin, calcium ions, fibrinogen were present in blood. A fourth factor thromboplastin (Tissue Factor Tf) was postulated to be contained within cells like platelets and leukocytes. Tf was postulated to be extruded during injury from damaged tissue cells. Tf reacted with calcium and prothrombin to form thrombin, which converted fibrinogen to fibrin strands of a blood clot [22].

- In 1935, a test was developed based on the classical theory to study defects in hemophilia patients [25]. This test entered clinical domain and goes by the name prothrombin time [22]. This measures the time required to form enough thrombin for clotting.

1.4.2 Modern Theory

- Two different models to initiate clotting were recognized. Clotting initiated in the intrinsic pathway due to contact with external surface. Paul Owren discovered discovered factor V [22]. This was followed by others and new clotting factors were discovered. Thomas Addis found out adding globulin fraction (factor VIII) improved delayed clotting time in hemophilia patients. However, the experimental evidence was ignored since it was inconsistent with existing theories [22]. Additional evidence surfaced due to the work of Patek and Taylor [27]. They identified antihemophilic globulin, now referred to as factor VIII.

- Paul Aggeler and others discovered Hemophilia B patients bled because of factor IX deficiency [22]. Other biochemists discovered several factors involved

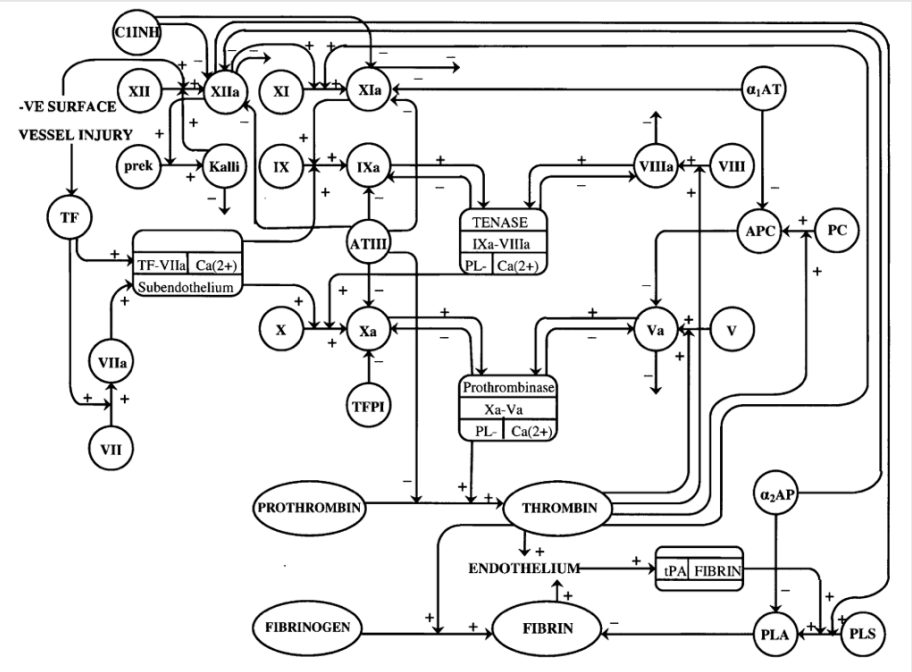

Figure 1.2: Modern theory of clotting. Detailed schematic of clotting reactions in the intrinsic and the extrinsic pathway (figure from [26]). This pathway considers initiation due to the extrinsic as well as the intrinsic pathway. Polymerized fibrin is the clot and its conversion from fibrinogen is catalyzed by thrombin. Actual coagulation is known to be more complicated and models considering hundreds of protein factors have been proposed.

in the clotting phenomena. Inhibitors antithrombin and activated protein C were discovered by studying thrombotic disorders such a venous thrombosis [22].

- Multiple active forms of thrombin were discovered [28]. Factors like V were isolated and characterized [29]. Kinetics of active complexes involved in clotting were studied. Thrombin generation simulation models [30, 31, 32] and platelet activation models were developed [33]. Faster and better experimental methods to estimate thrombin generation were designed [34]. Spatiotemporal models for dynamics of clotting were developed [35] followed by comprehensive models for blood coagulation [13].

1.4.3 Confluence

- There have been tremendous advancements in diagnosis, treatment [36], and modeling [15]. Panteleev and Hemker [12] indicate that the standard assays are not sensitive and specific for many major hemostatic disorders. Biochemists and hematologists acknowledge the need to consider aspects of geometry and flow [37].

- Taylor and Humphrey [38] reviewed open problems in vascular biomechanics. Patient-specific geometry modeling, simulations with more realistic boundary conditions, multiscale models that combine molecular mechanisms with clinical manifestation are some of the discussed problems. Work in the field of biomechanics borrows heavily from the results of biochemists and the models for blood coagulation that are currently in use are complex. A detailed depiction of clotting reactions is shown in Figure 1.2.

- Model simplification is considered a necessity from both the sides.

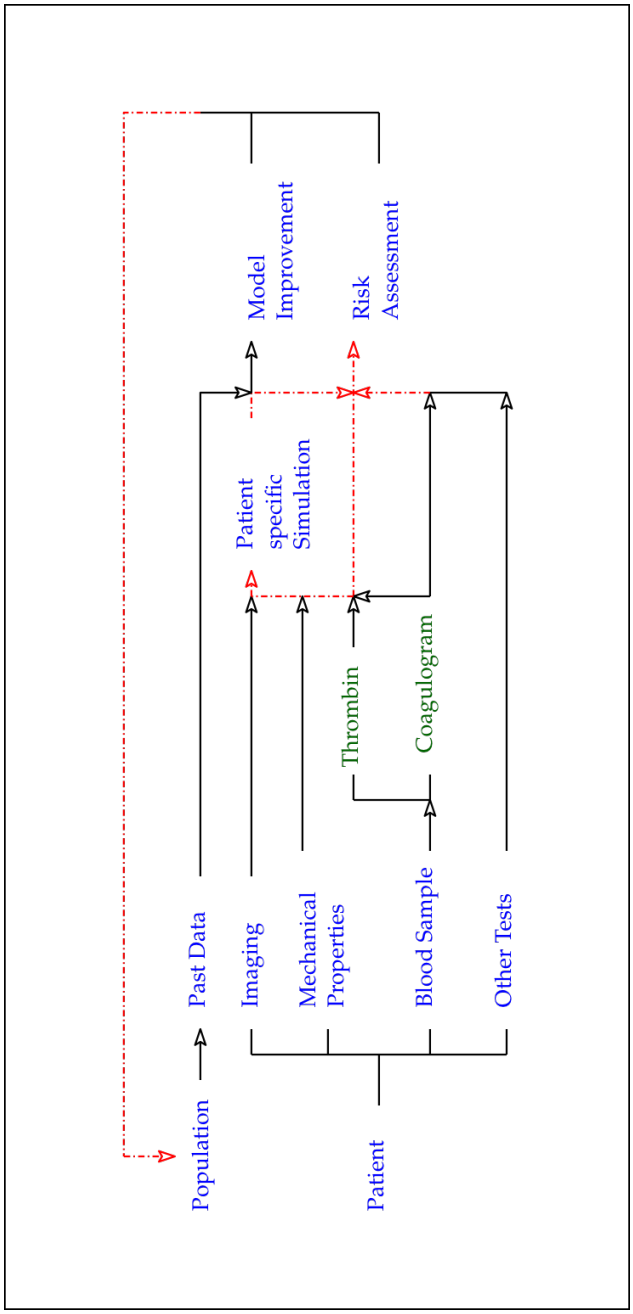

Figure 1.3: A schematic of process involved in patient risk assessment. Missing links are highlighted in red. Data that are the most insightful and useful (diagnosis, intervention, and cure primarily happens at this level) are shown in green.

This is one of the best times to be involved in blood coagulation research. There is a confluence of scientific fields. Tremendous improvements have been made in the treatment of cardiovascular diseases but there is a still a long way to go (see Figure 1.3).

1.5 Abnormalities in the Blood Coagulation System

Typically, abnormality in coagulation could occur when:

1. **one of the protein factors is missing or present when not needed**. Hemophilia or hypocoagulation has been extensively studied are usually due to missing factors. The most important inactive factors and inhibitors are shown in Table 1.1. Absolute deficiency in most these factors is either fatal or leads to extreme disorders.

2. **stoichiometry of certain factors are abnormal**. Plasma factor composition affects the dynamics of reactions that happen during and after coagulation [39, 40]. Adding a new dose of inactive plasma factors results in restoration of thrombin generation without clot initiation triggers [41].

3. **kinetics are abnormal**. Rates are often a combined effect of all factors. Moreover, the effect of rates is complicated due to complexity of chemical kinetics. It could further create complications in other pathways, systems and scales. For example, rates combined with diffusion could further determine the size of the clot or with convection determine occurrence of clot downstream.

The absence of protein factors can be easily identified and dealt with. However, abnormalities in stoichiometry or kinetics are harder to quantify. In most cases, changes in stoichiometry have to quantified before changes in kinetics are observed.

Table 1.1: Inactive plasma factors and associated diseases due to their deficiency.

Protein	**Associated disease due to deficiency**
Factor VII	Rare, hemophilia-like bleeding disorder
Factor X	Rare, bleeding disorders
Factor IX	Hemophilia B
Factor II	Bleeding disorders
Factor VIII	Hemophilia A
Factor V	Rare, mild form of hemophilia
TFPI	Thrombotic diseases
ATIII	Thrombotic diseases

A simplified depiction of the extrinsic pathway of blood coagulation is shown in Figure 1.4. Thrombin plays a central and a multifunctional role [18, 42]. Abnormalities in the coagulation system are reflected in thrombin generation curves and offer descriptive explanations. In the last two decades, sustained interest has been shown in empirical and computational thrombin generation assays [43, 44, 45, 46] to study the coagulation system under abnormal conditions and to use it for patient-specific diagnosis and treatment. The aim is to quantify hypo- and hypercoagulable states of blood using thrombin generation curves compared to those in healthy individuals [47, 40, 48].

Simulations of thrombin generation during blood coagulation in the extrinsic pathway is known to discriminate ACS and coronary artery disease (CAD) [49]. When the simulations results are compared for the thrombin generation parameters, samples from the two groups differ in a statistically different way. Thrombin generation is higher in ACS population suggesting hypercoagulation. Further, the blood coagulation reactions in acute myocardial infarction patients is known to be markedly modified compared to CAD patients [50].

Similarly, thrombin generation is higher in chronic obstructive pulmonary disease

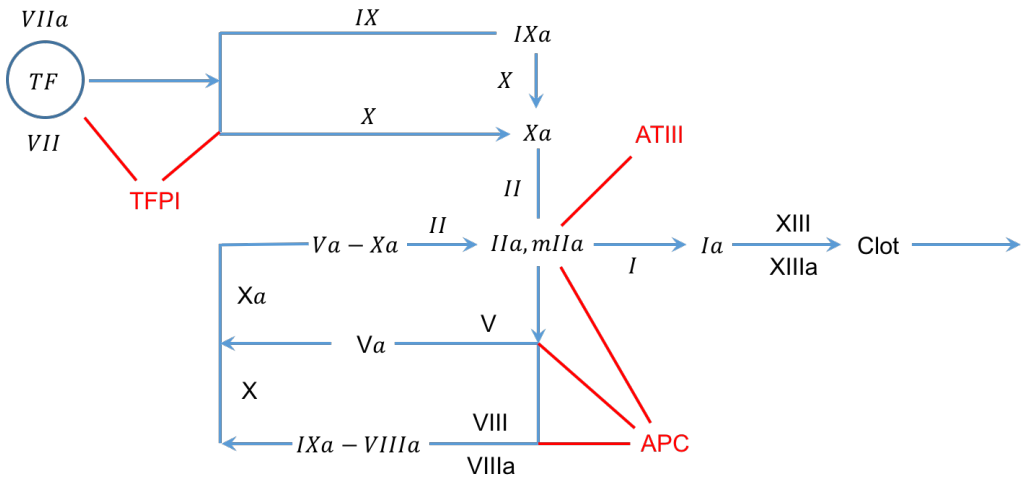

Figure 1.4: A schematic of the extrinsic pathway. We restrict our study to the class of models that focus on the extrinsic pathway. There is clot initiation due to Tf. Then the reactions proceed to the propagation phase where thrombin reaction rates explode is in a positive feedback loop due to tenase and prothrombinase activation. Thrombin catalyzes fibrin formation which is further stabilized into clot.

[51, 52], acute cerebrovascular disease [53], and rheumatoid arthritis [54]. Further, clot properties are known to be affected in such hypercoagulative systems [55].

1.6 Challenges in Modeling the Coagulation System

Many challenges centered around the chemical kinetics aspect arise in modeling and simulating the system:

- **Need to identify sensitive risk factors**: There is need for better phenotyping of coagulation system [56]. Available assays (referred to as the 'coagulogram') fall short of effectively characterizing the status of blood chemistry [12]. Hypercoagulable diseases like venous thromboembolism do not have readily identifiable and sensitive risk factors [57].

- **Need to identify useful aspects and methods to validate**: It is not clear

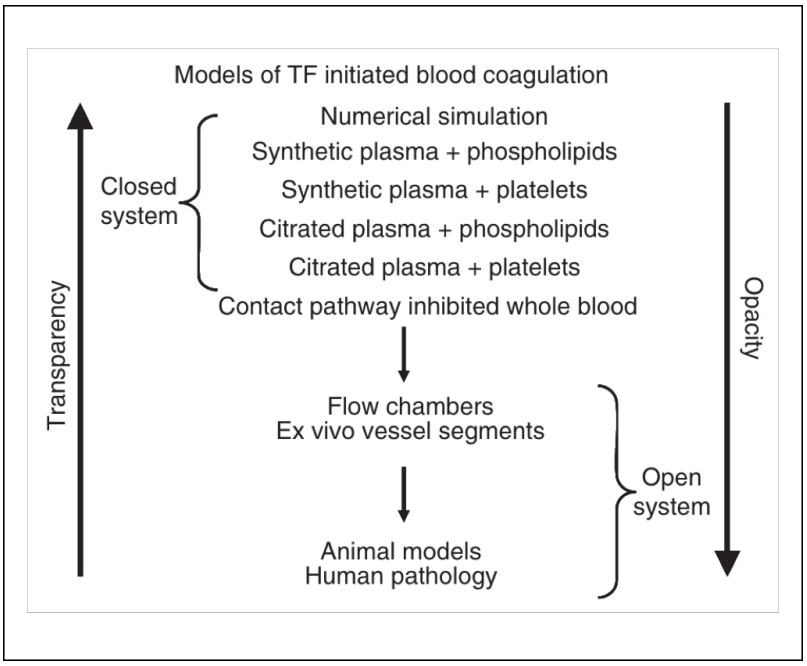

Figure 1.5: Models and transparency (figure from [58]). Simulations are considered to be more transparent compared to experiments in real systems. However, the opaqueness of real systems are often reflected as inadequacies in the simplified models. There is drastic room for improvement by establishing consistency between various systems. Simulations could be used to identify critical experiments. Iteratively, experiments could be used to correct our models and improve simulations.

which aspect of the simulation data is useful to make risk assessments even while using model results as a black-box. Obtaining insights is much harder due to a lack of established procedures to extract relevant aspects of the results for further analysis (see Figure 1.5).

There is uncertainty regarding the type of the blood used or viable for study (Platelet-rich plasma PRP, Platelet-poor plasma PPP, in vivo, in vitro, synthetic); our understanding of the various mechanisms involved, the functionality of different chemical species is still incomplete. The agreement between empirical results and model simulations even for thrombin generation curves is disputed (see Figure 1.6). For example, Hemker [59] mentions *'mostly used'* models ([60, 61, 62] do not match experimental curves of thrombin generation.

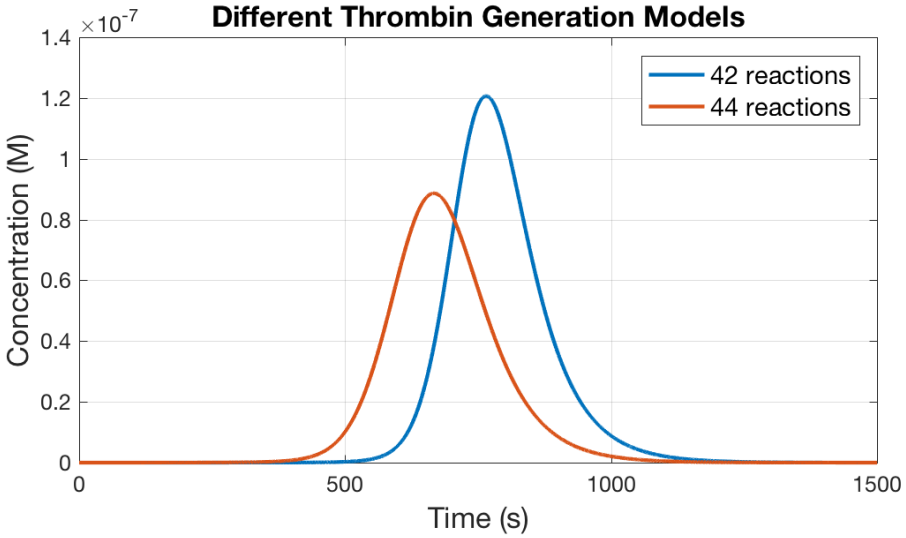

Figure 1.6: Two different thrombin generation models in the extrinsic pathway. Identifying critical aspects in model simulations based on clearly defined context such as disease classification would be a first step towards improving the models.

There is uncertainty in the measurement of various parameters like the rate

constants [63]. Some of the rate constants are not directly measurable. For example, many rate constants in the extrinsic pathway model were indirectly inferred [32, 60]. Further, there are at least two other versions even for the extrinsic pathway model [64, 65, 66]. This could be attributed to the fact that comparing model simulations with few experimental data does not suffice and the critical aspects of model simulations need to be clearly established.

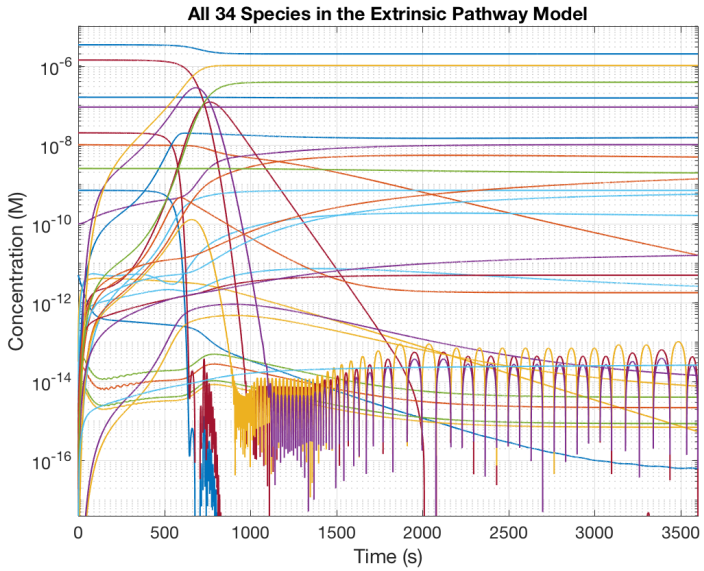

Figure 1.7: Concentration of different species during clotting in the extrinsic pathway. The concentration of various species varies by orders of magnitude. The poses stiffness issues in numerical schemes demanding small time steps and excessive conditioning.

- **Solving is hard and expensive**: Concentrations of protein factors and rates of reactions often vary by orders of magnitude (see Figure 1.7). The nonlinear chemical kinetics problem is modeled using reaction rates that have quadratic terms, the model is very stiff and solution trajectories are unstable in many

directions [67] (see Figure 1.8). The rates involve negative feedback loop or cycles in the reaction cascade [68].

Moreover, numerically solving stiff chemical kinetics is computationally expensive. The solvers for chemical kinetics are time and memory consuming when augmented with spatial and flow aspects. Coagulation is known to vary drastically in patients. A simplified model is desirable so that augmenting chemical kinetics in fluid flow solvers in patient-specific terms become feasible.

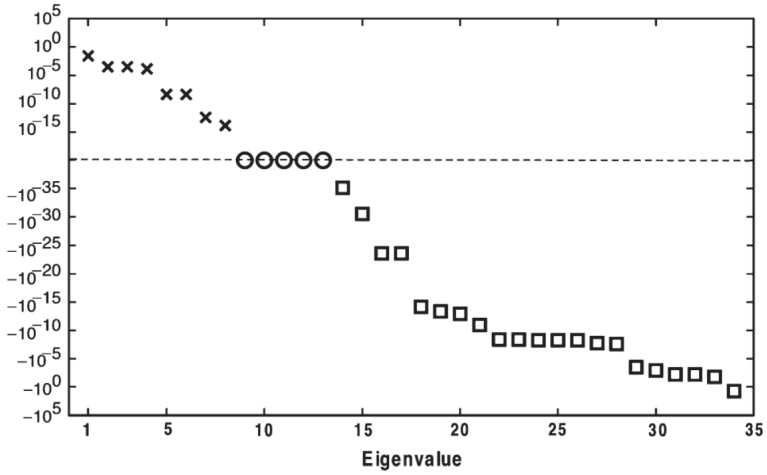

Figure 1.8: Eigenvalues of reaction rates in thrombin generation simulation. At any given point in time, the reaction trajectories are unstable in more than five directions (figure from [67]). This poses stability issues in numerical solutions.

Recent studies have tried to augment thrombin generation with convection-reaction systems [69, 70, 71]. In patient-specific studies of Papadopoulos et al. [71, 72], the focus is primarily on the effect of geometry. They use a simplified thrombin generation model to study the effect of vessel geometry. However, in order to understand the effect of reduced models, the abnormalities in coagulation system need to be

identified and quantified systematically.

We would like to have the following towards realistic patient-specific simulations:

- We would like to find critical features that are useful for diagnosis and that are possibly useful for understanding the complicated physiology behind coagulation in simple terms.

- We need a simplified chemical kinetics model that could be augmented with other models in order to perform useful patient-specific simulations.

The following question naturally arises: How do we simplify the model describing chemical kinetics? The simplified model,

- should be practically useful for making decisions.

- should be easier for patient-specific simulation and to augment with the simulations of other aspects like flow properties.

- should offer physiological insight.

We would like to understand and quantify the limitations of the reduced model. Further, given that the search space is huge and the experiments costly, we would like to identify potential candidates for further research. For example, based on a well defined purpose such as classification of diseases, we would like to quantify model performance.

We will make use of recent advances in machine learning algorithms and statistical learning theory to study the thrombin generation system.

1.7 Machine Learning

The core idea of machine learning is to use information about certain training samples to predict responses for new test samples (see Figure 1.9). The training and

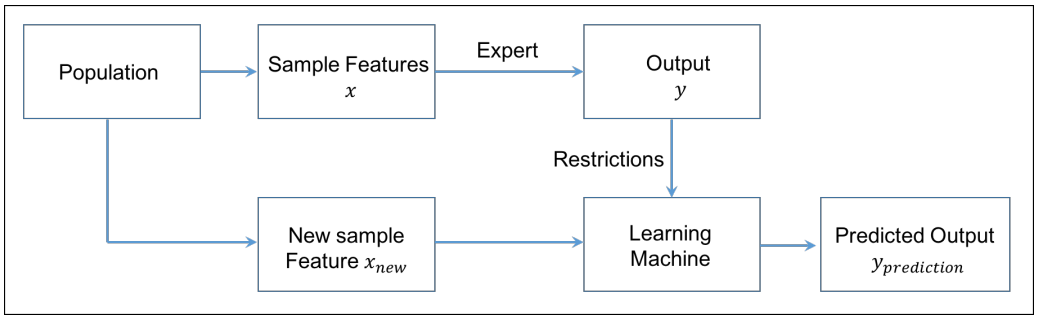

Figure 1.9: Schematic of the learning process. The core idea of machine learning is to use certain training samples to predict responses for new test samples.

the test samples are assumed to drawn from the same population. Machine learning tools are useful in systems where the involved physical laws are too complicated to model but data-driven predictions could be applied efficiently. Offering non self-evident solutions to ill-posed problems is the hallmark of learning theory [73, 74] and machine learning algorithms [75]. Learning theory and algorithms have been applied in a wide range of problems and fields such as bioinformatics, machine perception, medical diagnosis, economics, and social network analysis.

Our hypothesis is that it is possible to systematically use machine learning tools to obtain greater insight about the coagulation cascade. In particular, we will quantify and characterize useful information in high-dimensional data from solutions of equations involving nonlinear chemical kinetics model. We will do so by studying classification of ACS from CAD. We will use information on thrombin generation from a patient's blood sample to classify if a given patient has ACS. Further, we will use classification performance as a way to identify critical aspects of the model. Such critical differences could help us come up with simplified models.

1.8 Objective and Scope

Main algorithms and ideas used in this thesis include:

- Maximum entropy distributions for sampling patient data from population data.

- Density estimation using expectation maximization of Gaussian mixture models (GMM): In order to classify ACS using thrombin generation parameters, we need to model their densities. We model density of thrombin generation attributes using GMM. Finding GMM parameters is a hard nonlinear problem. Expectation maximization [76] offers a clever solution to the difficult nonlinear problem. The design of the algorithm and the guarantees of its convergence gives a flavor of ideas in machine learning. This is the starting point for many sophisticated applications as well as generalizations.

  We introduce the challenge of ill-posedness encountered in many learning problems via the example of parameter estimation in GMM. We briefly discuss how solutions are regularized in such a scenario.

- Classification using Random Forests: We use Random Forests [77, 78] to classify ACS from CAD using attributes from the full model. Random Forests exploits two key ideas to effectively deal with high-dimensional data: i) Use of ensembles of base learners, ii) and use of random subset selection. Both these aspects together avoid overfitting problems while dealing with sparse data. Moreover, the underlying base learner is nonparametric, and invariant to monotone transformations of data making it suitable to study chemical kinetics data. Further, Random Forests offer strong feature selection as well as error estimation tools.

We study blood coagulation using a model for the Tf-initiated extrinsic pathway developed by Hockin et al. [60]. Specific problems addressed in this dissertation include: i) simulation of patient-specific thrombin generation, ii) likelihood estimation

of thrombin generation parameters using expectation maximization, iii) classification of high-dimensional feature space using Random Forests, and iv) identification of critical aspects of the thrombin generation model.

Based on the insights gained from the results, **we also propose a simplified model for the dynamics of thrombin**.

1.9 Structure of the Dissertation

The structure of the dissertation is as follows:

1. Introduction in this chapter

2. We introduce the thrombin generation system. We describe patient-specific simulation and classification of thrombin generation.

3. We use expectation maximization algorithm and GMM to characterize and classify ACS and CAD based on summary parameters used to describe dynamics of thrombin [18, 45].

4. We extract features to characterize data from all the chemical factors in the model and use Random Forests to classify ACS and CAD [79]. We also perform feature selection in order to reduce the number of features used for classification.

5. We propose a simplified model for the dynamics of thrombin.

6. We discuss future research directions.

2. PATIENT-SPECIFIC SIMULATION OF THROMBIN GENERATION*

"These models are not a panacea nor are they a replacement for empirical fishing, but they are a useful thinking tool."

– Kenneth G. Mann, [58]

2.1 Chapter Outline

In this chapter, we describe the thrombin generation system in greater detail. In addition to patient-specific variation of chemical factors in thrombin generation systems, we describe how these variations are associated with ACS and CAD. We, also, describe patient-specific simulations and sampling required data for simulations.

2.2 Thrombin Generation

A schematic of the major events involved in clotting is shown in Figure 2.1. Clot initiation is modeled by three pathways:

1. **Intrinsic pathway**: Clotting in the intrinsic pathway is initiated when blood comes into contact with an external surface like glass.

2. **Extrinsic pathway**: Clotting in the extrinsic pathway is initiated by Tf. Tf proteins are embedded in the vessel walls and are hypothesized to be exposed to flowing blood due to injury. This pathway is the major cause of in vivo clotting. Tf activates factor VII to factor VIIa. This is followed by formation

*Part of this chapter is reprinted with permission from "Random Forests Are Able to Identify Differences in Clotting Dynamics from Kinetic Models of Thrombin Generation" by Jayavel Arumugam, Satish T. S. Bukkapatnam, Krishna R. Narayanan, and Arun R. Srinivasa. PloS one, e0153776, Copyright [2016] by Arumugam et al.

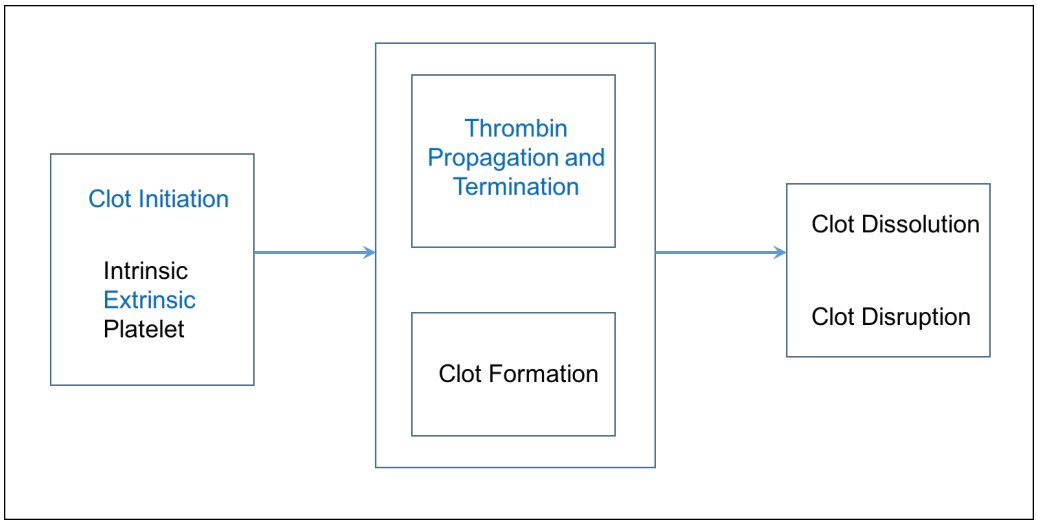

Figure 2.1: Major steps involved in blood clotting. We consider clot initiation in the extrinsic pathway along with thrombin propagation and termination.

of extrinsic tenase complex (Tf-fVIIa). The complex activates factors IX and X to form activated factors IXa and Xa.

3. **Platelets**: Clotting is also activated by platelets. In hypercoagulative blood, spontaneous clotting could occur due to platelet aggregation.

Tissue Factor Pathway Inhibitor (TFPI) regulates clot initiation. Activated factor Xa activates prothrombin to form thrombin. If thrombin concentration is above a certain threshold, clotting propagates via a different set of reactions [80]. This control mechanism likely functions to avoid excessive clot due to false triggers. Thrombin activates inactive factors VIII and V. This results in the formation of intrinsic tenase and prothrombinase complexes which further activate thrombin. Thrombin catalyses formation of Ia. Ia is stabilized into stable clot by activated XIIIa. Thrombin and other active factors are inhibited by ATIII and APC.

A schematic of the Tf-initiated extrinsic pathway model we used is shown in Figure 2.2. The model does not account for the effect of the inhibitor APC. We

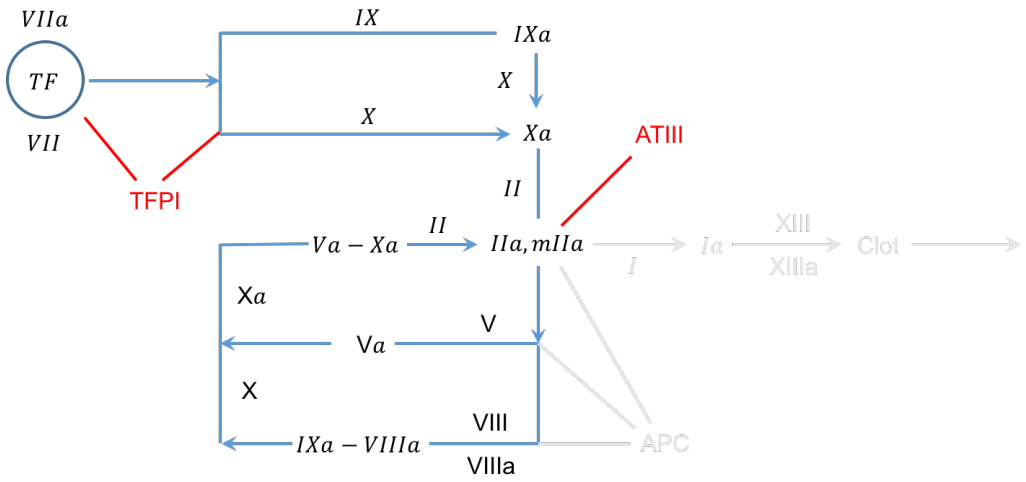

Figure 2.2: Schematic of the extrinsic pathway model for thrombin dynamics used in this work. This is the simplest nontrivial patient-specific model that has demonstrated practical use.

also note that there are at least two other versions of this extrinsic pathway model [64, 65, 66]. We use the simplest nontrivial version for this study [60]. There are 34 species undergoing 42 reactions in this model.

Thrombin generation upon tissue factor initiation is usually monitored using activity of thrombin and thrombin-antithrombin (TAT) activity. The model accounts for dynamics of two forms for thrombin (refer to Figure 2.3). One form is more active compared to the other and net thrombin activity is appropriately defined.

2.3 Patient-Specific Simulation

Thrombin generation varies from one patient to the other. In addition, there is often considerable variation within a patient. The source of the variation is modeled as changes due to plasma factor composition, i.e., the change in concentration levels of the inactive protein factors in the blood before clot initiation.

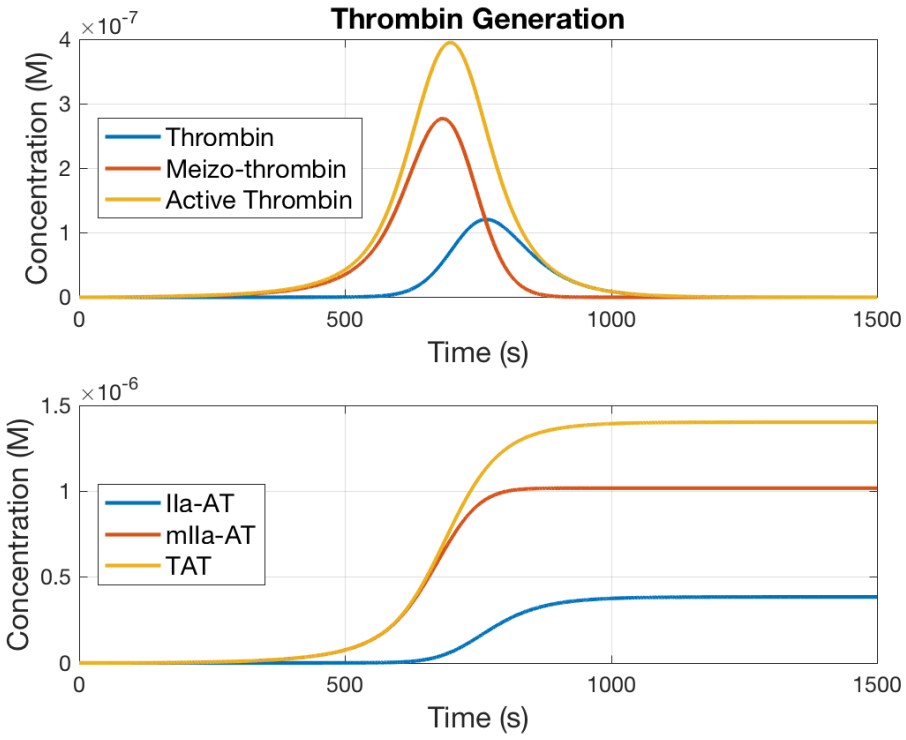

Figure 2.3: Two forms of thrombin are modeled. Active thrombin is defined based on activity of thrombin measurement (which is IIa + 1.2 mIIa). IIa is the alpha-thrombin and mIIa is the meizo-thrombin. Thrombin is inhibited by antithrombin ATIII. Thrombin-ATIII (TAT) complex is often measured to infer clotting properties. It is essential to account for the two forms in order to satisfy stoichiometry in simplified models.

Table 2.1: Physiological mean plasma factor composition. VIIa is set at 1 % of factor VII. Data reported in [60, 66].

Protein Factor	**Mean Value**	**Normal Range**
	(M)	(Percentage)
TF	5.0E-12, Varied	Controls trigger level
VII	1.0E-08	60 - 140
X	1.6E-07	60 - 140
IX	9.0E-08	69 - 151
II	1.4E-06	60 - 140
VIII	0.7E-09	64 - 232
V	2.0E-08	60 - 140
TFPI	2.5E-09	46 - 171
ATIII	3.4E-06	88 - 174

Mean physiological values of the eight initial inactive coagulation factors that are considered in this model are shown in Table 2.1. There is considerable variation of concentration of these factors in a given population. Thrombin generation has been used to phenotype such variations. For example, deficiency of factor VIII alone is not known to cause serious bleeding. Composite effect of all the protein factors determines the propensity of blood to clot. This is reflected in thrombin generation curves. Thrombin generation simulation curves in two Hemophilia A patients with factor VIII deficiency in Figure 2.4. Further, treatment is usually known to affect these results drastically.

We are concerned about thrombin generation in ACS and CAD patients. Simulations of thrombin generation is known to discriminate ACS and CAD [49]. When the simulations results are compared for the thrombin generation parameters, samples from the two groups differ in a statistically significant way. Thrombin generation is higher in ACS population suggesting hypercoagulation. Thrombin generation curves for the mean plasma factor composition in the two populations are shown in Figure 2.5.

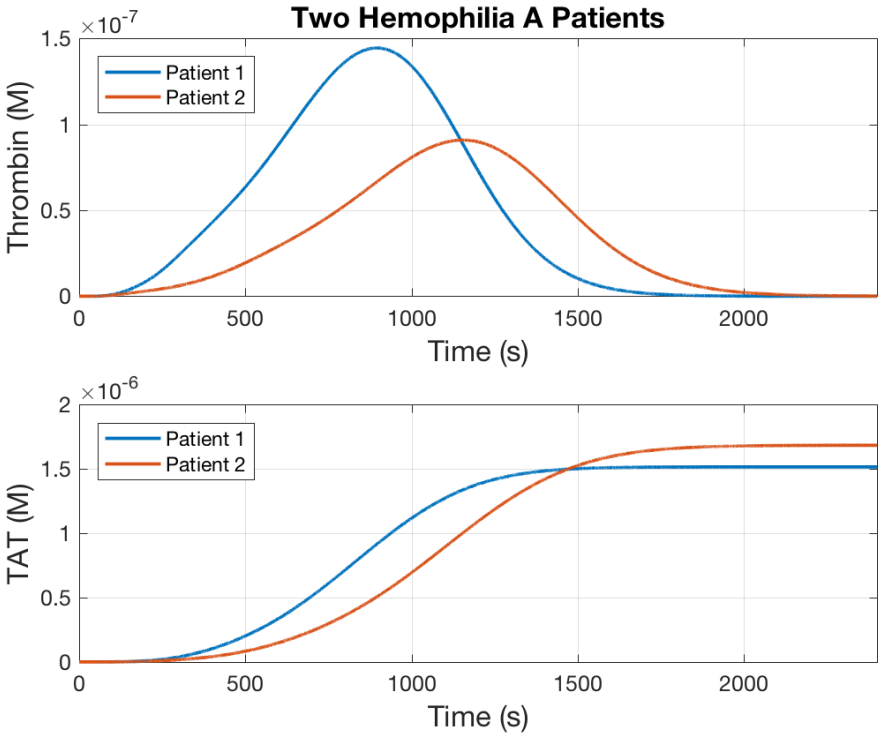

Figure 2.4: Thrombin generation in two hemophilia patients. Hemophilia A patients with factor VIII deficiency (set to 1 % of physiological mean in simulations). Though both the patients had initial factor VIII percentage value to be 1% of the physiological mean, the thrombin generation is significantly different due to other changes in the other inactive factors. This information has potential utility to monitor hemophilia treatment via recombinant factor VIII administration [81, 66].

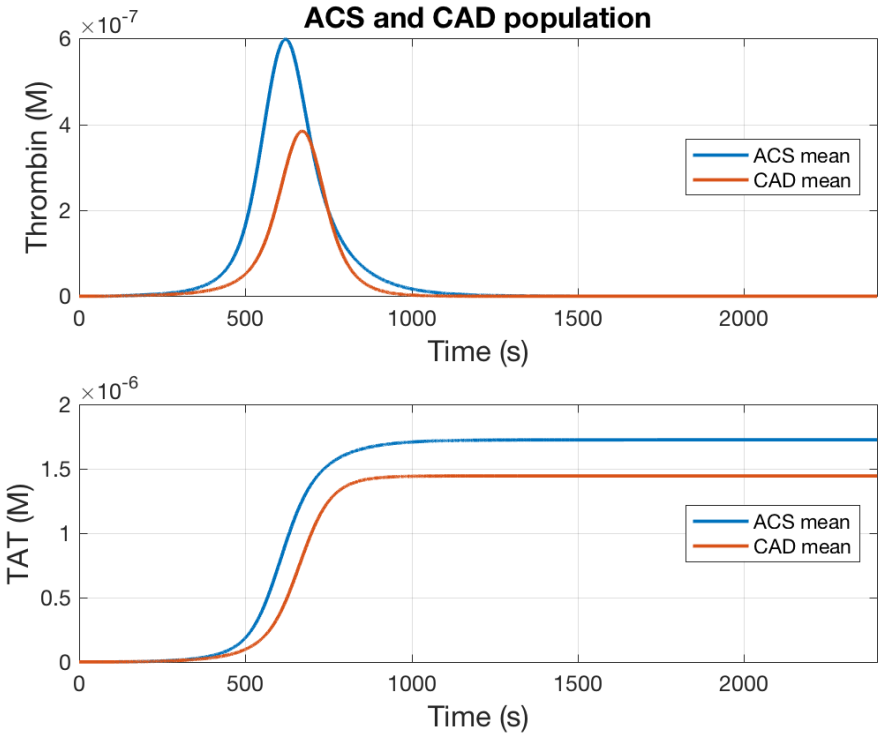

Figure 2.5: Thrombin in generation in ACS and CAD population mean. The net effect of changes in initial reaction is comprehensively captured in thrombin generation rates. Further, thrombin generation is higher in ACS population compared to CAD population. We will extract appropriate features to characterize the curves and describe such differences.

Most of the thrombin generation studies have been conducted to identify markers that differentiate groups in a statistically significant way. Given that, only the mean and standard deviation data for the groups are usually reported in these studies. In this work, we are interested in patient-specific classification instead of group comparison, i.e., we want to classify if a given patient likely belongs to one group compared to the other(s). Such simulations need patient-specific plasma factor composition. Using appropriate tools, we numerically sample such patient-specific data from reported population data. We describe our sampling procedure next.

2.4 Sampling Using Maximum Entropy Distributions

To get the initial condition data for the clotting model, we used reported mean and standard deviation data of the procoagulant and anticoagulant factor percentages in ACS and CAD populations [49]. To generate samples for these non-zero factors from the mean and standard deviation data, we use the maximum entropy principle [82]. The idea is to obtain a distribution that maximizes information entropy [83] subject to known constraints. Information entropy of a probability distribution $p(x)$ of a random variable x is defined as,

$$H = -\int p(x)log(p(x)). \tag{2.1}$$

The principle essentially restricts the class of probability distributions to those satisfying the given constraints by looking for functions that maximize,

$$\hat{p}(x) = \underset{p(x)}{\operatorname{argmin}} \; -\int p(x)log(p(x)), \tag{2.2}$$

$$s.t. \quad f_i(p(x)) = 0, \quad i = 1, \ldots, N. \tag{2.3}$$

where $f_i(p(x))$ are the known constraints.

When the constraints are moments of the probability distribution function, i.e., $f_i(p(x)) = E[g_i(x)]$ where $g_i(x)$ is an arbitrary function and $E[.]$ is the expectation, the solution could be expressed as,

$$\hat{p}(x) = \frac{1}{Z} \exp[\sum \lambda_i g_i(x)], \tag{2.4}$$

$$Z = \int \exp[\sum \lambda_i g_i(x)], \tag{2.5}$$

where Z is the partition function and λ_i are Lagrange multipliers which are found based on the given constraints.

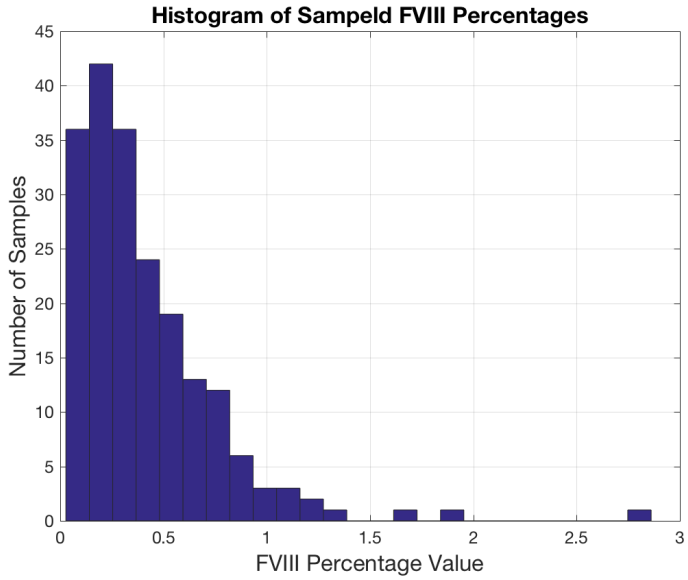

Figure 2.6: Sampled factor VIII values in hemophilia patients. Significance of using lognormal distribution of sampling initial factors.

By the maximum entropy principle, the probability distribution that best represents a positive random variable given mean and standard deviation is the log-normal

distribution [84]. Essentially mean, standard deviation, and nonnegativity are the three prior constraints we imposed based on the data available for plasma factor composition. The samples for each initial factor level were generated from log-normal distributions.

Using log-normal ensures that the sampled values are positive, for example, initial factor percentages sampled for hemophilia patients are shown in Figure 2.6. Also, unlike the symmetric Gaussian distribution, the lognormal distribution is skewed toward zero.

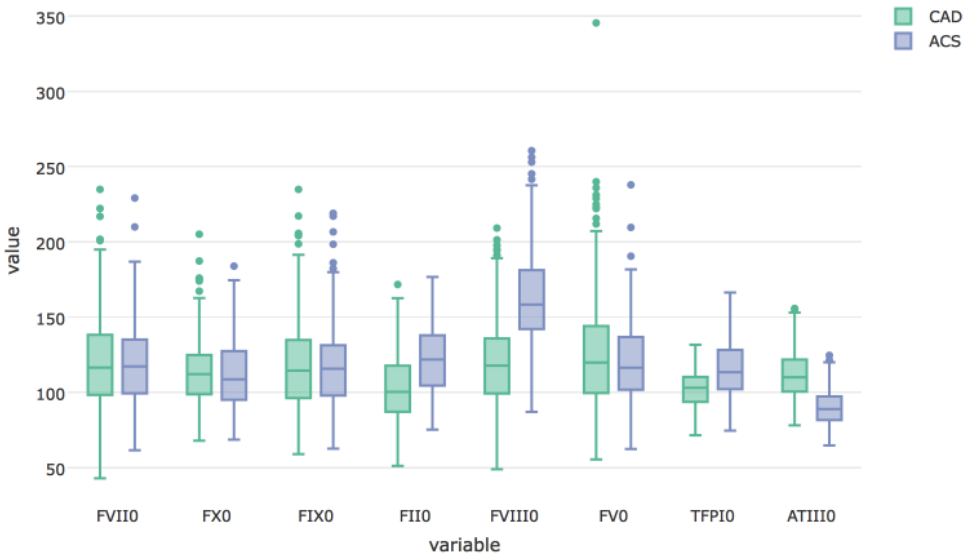

Figure 2.7: Plasma factor composition in ACS and CAD patients. Variation of sampled initial factor levels in ACS and CAD (data obtained from [49]).

We sampled 200 sets of percentage values for the initial coagulation factors in each class (ACS and CAD). Box plots for sampled data for ACS-CAD population are shown in Figure 2.7. Compared to CAD data prothrombin (FII), factor VIII, and

TFPI are higher and ATIII is lower in ACS data (see Figure 2.8 for scatter plots). These percentage values were scaled using the physiological mean values [60] and we obtained thrombin generation parameters by solving the chemical kinetics problem for each sample.

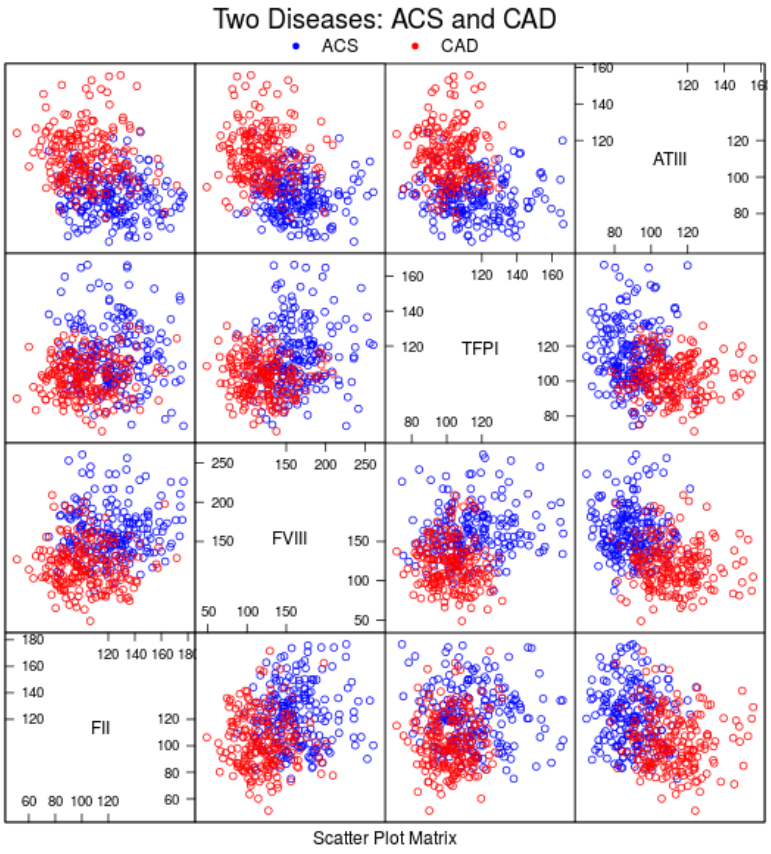

Figure 2.8: Sampled data of plasma factor composition in ACS and CAD. Only those factors that significantly differ in the two groups are shown. Note that the data are almost isotropic because our sampling assumes that the factors are independent of each other.

2.5 Piecewise Polynomial Representation of Simulation Profiles

Simulations were carried out using MATLAB[1]. Thrombin generation simulations were initiated with 5 pM trigger TF. Solution profiles for all chemical species were obtained for 3600 seconds using *'ode15s'* stiff solver available in the standard library. The method finds solutions to the differential equations $y' = F(t, y)$ by approximating it using a numerical differentiation formula of the following form [85],

$$(1-\kappa)\gamma_k \left(y_{n+1} - y_{n+1}^{(0)}\right) + \sum_{m=1}^{k} \gamma_m \nabla^m y_n - hF(t_{n+1}, y_{n+1}) = 0. \tag{2.6}$$

The implicit approximation is solved using a simplified Newton method [85]. The absolute tolerance for the method was set as 1e-15 M for all variables in the model and numerical convergence was corroborated (see Figure 2.9).

We normalized the simulation profiles by their respective physiological mean peak values and fit them with piecewise cubic hermite interpolating polynomials (PCHIP) [86] using *'pchip'* function. Data in each profile was divided into pieces (time intervals), and a cubic polynomial was fit in each piece while ensuring smoothness across pieces. PCHIP technique ensured the resulting interpolation changed monotonically in each interval, thereby avoiding spurious oscillations inherent in a regular spline interpolation.

Approximation using 14 pieces and using regular spline interpolation are shown in Figure 2.10. Even in the context of interpolation, we can notice one of the models is too complex and starts overfitting. By ensuring monotone changes between knots or the approximation points, PCHIP essentially looks for a restricted class of solutions compared to cubic spline approximation. This problem is more difficult in case of

[1]MATLAB 8.5.0, The MathWorks, Inc., Natick, Massachusetts, United States.

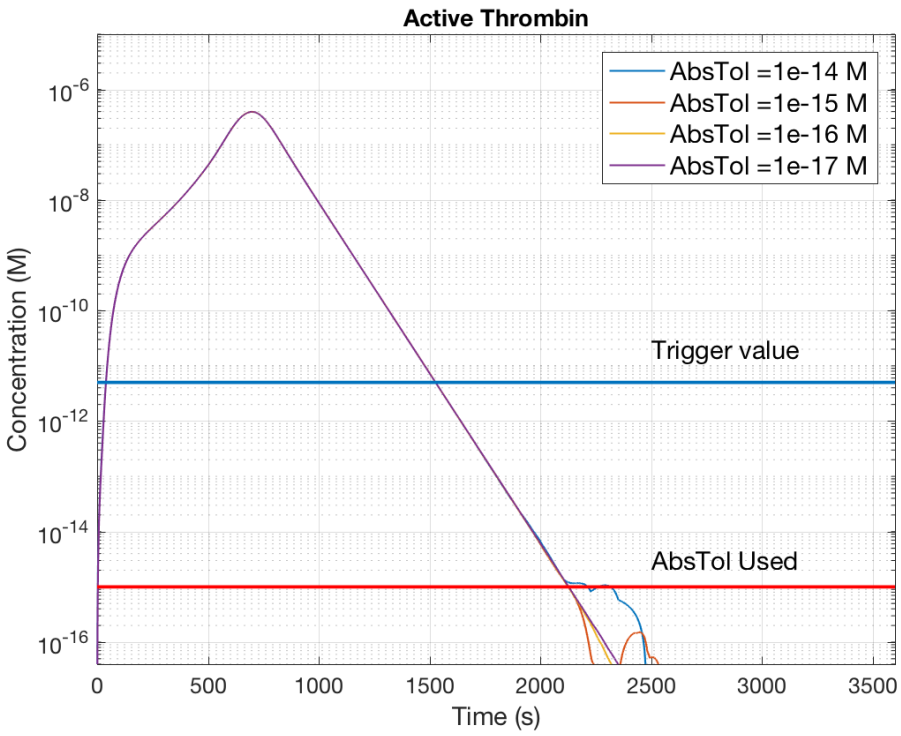

Figure 2.9: Convergence check for the numerical solution. Solution profiles for active thrombin obtained using different values of absolute tolerance in the numerical solution scheme. We used an absolute error tolerance of 1e-15 M for the simulations. Differences and oscillations below the error tolerance were neglected.

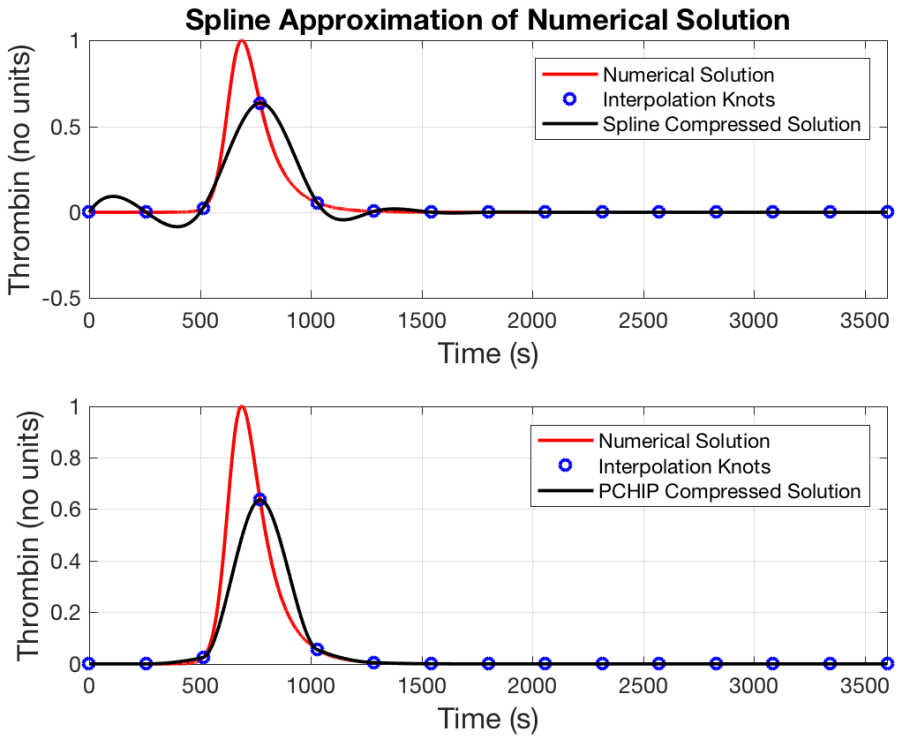

Figure 2.10: Piecewise polynomial representation of simulation data. By ensuring monotone changes between knots or the approximation points, PCHIP essentially looks for a restricted class of solutions compared to cubic spline approximation.

regression where we will have to account for data noise. One has to restrict the solutions by compromising between the available sparse and noisy data and model complexity.

We used 139 pieces - each of length approximately 26 seconds. This captured fast changes, such as the time it takes for Tf-fVIIa to reach its first peak since addition of the trigger, reasonably well. PCHIP representation serves two purposes: i) it efficiently stores large amounts of simulation data; and ii) since the polynomials represent data well, the coefficients of the polynomials could act as features for classification.

2.6 Conclusion

We have addressed the problem of patient-specific sampling using maximum entropy distributions. We assumed that the data were independent of each other. Interdependence of factors in thrombin generation has often been overlooked in the literature. If not posed and solved properly, this is a hard problem. While introducing patient-specific classification in the next chapter, we also address the problem of describing and inferring such dependence without the independence assumption.

3. DENSITY ESTIMATION USING EXPECTATION MAXIMIZATION

"I have had my results for a long time: but I do not yet know how I am to arrive at them."

– Johann Carl Friedrich Gauss[1]

3.1 Chapter Outline

We use expectation maximization of GMM to characterize thrombin generation parameters. We use it to classify ACS and CAD. We also discuss alternative approaches to solve these problems using non-self evident restrictions.

3.2 Simulation of Thrombin Generation Summary Parameters

Brummel-Ziedens et al. [49] studied alterations in thrombin dynamics between ACS and CAD. Features of thrombin profile like maximum value, area under the curve, and maximum rate were higher in ACS than CAD, suggesting hyper-coagulability. The question we address is the following: using simulation results of the thrombin generation curves, can we find the probability that a new curve is from an individual having ACS instead of CAD? Effective answer to this question could be used to screen patients for better monitoring.

We extracted the following features that characterize active thrombin [49] (see Figure 3.1):

1. time to reach 2 nM
2. area under the curve
3. maximum level reached

[1]quoted by A. Arber in 'The Mind and the Eye' 1954

4. maximum rate of generation

5. time to reach maximum value

6. time to reach maximum rate.

Time to reach 2 nM of thrombin in the extrinsic pathway is related to prothrombin time; area under the curve is related to the thrombin generation potential [87]. Further, thrombin generation measurements could be used to obtain the other features.

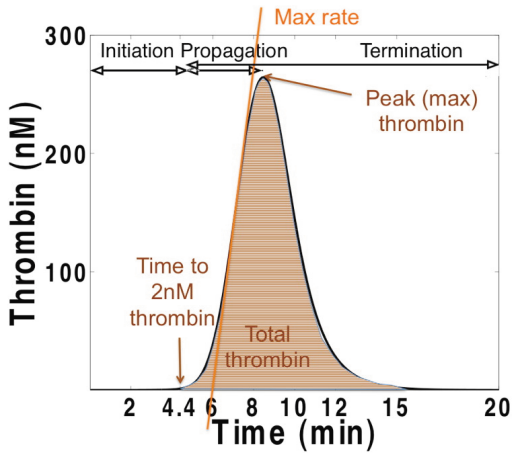

Figure 3.1: Thrombin generation summary parameters. Figure from [66]. Time 2 nM of thrombin and area under the curve are related to prothrombin time and endogenous thrombin potential respectively.

To better explain the issues with density estimation and to introduce more advanced algorithms, we consider just two of the summary parameters namely 'T2nM' (time to reach 2 nM from the start of the simulation) and 'Maximum Rate' (maximum rate of activation in active thrombin). The data was nondimensionalized using

the maximum values from the ACS population. Nondimensionalized thrombin generation summary parameters are shown in Figure 3.2. We proceed with building a classifier to distinguish data points with regards to different disease conditions.

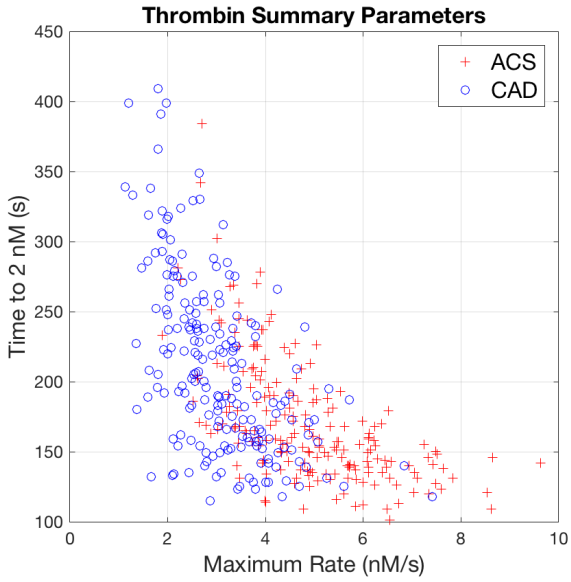

Figure 3.2: Maximum rate and time to 2 nM in thrombin generation. Summary parameters of thrombin generation for ACS and CAD population. Time to reach 2 nM of thrombin is faster in ACS population compared to CAD population. Similarly, maximum rate of thrombin generation is higher in ACS population. This suggests tendency of blood to clot more in ACS population.

3.3 Gaussian Mixture Models

To answer the above question, we classify given data from thrombin generation simulation into ACS or CAD group using Gaussian Mixture Model (GMM). Let

$$x = [\text{T2nM}, \text{MaxRate}]^T \tag{3.1}$$

denote the vector representing thrombin generation parameters. We consider soft classification where each data point has probability of belonging to either class. Given two classes G_1 and G_2 which are mutually exclusive, the posterior probability density of a group is,

$$P(G_i|x) = \frac{P(G_i)P(x|G_i)}{P(G_1)P(x|G_1) + P(G_2)P(x|G_2)}, \quad i = 1, 2. \tag{3.2}$$

where $P(G_i|\mathrm{x})$ is the posterior class probabilities for a given data point, $p(G_i)$ is the prior probability of the classes, and $P(\mathrm{x}|G_i)$ is the likelihood of the data conditional on the class.

If we can estimate $P(G_1|x)$, we can find the posterior using a suitable prior. What we know is only samples of data from the two classes. We construct a GMM for each class using the known samples of data. The likelihood function for a group $P(x|G)$ is approximated using a multivariate GMM,

$$p(x|\mathrm{ACS}) = \sum_{i=1}^{K} \alpha_i^{\mathrm{ACS}} \mathcal{N}(\mu_i^{\mathrm{ACS}}, \Sigma_i^{\mathrm{ACS}}) \tag{3.3}$$

$$p(x|\mathrm{CAD}) = \sum_{i=1}^{K} \alpha_i^{\mathrm{CAD}} \mathcal{N}(\mu_i^{\mathrm{CAD}}, \Sigma_i^{\mathrm{CAD}}) \tag{3.4}$$

where K is the number of components in the GMM and $\mathcal{N}(\mu_i, \Sigma_i)$ is multivariate normal distribution with mean μ_i, covariance Σ_i, and α_i is the probability of each component. An example is shown in Figure 3.3.

3.4 Expectation Maximization

The problem of approximating the likelihood function reduces to finding the parameters $\theta = \{\alpha_i, \mu_i, \Sigma_i\}$ of the GMM that explain the data the best. A standard way

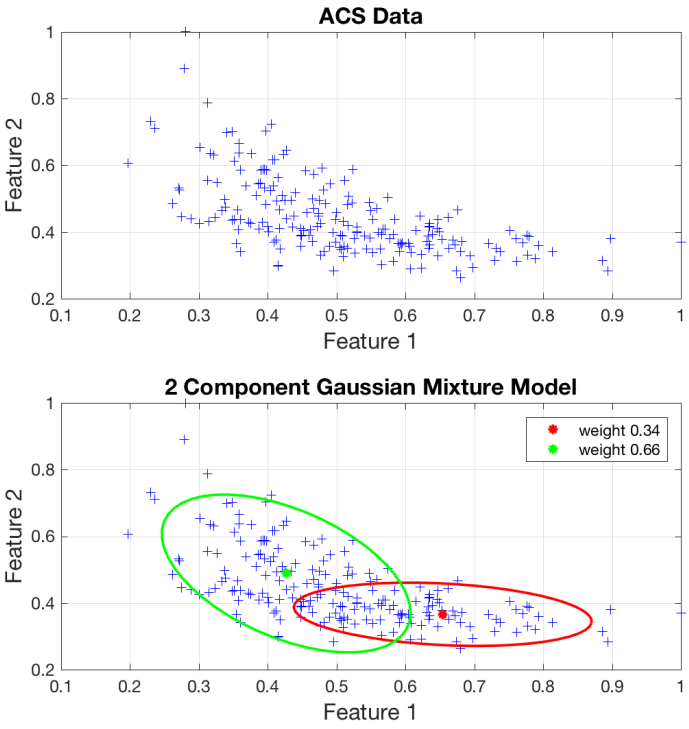

Figure 3.3: An example of a 2 component GMM.

to pick the parameters is by maximizing the likelihood function (or equivalently the log of it) called the maximum likelihood estimate (MLE). This is a simple problem if we know how each data point is generated, i.e., information pertaining to component identity. However, both this information as well as the parameters of GMM are not available. This makes the resulting likelihood function non-convex and a closed form MLE for this function is not known. In such a scenario, an efficient way to find the parameters is by using the expectation maximization algorithm [88] and [76].

Expectation Maximization (EM) algorithm finds the solution iteratively by approximating the likelihood function by a convex function which is a tight lower bound and then maximizing it. Let y be the incomplete data which denotes component identity for each data point. The algorithm iterates between two steps:

1. **Expectation step**: Find expectation of the log-likelihood $Q(\theta, \theta_{i-1})$ using the parameters θ_{i-1} from the previous step,

$$Q(\theta, \theta_{i-1}) = E[\log(p(x, y|\theta))|y, \theta_{i-1}] \tag{3.5}$$

2. **Maximization step**: Find new value θ_i for the mixture parameters by maximizing the above likelihood.

$$\theta_i = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta_{i-1}) \tag{3.6}$$

3.5 Classification Results

The results for likelihood estimation are shown in Figure 3.4 for ACS and CAD respectively. We use non-informative priors for class probabilities, i.e., the prior probability of each class is the same ($p(G_1) = p(G_2) = 0.5$). We consider the

posterior probabilities of the 400 samples used to train the GMM. The samples are ordered such that the first 200 are from the ACS class and the last 200 belong to the CAD class.

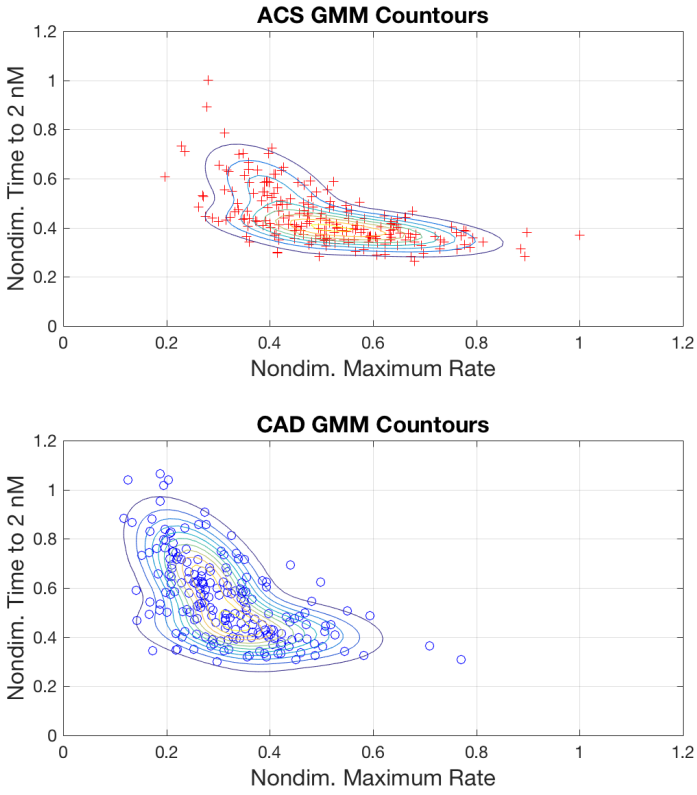

Figure 3.4: Contours of likelihoods for the thrombin generation parameters. GMM contours of the likelihood function for the thrombin generation parameters estimated using the EM algorithm.

The predicted probabilities that a sample belongs to the ACS class and CAD class is shown in Figure 3.5. The predicted probabilities that the sample belongs to ACS are high for the first 200 points and low for the last 200 points. The test accuracy was calculated on 40 randomly sampled data from each class that was not

used for training. The mean classification accuracy was 77 %. Due to the use of sampled data for the plasma factor composition, we do not further distinguish errors in each class separately.

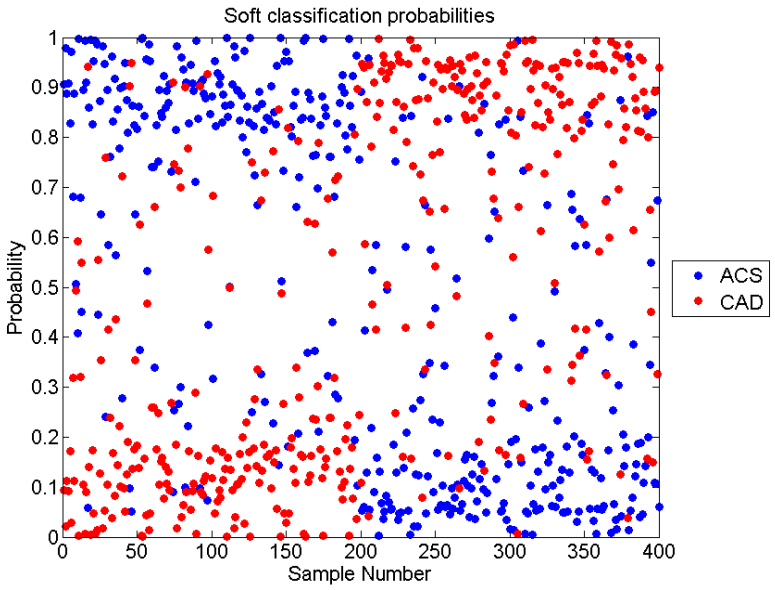

Figure 3.5: Predicted ACS/CAD probabilities. For each sample the predicted probability that it belongs to ACS (blue) or CAD (red) group is shown. The first 200 samples are from ACS. Samples 201 to 400 are from the CAD group. The first 200 samples are predicted to have a high probability of belonging to class ACS and CAD. The mean test accuracy was close to 77 %.

3.6 Paradigm Shift

Expectation maximization is a powerful algorithm. In mechanics, it is useful in the context of deformable surface tracking [89, 90] and discrete optimization formulations of certain hard mechanics problems such as buckling of beams constrained to a tube [91, 92].

GMMs face the curse of dimensionality if used to model features of all protein

factors using sparse data. For a full GMM, the number of parameters grow as $(K-1)+Kp+Kp(p+1)/2$ where K is the number of components and p is the dimension of the feature space [93]. This results in ill-conditioning of the parameter estimation problem. The problem is worse in case of nonparametric extensions such as Dirichlet process mixture models [94] which does not restrict the number the mixture components. Notice that this is not a problem with the EM algorithm but with the M-Step of the EM algorithm where the parameters are estimated.

Density estimation is not a necessary step in order to make classification. Statistical learning has made tremendous advancements using this idea (see Figure 3.6). Parameter estimation in high-dimensional problems is often ill-conditioned. This demands better algorithms and approaches. Non self-evident restrictions and solutions to such ill-posed problems is the hallmark of machine learning and statistical learning theory. Some of the approaches include using ensembles [77, 95], exploiting sparsity of data [96], efficient use of sampling [97], inference [98], subsets and sparse data [99, 77], better and efficient models for the covariance structures [100].

> "If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem."
>
> – Vladimir N. Vapnik, Statistical Learning Theory [73].

3.7 Conclusion

We used GMM to estimate densities of thrombin generation summary parameters. The approach described in this chapter can be used to systematically classify coagu-

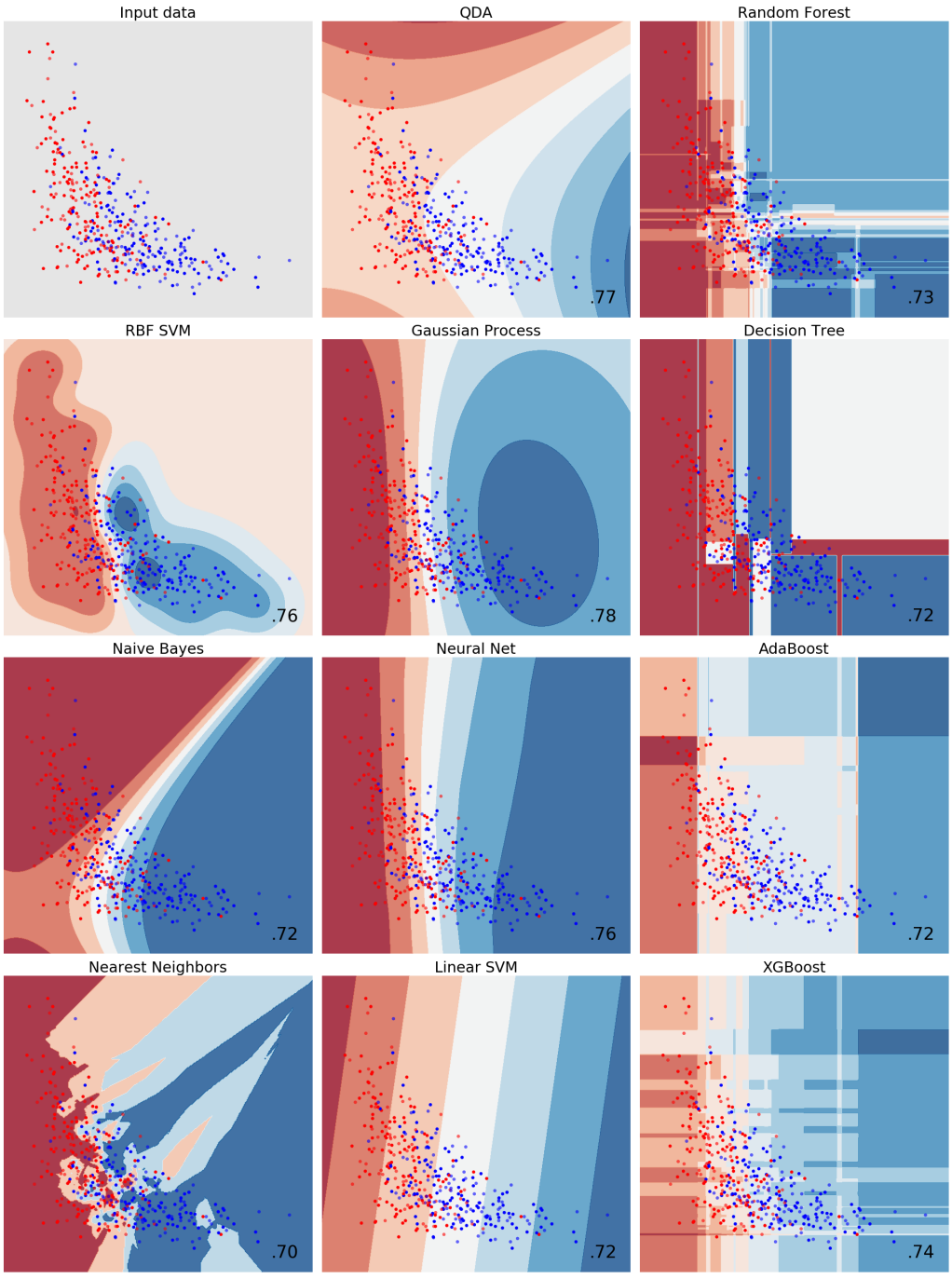

Figure 3.6: Different classifiers, mean test accuracies, and their decision boundaries. Most of these methods demand lots of parameter tuning. Further issues include model selection and significant feature selection.

lation disorders using data from thrombin generation parameters [45], plasma factor composition [49], and thromboelastography [101]. Estimation of densities of plasma factor composition could be one the most important applications of the methods used in this chapter.

We, also, discussed how classification in higher dimensions demands better strategy and regularization is one of them. We choose Random Forests to proceed further with the high-dimensional classification problem.

4. HIGH-DIMENSIONAL CLASSIFICATION PROBLEM*

"The limits of my language mean the limits of my world."

– Ludwig Wittgenstein, *Tractatus Logico-Philosophicus (1922)*

4.1 Chapter Outline

Random Forests is a nonparametric classification method that stands up to its name. Perhaps, a lot more. Unlike GMMs, Random Forests are effective in dealing with high dimensional data. Further, we use Random Forests to classify and to find significant aspects in the thrombin reaction network. In particular, we find significant chemical species and their location in time during clotting that could be useful for classification.

4.2 Introduction

Current efforts towards patient-specific characterization include differentiating systemic changes to blood coagulation in ACS from CAD populations [50]. Blood is observed in a hyper-coagulable state after ACS [102]. Brummel-Ziedens et al. [49] studied alterations in thrombin dynamics between ACS and CAD. Features of thrombin profile like maximum value, area under the curve, and maximum rate were higher in ACS than CAD, suggesting hypercoagulability.

The nature or extent of the hypercoagulability[1], as well as its relation to and its presence before the acute condition are not well understood. This could be attributed

[1] A recently published review article on this [103].

*Part of this chapter is reprinted with permission from "Random Forests Are Able to Identify Differences in Clotting Dynamics from Kinetic Models of Thrombin Generation" by Jayavel Arumugam, Satish T. S. Bukkapatnam, Krishna R. Narayanan, and Arun R. Srinivasa. PloS one, e0153776, Copyright [2016] by Arumugam et al.

to at least two reasons: i) lack of assays to efficiently and effectively determine the status of blood chemistry [12]; and ii) lack of adequate statistical and mathematical tools to understand blood coagulation system involving large numbers of variables.

Recently there have been attempts to study changes in factor Xa (fXa), in another hyper-coagulable condition - deep vein thrombosis, using computational models [104]. Features similar to those used for thrombin were used to describe fXa. We have good prior knowledge about thrombin and fXa, which are both active chemical species that play significant roles in clotting. Naturally, the following questions arise:

1. Do the dynamics of any other chemical species change significantly?
2. Are there better features to characterize changes in the system?
3. Can we efficiently assay the entire system without losing much information pertaining to classification?

We study blood coagulation using a model for the Tissue factor(Tf)-initiated extrinsic pathway developed by Hockin et al. [60]. The model uses a system of ordinary nonlinear differential equations to describe dynamics of thrombin evolution. The model has copious empirical validation and has been previously used for risk analyses between ACS/CAD [49]. The number of chemical species involved is large (34 in this case), and their responses are varied, typically requiring large numbers of features to represent the time profiles.

We use a non-parametric statistical learning algorithm - Random Forests [77] to classify ACS and CAD populations. Random Forests can be used to capture highly nonlinear class boundaries, and is robust to outliers in data and to lots of noisy features. Random Forests technique allows us to filter significant species and find their critical aspects. Moreover, unlike the current use of isolated features for group

comparisons prevalent in thrombin generation literature [45], use of Random Forests here exploits the role that interactions of features play in order to classify data into various groups.

4.3 Feature Extraction from Simulation Profiles

The central idea of the scheme is to consider the data points (simulation profiles) as a "noisy-image" in a very high-dimensional space from which we try to extract features with semantic attributes like "concentration is high", "concentration profile is sharply curved", etc. This was implemented by extracting different kinds of features and using them in the classification study to characterize the system.

In order to capture the dynamics of each species at different times during the simulation, we use the PCHIP coefficients as features. Since there is a lack of classification study to compare this work with, we used classification results of the plasma factor composition (initial conditions data used for the simulation), and the features that are conventionally studied to compare with the performance of PCHIP features considered here. Moreover, we study a fourth set of features which have the possibility of direct experimental observation.

The list of features we extracted and used for classification include the following four sets:

1. **PCHIP features to characterize dynamics** - this set includes 18904 PCHIP coefficients obtained during data representation. This set uses two datasets [49, 60] as described at the beginning of the section Methods. We used 139 pieces - each of length approximately 26 seconds (the representation is shown in Figure 4.1). For a given species, there are 4 coefficients in each time interval. The coefficients are such that the fit polynomial in an interval starting at t_i has the form $C_{i3}(t-t_i)^3 + C_{i2}(t-t_i)^2 + C_{i1}(t-t_i) + C_{i0}$. These coefficients

have information pertaining to function values and derivatives up to 3 orders at time t_i. Information in second and third derivatives is expected to be weak as PCHIP enforces monotonicity. Variables corresponding to the two forms of thrombin, IIa (alpha-thrombin) and mIIa (meizothrombin), were interpolated separately.

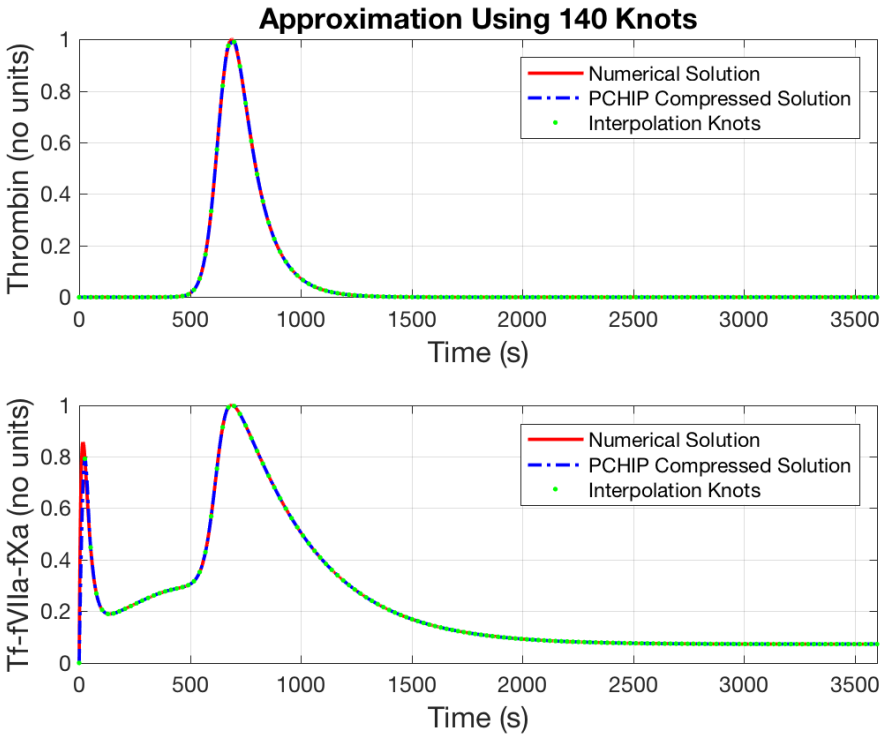

Figure 4.1: Spline representation of simulation data. The number of approximation knots were chosen to be 140 based on the response of one of the fastest reacting variables Tf-fVIIa-Xa.

2. **Plasma factor composition** - this set consists of 8 non-zero initial condition percentage values of procoagulant and anticoagulant factors used for model simulations [49]

3. **Conventional features** - this set consists of 11 features used to characterize active thrombin [49] and fXa profiles [104]. This includes time to reach 2 nM (for active thrombin), area under the curve for active thrombin and fXa, maximum level reached by active thrombin and fXa, maximum rate in active thrombin and fXa profiles, time to reach those maximum levels for active thrombin and fXa, and time to reach maximum rates for active thrombin and fXa. Data from two datasets [49, 60] are used in this set.

4. **Moving averages of concentration values** - this set consists of 200-second moving average (200s-MA) features extracted at uniform time intervals. For each chemical species, we extracted these 18 time-averaged function values of simulation profiles ($1/200 \int_t^{t+200} x(t)dt$, where $x(t)$ is concentration of a given species) at every 200 seconds starting at 100 seconds. 612 such features were extracted from all species (18 each for 34 species). This set makes use of two datasets [49, 60]. These features localize significance of each species in a time frame of about 3 minutes. Moreover, averaging over time gives a more robust feature with respect to time lags and noise imposed by model and model parameters.

These features are used as inputs in the Random Forests classification algorithm, which outputs group identity (ACS/CAD).

4.4 ACS/CAD Classification using Random Forests

The core objective of any classification method is to label a collection of data/measurements using certain features [105]. Here we use Random Forests [77] which is formed by aggregating an ensemble of decision trees [106].

4.5 Decision Tree

A decision tree [107, 108, 106] divides the feature space into a number of non-overlapping regions. The regions have an equivalent tree representation in which each node is a decision rule regarding class identity. Such trees are nonparameteric and assume no particular form of the data. The task of the tree algorithm is to frame decision rules that suit the data. Such decision rules are invariant to all monotone transformations in the data [106]. Once a tree is formed, data points with unknown classes are assigned a class based on these decision rules. Decision trees have been used in the study of thrombin generation systems [109].

However, a simple tree structure is sensitive to perturbations in the data [110] which could propagate down the tree and lead to very different class labels. The random forest technique [111, 77], which uses an ensemble of trees and aggregates the results, offers a solution to this problem.

4.6 Random Forests

In Random Forests, the learning process of each tree involves two types of random subset selection. First, each tree in the ensemble is built with a random subset of the training data. The other subset which is kept 'out' is called as the out-of-bag (OOB) samples. These OOB samples are used for finding internal estimates such as error rates. Second, each decision rule in a tree is made only using a random subset of all features. This avoids the classification results being unduly biased by a few sensitive features most of the time. Such classification results aggregated from many trees can capture complex and highly nonlinear class boundaries. It is well known [112] that the method avoids overfitting of the training data, a feature which is vital when there is limited or scarce data.

Random Forests methods are known to perform well in a variety of fields such

as in gene selection in microarray data [113], and in functional studies of chemical compounds [114]. In empirical studies, Random Forests compares well with other classification algorithms [115, 116], and performs consistently well in high-dimensions [117]. Use of Random Forests in clinical studies include study of blood proteins in Alzeimer's disease [118, 119].

A key feature of the Random Forests approach is their ability to provide reliable internal estimates to monitor error rates, and it has sharp measures to rank significance of features. In particular, we made use of OOB error rate and mean decrease in Gini index (MDGini) (see below). Since this error rate does not involve data used in training a given tree, using this error rate provides inherent cross validation [120].

- **OOB Error Rate:** OOB samples are used to find error rates for each tree in the ensemble, and all such error rates are averaged to get the OOB error rate. Empirical studies suggest OOB error rates are good estimates for generalization error [120, 77]. We used OOB error rates to assess the accuracy of the classifiers which are reported as percentages of (1.0 − OOB error rate).

- **Feature Significance Measure - MDGini:** Feature significance was interpreted using a Random Forest importance measure known as 'Mean Decrease in Gini index' [78]. Typically, the decision rules in the trees are not pure in the sense that the corresponding region in feature space is heterogeneous; i.e., there is a mix of data points from all classes (in our case 2). Gini index (or Gini impurity) [110] for a decision rule is a measure of this mix; it is zero only when the decision rule is perfect (the region is homogenous). It is maximum when the mix is the highest (half-and-half mix from both the classes).

  MDGini involves randomly permuting OOB sample data corresponding to the decision rule in a tree, and estimating the change (decrease) in Gini index. If

the decrease is high while perturbing a feature, it suggests that the classification is highly dependent on that particular feature. This provides an information-theoretic feature significance measure. It inherits the invariance property of the decision rules, i.e., absolute values of the features do not matter. This is a very sharp feature significance measure (see Figures 1, 2 and 6 in [121]). We use MDGini here to find even minute differences that are significant between ACS/CAD.

We used the 'randomForest' package in R [122] for our analysis. For each Random Forest classifier, 501 trees are used in the ensemble. To account for statistical variation between runs, we report mean and standard deviation (SD) of classification accuracies based on 50 runs.

4.7 Classification Performance of the Entire System

Classification using initial factors has a mean accuracy of 88.13% (Table 4.1). Conventional features of fXa and active thrombin classify with lower mean accuracies, 82.58% and 81.04%, respectively. Using all PCHIP coefficients and all 200s-MA values result in classification accuracies of 88.59% and 88.78% respectively, which are slightly better than using 8 initial factors. At this point, one might wonder if combinations of initial conditions suffice to characterize the system. However, we note that the same set of initial conditions could give different dynamics if the reaction network is perturbed (say, rate constants are changed due to a drug or a mutated form of a coagulation factor). Hence, studying initial conditions might not suffice to characterize the dynamics of the system. Moreover, studying the dynamics of chemical species gives more physiological insight about the underlying process.

Classification accuracies quantify the information in various features with respect to ACS/CAD classification. Although, minimum and maximum accuracies in Ta-

Table 4.1: Classification accuracies (%), mean (SD), of different sets of features. PCHIP and moving average features classify better than conventional parameters, and slightly better than all nonzero initial conditions. Every year $\sim$ 660,000 Americans have a coronary event [2]. A 7% improvement in classification accuracy suggests $\sim$ 46000 patients could be screened better every year in US.

Random Forest Classifier		**Mean (SD)**
PCHIP Features		
	All PCHIP Coefficients	88.59 (0.36)
Plasma Factor Composition		
	8 Initial Conditions	88.13 (0.49)
Conventional Features		
	fXa	82.58 (0.53)
	Active Thrombin	81.04 (0.46)
Moving Averages		
	All 200s-MA	88.78 (0.32)

fXa - factor Xa; PCHIP - piecewise cubic hermite interpolating polynomials; 200s-MA - 200-second moving average.

ble 4.1 vary over a small range ($\sim$ 81%-89%), they offer a potent way to compare features and quantify relevant differences. Loss of 11% accuracy in the initial conditions classifier is due to the overlap in the initial condition data used (though the means were significantly different for prothrombin, factor VIII, tissue factor pathway inhibitor, and antithrombin [49], the samples from lognormal distributions used for this study overlapped). Also, the best possible accuracy is restricted by the choice of features.

4.8 Selection of a List of Significant Species

We robustly selected a list of species that behave differently in ACS/CAD. We based our selection heuristics on three criteria and selected five species:

1. fXa and IIa were selected due to their known significance.

2. Tf-fVIIa-fXa, Tf-fVII-fX - these species had high averages for MDGini values

in the classifier built with all PCHIP coefficients. MDGini values for each species were sorted and the highest ∼10% of the values were used for selection criterion. Use of just one of the highest MDGini value for each species would be too biased and prone to noise; use of all MDGini values caused huge variation in the values, blurring out differences between species.

3. fIXa-fVIIIa-fX - this species had the highest significance during the last 600 seconds of the simulation in the classifier built with all PCHIP coefficients. Similar to selection criteria 2, selection was based on averages of highest ∼10% MDGini values. This criterion was used since the fate of such a chemical species is highly uncertain after the end time of simulation and calls for better scrutiny.

For criteria 2 and 3, MDGini values were obtained from the classifier built using all PCHIP coefficients as it had information pertaining to both function values as well as information about derivatives at a fine time scale. See Figures 4.2 and 4.3 for Box plots for these MDGini values.

The five filtered species were further studied by resolving their significance over time. MDGini for these species obtained using the classifier built with all 200s-MA values is shown in Figure 4.4. Tf-fVIIa-fXa and Tf-fVIIa-fX are most significant around 1200 seconds after clot initiation. Classification accuracies of these individual species using 200s-MA values are tabulated in Table 4.2. 200s-MA values of Tf-fVIIa-fXa, Tf-fVIIa-fX, IIa, and fIXa-fVIIIa-fX classify better than conventional features (Table 4.1).

A single feature from Tf-fVIIa-fXa classifies with accuracy 78.86%, which suggests that significance of Tf-fVIIa-fXa is best localized in time. This can be seen in Figure 4.4 as well as in Figure 4.5. Around 1200 seconds in Figure 4.5, the mean of the CAD group is outside the 90% quantile of the ACS group. Tf-fVIIa-fX visually

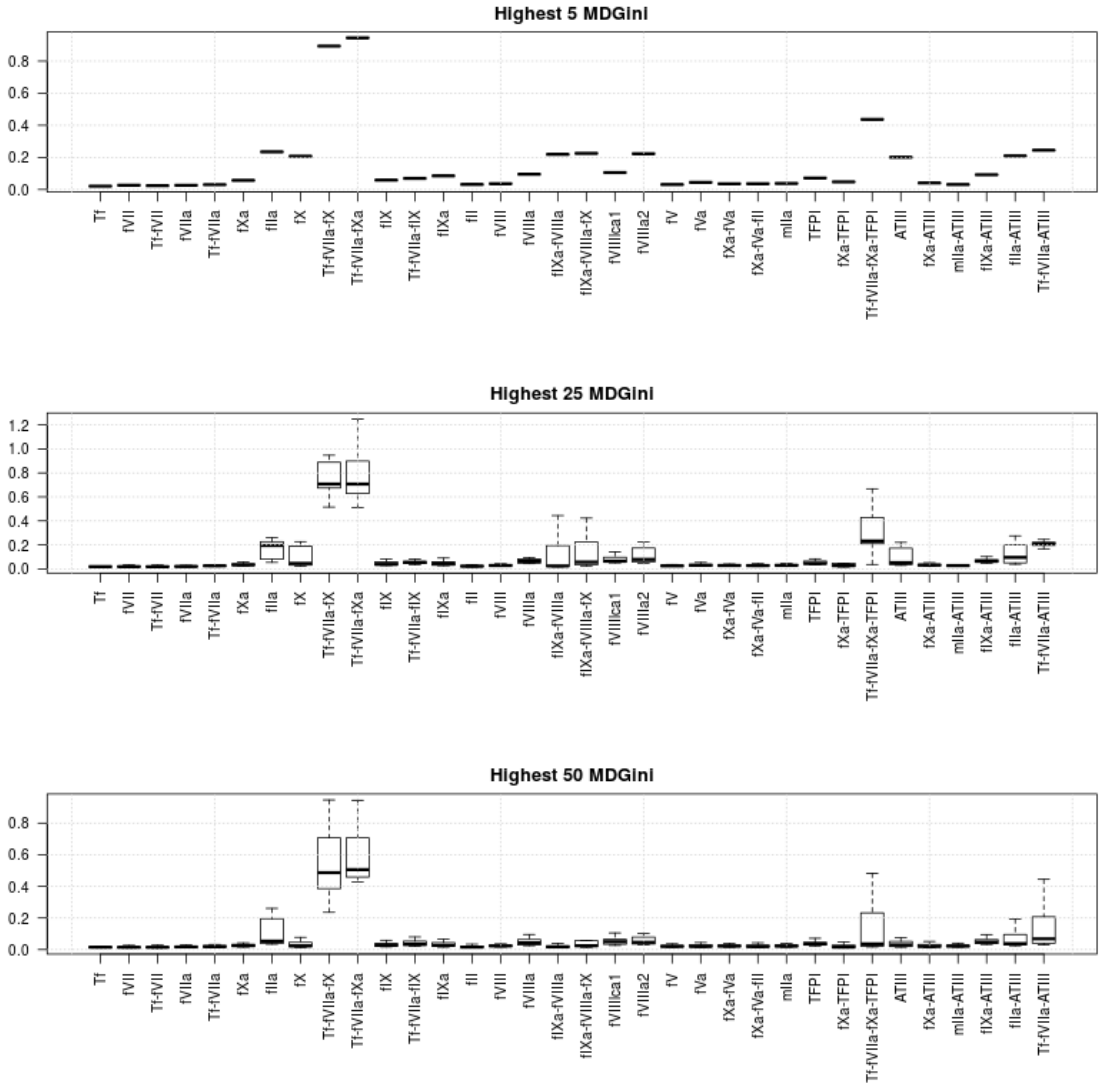

Figure 4.2: Feature significance during the entire simulation. Box plots of MDGini values for the PCHIP coefficients for each species. Tf-fVIIa-Xa and Tf-fVIIa-X stand out from the rest of the variables. MDGini values were obtained from the classifier built with all PCHIP coefficients so that their relative importance could be compared for filtering.

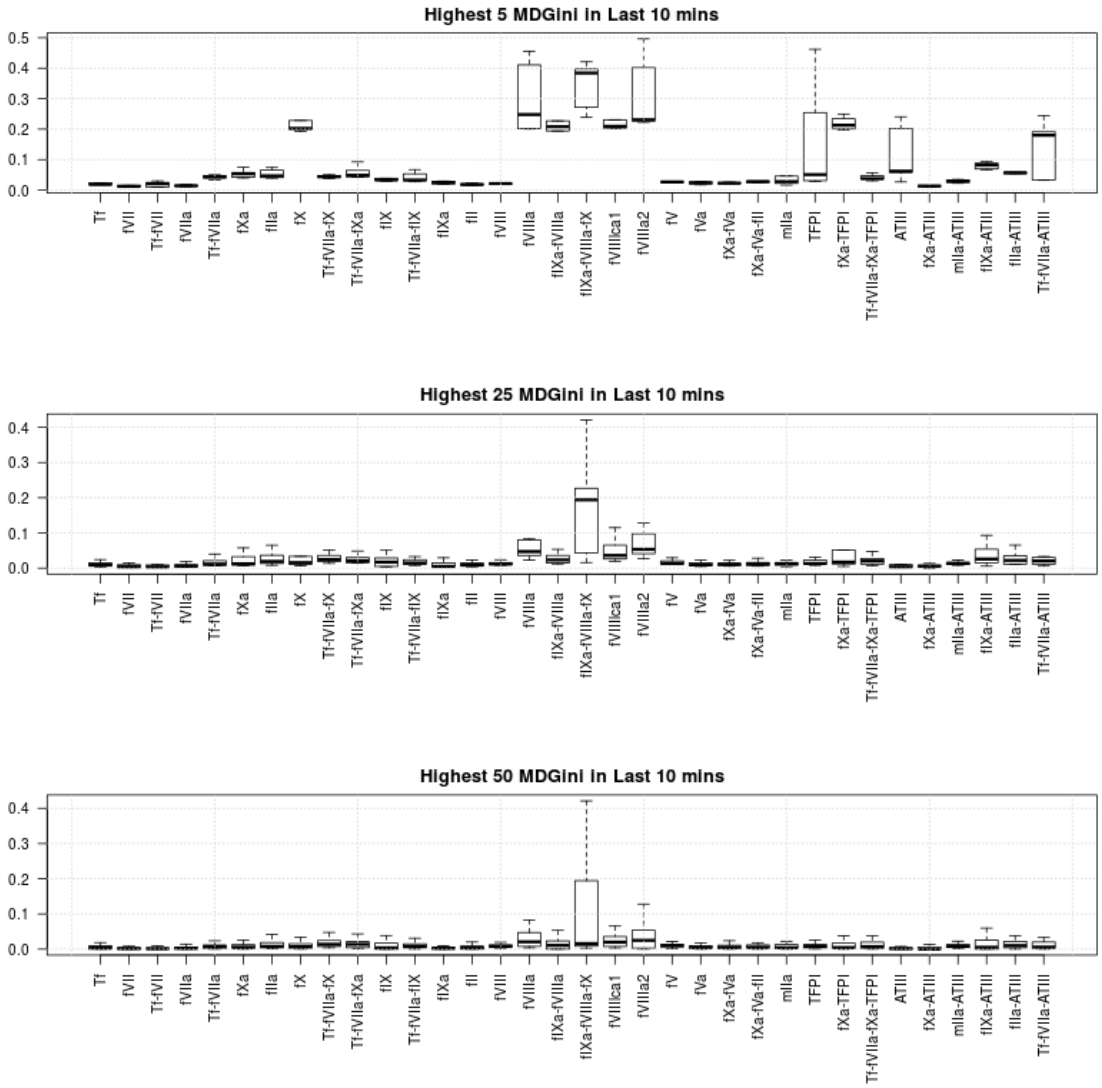

Figure 4.3: Feature significance at the end of simulation. Box plots of MDGini values for the PCHIP coefficients taken from the last ten minutes of the simulation. Unlike 4.2, many species appear significant based on 5 MDGini values. Average of 25 MDGini values makes fIXa-fVIIIa-fX stand out.

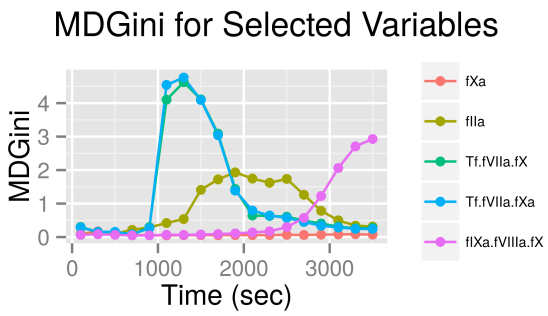

Figure 4.4: MDGini variation (in the classifier built with 200s-MA features) with time for the five selected chemical species. Tf-fVIIa-fXa and Tf-fVIIa-fX are most significant during 1000-1600 seconds, and IIa during 1400-2500 seconds from the addition of the trigger. Significance of fIXa-fVIIIa-fX increases monotonically and remains most significant at 3600 seconds suggesting that it is a long-lived species. 200s-MA - 200-second moving average; MDGini - Mean Decrease in Gini index; Tf-fVIIa-fXa - Tissue factor-factor VIIa-factor Xa; Tf-fVIIa-fX - Tissue factor-factor VIIa-factor X; fIXa-fVIIIa-fX - factor IXa-factor VIIIa-factor X; IIa - activated alpha-thrombin.

Table 4.2: Classification accuracies (%), mean (SD), for 200s-MA values of selected species. Classification using all 18 200s-MA features of Tf-fVIIa-fXa, Tf-fVIIa-fX, fIXa-fVIIIa-fX, and IIa result in similar accuracies. Classification accuracies of the best 3 and the best feature from each species indicate significance is most localized in Tf-fVIIa-fXa.

Species	**All 18**	**Best 3**	**Best 1**
Tf-fVIIa-fXa	83.96 (0.36)	83.18 (0.47)	78.76 (0.08)
Tf-fVIIa-fX	84.23 (0.40)	82.57 (0.38)	77.32 (0.11)
fIXa-fVIIIa-fX	83.80 (0.47)	78.78 (0.56)	75.26 (0.04)
IIa	84.44 (0.59)	75.86 (0.53)	74.26 (0.04)
fXa	82.07 (0.67)	71.36 (0.81)	53.26 (0.08)

MDGini values from the classifier built with all 200s-MA values were used to choose the subset of best features for each species. Tf-fVIIa-fXa - Tissue factor-factor VIIa-factor Xa; Tf-fVIIa-fX - Tissue factor-factor VIIa-factor X; fIXa-fVIIIa-fX - factor IXa-factor VIIIa-factor X; IIa - activated alpha-thrombin; MDGini - Mean Decrease in Gini index.

behaves in a similar way. This behavior contrasts with a species like fXa (Figure 4.5).

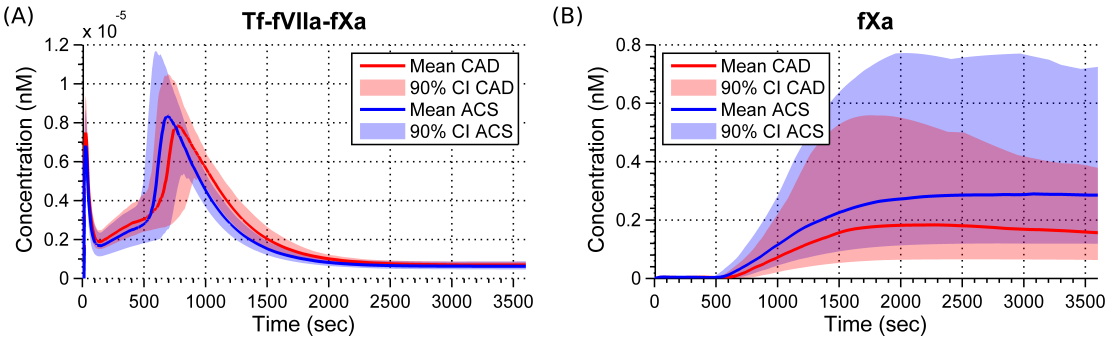

Figure 4.5: Means and 90% quantiles for Tf-fVIIa-fXa and fXa simulation profiles in ACS and CAD populations. A: Tf-fVIIa-fXa concentration profiles from the two groups split significantly from about 1000 to 1500 seconds. B: In fXa concentration profiles, there is a huge variation in both ACS and CAD populations. However, the profiles from the two groups overlap and make the features of this species weak for classification. Tf-fVIIa-fXa - Tissue factor-factor VIIa-factor Xa; fXa - factor Xa.

Among the five species in Table 4.2, fXa has the lowest accuracy of 82.07%. The best feature of fXa classifies with an accuracy of 53.26% which is marginally better than random guessing (50%). This indicates that all the features of fXa are weak. This is due to a huge overlap between the function values in the two groups (Figure 4.5). For fXa, conventional features and 200s-MA values classify with an accuracy of about 82% due to the ways in which these weak features interact. Moreover, conventional features of fXa classify marginally better compared to its 200s-MA values due to lack of time information (time to maximum level, rate, etc.,) in 200s-MA values. This suggests that classification accuracies of every species could be further intensified by considering more features based on time information, in particular, time delay.

Concentration profiles for fIXa-VIIIa-fX in the ACS group appear to live longer

(Figure 4.6). Recent studies suggest existence of active circulating particles in blood (long-lived active species) to be a primary mechanism leading to spontaneous clotting in hyper-coagulable blood [123]. As in the case of fIXa-fVIIIa-fX, the computational approach taken here could be tuned to help identify such long-lived differences under various perturbed conditions of the reaction network. Next, we further this discussion using IIa.

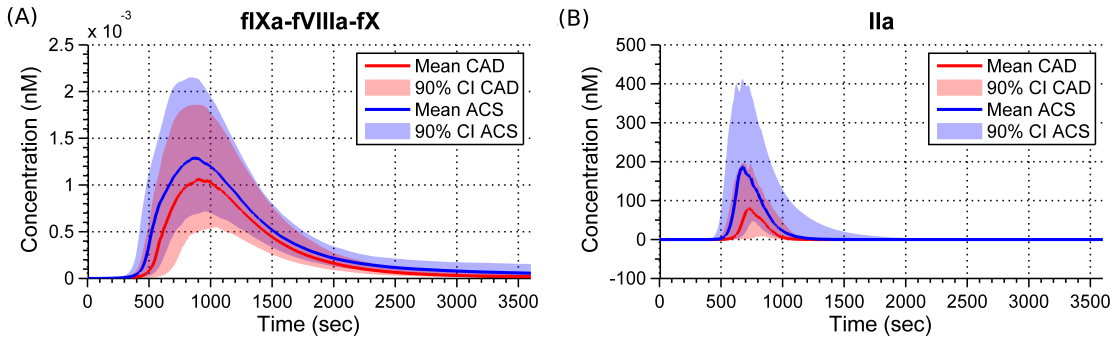

Figure 4.6: Means and 90% quantiles for fIXa-fVIIIa-fX and IIa simulation profiles in ACS and CAD populations. A: fIXa-fVIIIa-fX profiles show that this species is more long-lived in ACS than CAD cases. B: Though IIa concentration profiles appear to reach zero by 2000 seconds, MDGini suggests that the dynamics between the two groups is most significant during that time. fIXa-fVIIIa-fX - factor IXa-factor VIIIa-factor X; IIa - activated alpha-thrombin; MDGini - Mean Decrease in Gini index.

200s-MA values of IIa classify better compared to conventional parameters of active thrombin. It is most significant starting at about 1500 seconds by which time its values are in the order of pM. Changes in this region are usually not considered in conventional features. IIa concentration profiles appear to have reached zero by 2000 seconds (Figure 4.6). However, given the information-theoretic nature of MDGini, it is able to differentiate IIa at regions beyond what is considered as the

termination phase of clotting. Given that biological systems are complicated enough where pico moles of certain chemical species could initiate clotting, and perhaps subsequently determine life or death, we do not overlook such a difference here. Regarding precision, we encountered negative concentration values in IIa in the order of 1E-19 (M). The precisions of the numerical solution and PCHIP approximation are possibly inadequate at this scale.

4.9 Classification Performance of a Few Combinations of Species

Our objective is to find a small combination of features (localized regions in the state space of the model, and labelled in time) which discriminate ACS and CAD well. Classification performance of a few combinations of the selected species is shown in Table 4.3. Average values of Tf-fVIIa-Xa, IIa, and fIXa-fVIIIa-fX at specific times classify with about 87% accuracy. This is better than conventional features or measuring any single species, and is close to using all features considered. For illustration, Figure 4.7 shows a single decision tree built with just two of the best features from fIXa-fVIIIa-fX and Tf-fVIIa-fXa. Typical of chemical kinetics, the region spanned by the data is localized, suggesting dynamics in a low dimensional manifold. The localized separation of the two groups in 2 dimensions of 200s-MA values is seen in the figure.

As our study indicates, we are now in a position to answer the two questions at the beginning of the chapter as follows:

1. There exists chemical species that can be used to classify ACS/CAD better than what can be achieved using thrombin and fXa. Our primary list includes Tf-fVIIa-fXa, Tf-fVIIa-X, fIXa-fVIIIa-fX, and IIa. Tf-fVIIa-fXa is the most important species in our list.

2. Conventional features from active thrombin and fXa can be used to classify

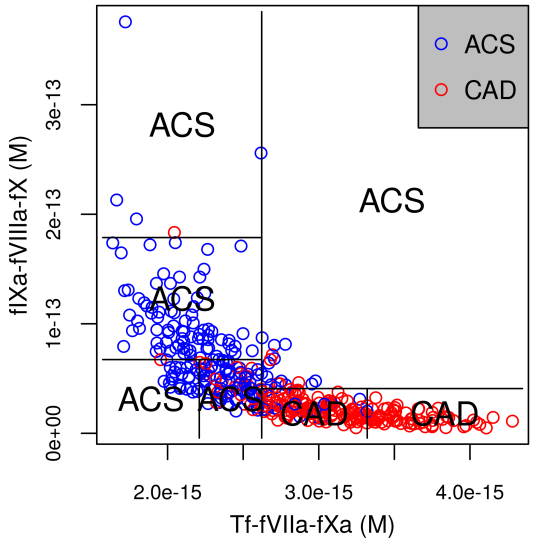

Figure 4.7: Illustrative decision tree. A single decision tree built with just two of the best 200s-MA features, one each from fIXa-fVIIIa-fX and Tf-VIIa-Xa, is shown. ACS and CAD populations separate well in just those two features. fIXa-fVIIIa-fX - factor IXa-factor VIIIa-factor X; Tf-VIIa-Xa - Tissue factor-factor VIIa-factor Xa; 200s-MA - 200-second moving average.

Table 4.3: Classification accuracies (%), mean (SD), for classifiers built using combinations of best 200s-MA features. An efficient way to assay the entire system is by measuring three species at three specific time intervals of 200 seconds. Tf-fVIIa-fXa, IIa and fIXa-fVIIIa-fX make the best combination.

Combination	**Mean (SD)**
Tf-fVIIa-fXa, IIa, and fIXa-fVIIIa-fX	87.16 (0.39)
Tf-fVIIa-fX, IIa, and fIXa-fVIIIa-fX	87.03 (0.40)
Tf-fVIIa-fXa, Tf-fVIIa-fX, and fIXa-fVIIIa-fX	85.58 (0.39)
Tf-fVIIa-fXa, Tf-fVIIa-fX, and IIa	84.90 (0.50)

Tf-fVIIa-fXa at 1400-1600 sec; Tf-fVIIa-fX at 1400-1600 sec; IIa at 1800-2000 sec; fIXa-fVIIIa-fX at 3400-3600 sec; Tf-fVIIa-fXa - Tissue factor-factor VIIa-factor Xa; Tf-fVIIa-fX - Tissue factor-factor VIIa-factor X; IIa - activated alpha-thrombin; fIXa-fVIIIa-fX - factor IXa-factor VIIIa-factor X.

with an accuracy of 81% and 82.6%. There are better features to characterize the system compared to conventional summary parameters, such as initial conditions (plasma factor composition), which result in a classification accuracy of 88.1%. However, plasma factor composition might not capture many attributes of the reaction network. The entire system, when represented using PCHIP coefficients and 200s-MA values, can be used to classify with accuracies of 88.6% and 88.8%. There could be a lot more going on in the system other than changes in thrombin and fXa. For example, activity of IIa (activated alpha-thrombin) was significantly different beyond the termination phase. Long-term activity of such active species warrants better scrutiny.

3. The entire system could be efficiently assayed by measuring a few combinations of species at well-specified times. For example, concentrations of 3 chemical species, namely IIa, Tf-fVIIa-fXa, and fIXa-fVIIIa-fX, averaged over specific time windows (see Table 4.3) chosen relative to the time of trigger (Tf), could be used to classify ACS/CAD to an accuracy of about 87.2%. This is a 7.6% improvement in classification accuracy over using the conventional summary parameters of thrombin.

4.10 Conclusion

We found high-dimensional feature representations for computational solution profiles and studied how combinations of these features could be used to classify ACS/CAD. We modified and tuned the tools offered by Random Forest to fit our purpose. The species we studied are limited by the species considered in the extrinsic pathway model. To the best of our knowledge, this is the first study in the literature to find such localized regions labelled in time and in very low dimensions of the state space that could be associated with ACS. Further validation of the classification

scheme is contingent upon the availability of more detailed data on these two cases. Such localized and effective combinations, which are also easily measurable, could make good global assays for the thrombin generation system.

While the random forest technique is a well accepted method for classification in the statistical learning field and has been used in clinical studies, this is the first study in the literature to apply it to classify ACS/CAD using numerical simulations of the thrombin generation system. The approach shows promise in characterizing hypercoagulability and predicting ACS. Our results open up a way to globally phenotype the thrombin generation system and include specific suggestions for experimental assays to classify ACS/CAD. Currently, measuring some of the recommended chemical species, especially at such low concentration values, may not be practical. However, using models to study combinations of triggers through this approach can reveal measurable chemical species. Moreover, current studies of ACS/CAD classification are restricted to reporting only mean and standard deviation data of plasma factor composition. Wide availability of more raw data would help researchers from diverse fields to study the thrombin generation system and the coagulation cascade. In the next chapter, we focus on modeling dynamics of protein factors that are easily measurable.

5. SIMPLIFIED THROMBIN GENERATION MODEL

"Everything should be made as simple as possible, but not simpler."

– Albert Einstein, *(A longer version is attributed to)*

5.1 Chapter Outline

We introduce the extrinsic thrombin generation model; we describe the proposed simplification; we describe the parameter estimation and prediction process; and then we study the performance of the model. In the process, we highlight a feature that of thrombin dynamics that is crucial for studying blood transport in realistic geometries.

5.2 Need for Model Simplification of Chemical Kinetics

We would like to study, understand, and model the different physiological aspects that cause abnormalities in coagulation. Patient-specific geometry modeling, simulations with more realistic boundary conditions, multiscale models that combine molecular mechanisms with clinical manifestation are some of the open problems discussed in vascular biomechanics [38].

Concentration of species involved in coagulation and rates of reactions vary by orders of magnitude. This requires stochastic methods [124] to account for low concentrations of species properly [125, 126]. The nonlinear chemical kinetics problem is modeled using reaction rates that has quadratic terms, the model is very stiff and solution trajectories are unstable in many directions [67]. The rates involve negative feedback loop or cycles in the reaction cascade [68]. This poses challenges in coupling chemical reactions with flow simulations.

Moreover, there is uncertainty in the parameters of the chemical kinetics model.

The plasma factor composition varies drastically in and within patients. Many of the rate constants are inferred indirectly rather than being directly measured. This demands multiple simulations typically in a bayesian framework. Models used to simulate clot in small two dimensional regions ($\sim$100 μm) consider the dynamics of many reactants [127]. These are inappropriate for simulations in realistic 3 dimensional flow conditions in arteries, say, in order to study atheroscelerosis or thrombosis [127]. Further, there are strong gradients near the boundaries where mass transport happens. This requires extremely fine grids near the boundaries [70] accounting for complex chemical kinetics models have made realistic patient-specific simulations an open problem. Reduced dimensional simplified chemical kinetics models will help to perform patient-specific simulations.

5.3 Background Literature

Papadopoulos et al. [127] suggested a phenomenological model for thrombin generation. Based on the mechanism of thrombin generation, they propose a simplified thrombin generation model using four reactions. The reactions include dynamics of thrombin, prothrombin, platelets, and activated platelets. Using the assumption of fast platelet activation, they derive analytical expression for thrombin generation. These are similar to thrombin generation functions prescribed by Hemker et al. [63]. The model essentially fits patient-specific thrombin generation profiles and the effect of the plasma factor composition and inhibitors on the dynamics of thrombin generation were not emphasized.

This motivated Sagar et al. [128] to come up with a dynamical model for thrombin generation using a hybrid strategy. The strategy combines differential equations and several logical rules to model thrombin generation. They design their approach to model systems where mechanistic insights are poor and experimental interrogation

is difficult. This results in reduced order model that has rates of the product of Hill-like terms [129] and activation functions that act as the logical rules, i.e.,

$$r_i = \frac{k_i x_i^{\eta_i}}{1 + k_i x_i^{\eta_i}} \min\left(\frac{k_j x_j^{\eta_j}}{1 + k_j x_j^{\eta_j}}, \frac{k_m x_m^{\eta_m}}{1 + k_m x_m^{\eta_m}}\right) \tag{5.1}$$

where r is the reaction rate, k and η are parameters, and x pertains to protein concentration or activity. Though the model shows great performance, the transparency in the mechanistic models such as [63] and [127] is lost. We seek a middle ground between the two simplified models where we find a dynamical model that makes use of the mechanistic knowledge of blood coagulation and is able to account for changes in plasma factor composition.

We suggest a simple phenomenological model for thrombin generation:

1. We model the stoichiometry of certain important chemical species. The model is based on the classically viewed initiation, propagation, and termination of thrombin generation. Hence the chemistry involved in the simulations should offer physiological insight.

2. The functionality of the parameters are evident and different aspects of thrombin generation are easily alterable.

3. A good model should be able to capture the necessary rich behavior of the phenomena as well as generalize well in order to predict important qualitative and quantitative responses. The model we propose is able to predict certain important changes in thrombin generation due to changes in prothrombin and antithrombin concentration.

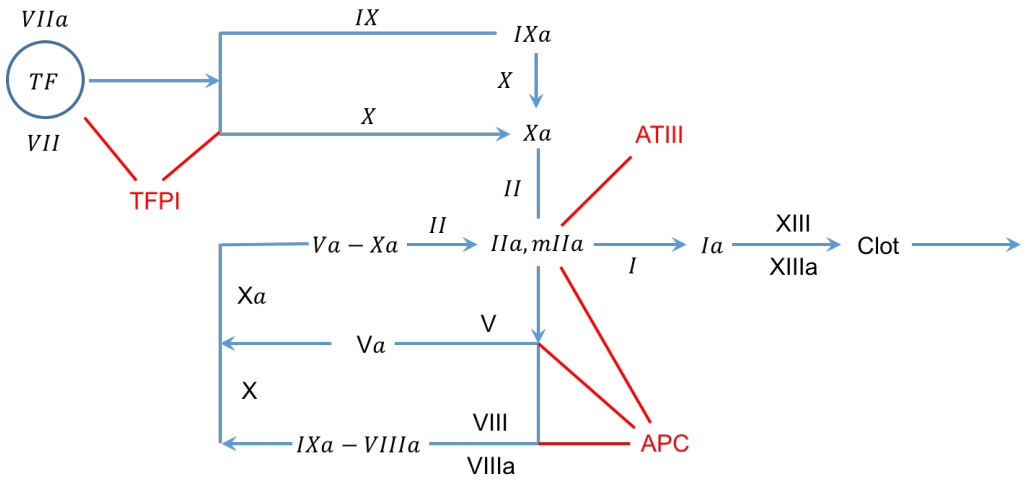

Figure 5.1: Schematic of the extrinsic pathway. We propose a simplified model for thrombin reduction. Note that events that occur after thrombin generation result in changes of mechanical properties.

5.4 Extrinsic Thrombin Generation

Figure 5.1 shows a schematic of the key elements of the extrinsic pathway involved during clotting. We consider the extrinsic pathway because hemostasis occurs due to tissue factor initiation. Further, we simply simplify thrombin generation. Given that flow properties affect and are affected by fibrin formation, such a simple thrombin generation model factors out the two phenomena. We will focus on an extrinsic parameter model as hemostasis occurs due this pathway. In particular, we study blood coagulation using a model for the Tissue factor(Tf)-initiated extrinsic pathway developed by Hockin et al. [60]. A schematic of the model used is shown in Figure 5.2.

We exploit two key ideas for model simplification:

1. We separate the parameters in the initiation and termination phase of throm-

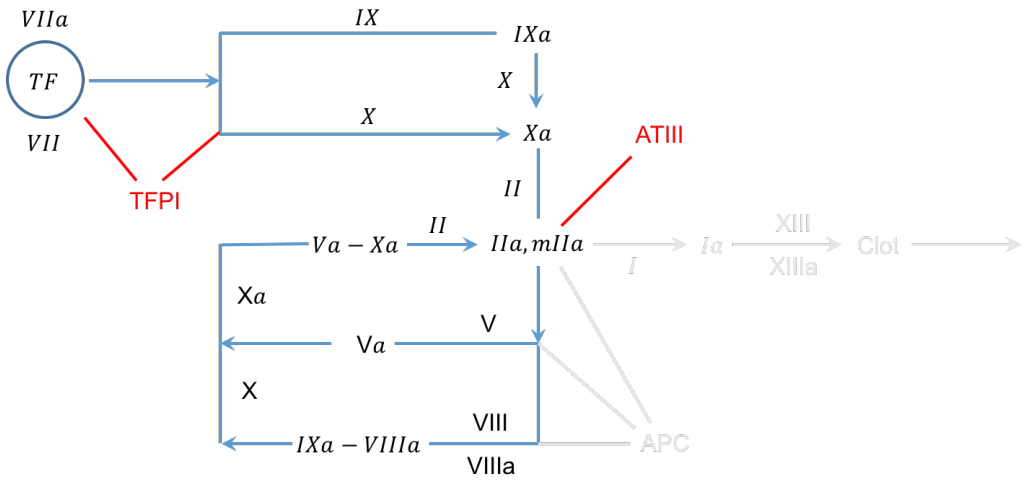

Figure 5.2: Schematic of the extrinsic pathway model for thrombin generation used in this work. There are essentially three elements in this network, namely, i) thrombin initiation; ii) thrombin propagation; and iii) thrombin inhibition.

bin using a switching model. The model switches from initiation to propogation/termination phase based on the concentration of thrombin. A similar idea was exploited in [127]. Though [128] also models switching using logical rules in the transfer function, they are not as explicit as in our model.

2. We model the effect of patient-specificity of thrombin generation using variations in the rate constants of the model. We show that training the model for a specific choice of initial condition (physiological mean) is able to predict qualitative responses of changes in prothrombin and antithrombin concentration.

We use the traditional simplification of the thrombin generation cascade:

1. **Thrombin Initiation**: Tissue factor activates prothrombin to form thrombin.

2. **Thrombin propagation**: Given that sufficient amount of thrombin is activated, clotting propagates via a different set of reactions.

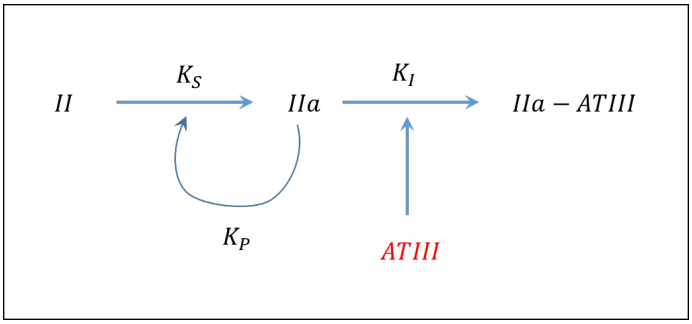

Figure 5.3: Schematic of the simplified model proposed in this study. There is fuel prothrombin, the key enzyme thrombin, the inhibitor antithrombin, and the by-product anti-thrombin. K_S is a rate constant that models initiation due to injury. K_P is the rate of thrombin propagation. K_I is the rate of thrombin inhibition.

3. **Thrombin inhibition**: Finally normal hemostasis requires that thrombin generation is controlled so that clot is localized.

We describe kinetics for prothrombin, thrombin, antithrombin, and thrombin-antithrombin (Figure 5.3) using the following set of rates:

$$\begin{aligned}
\frac{d}{dt}[\text{II}] &= -K_S - K_P[\text{II}][\text{IIa}] \\
\frac{d}{dt}[\text{IIa}] &= K_S + K_P[\text{II}][\text{IIa}] - K_I[\text{IIa}][\text{AT}] \\
\frac{d}{dt}[\text{AT}] &= -K_I[\text{IIa}][\text{AT}] \\
\frac{d}{dt}[\text{IIa-AT}] &= K_I[\text{IIa}][\text{AT}]
\end{aligned} \tag{5.2}$$

where we model initiation using the rate constant K_S, propagation using the rate constant K_P, and inhibition using the rate constant K_I. In order to be stoichiometrically consistent, our [IIa] is the sum of both forms of thrombin in the full 34 variable model (the variables used in our model are drawn in continuos lines in the full model simulation as seen in Figure 5.4). [IIa-AT] in our model is the sum of the

two antithrombin complex formed due to inhibition[1].

We propose the following switching rules that changes the response of the model during initiation and propagation/termination.

$$K_S = \begin{cases} k_s > 0, & [IIa] < 2nM \text{ and injury} \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

$$K_I = \begin{cases} k_{i2} > 0, & [IIa] < 2nM \\ k_{i1} > 0, & \text{otherwise} \end{cases} \tag{5.4}$$

$$K_P = \begin{cases} 0, & [IIa] < 2nM \\ k_p > 0, & \text{otherwise} \end{cases} \tag{5.5}$$

Essentially, thrombin propagation occurs if [IIa] crosses a threshold. For normal clotting, rate of propagation is expected to be orders of magnitude higher than that of rate of initiation, i.e., $k_p \ll k_s$. We also use two different inhibition rate constant k_{i1} and k_{i2} so that rate of inhibition could be separately controlled during the two phases.

5.5 Estimation of Model Parameters

We carried out the simulations of the full model using the Tf-initiated clotting model [60]. In the simulations, clotting was initiated with 5 pM and the plasma factor composition was set to physiological mean values [60] (Table 2.1). We used

[1]there are other antithrombin complexes formed in the full model but they are 3 orders of magnitude smaller than [IIa-AT] and we neglect them.

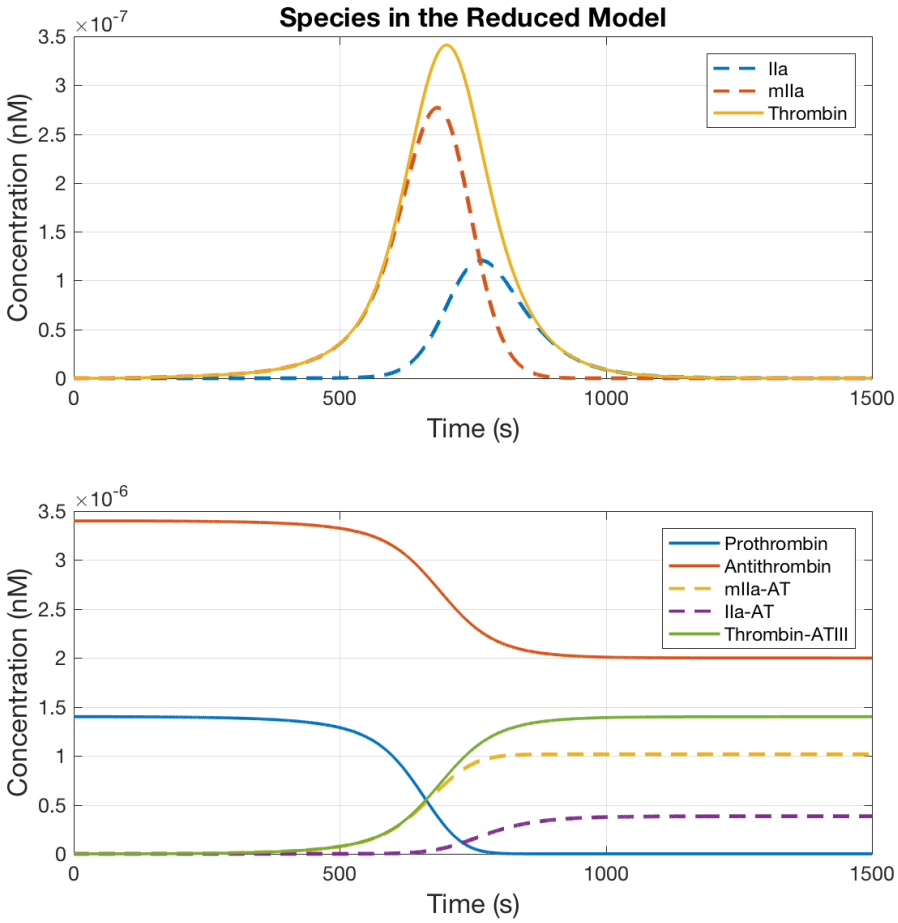

Figure 5.4: Stoichiometry of thrombin and antithrombin. Only the species drawn with a continuous straight line are considered in the model. This ensures stoichiometric consistency. [IIa] or thrombin in our model is sum of the two forms of thrombin in the full model. Similarly, [IIa-AT] or thrombin-antithrombin in our model the sum of the two by products of thrombin inhibition due to antithrombin in the full model.

the data from the full model to fit parameters for the reduced model. Particle swarm optimization was used for parameter estimation and we obtained one set of parameters for the physiological mean composition.

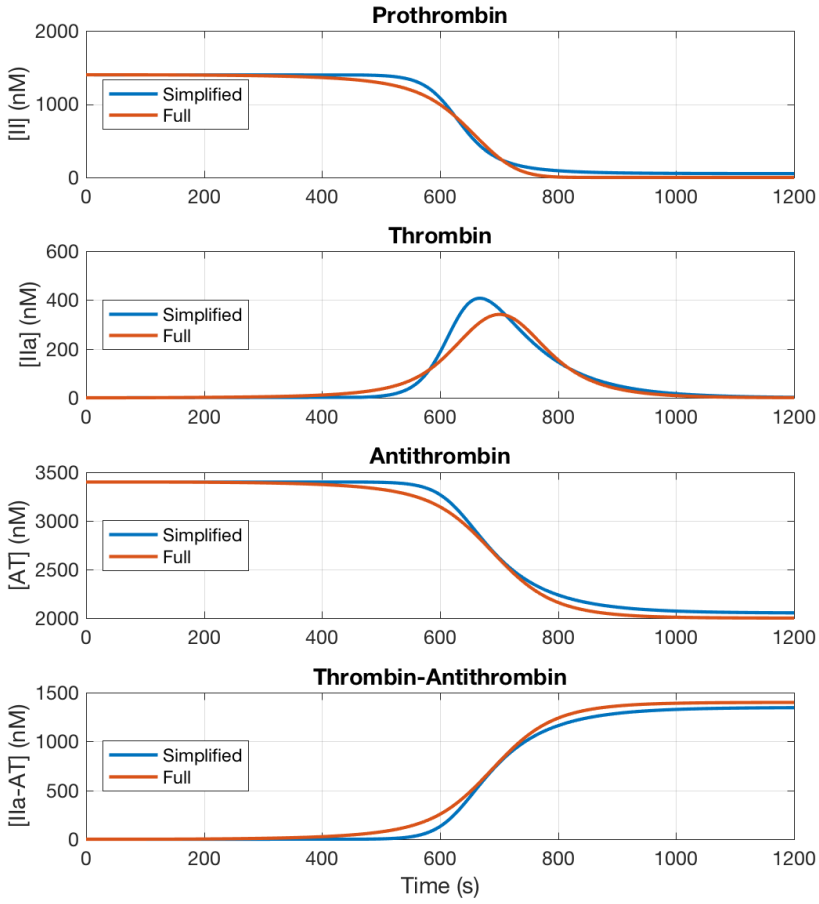

Figure 5.5: Comparison of all the species modeled in the simplified model. The simplified model captures the full models behaviour very well. There is a slight mismatch in the maximum amount of thrombin generation. This could be improved by also choosing the clot propagation threshold ([IIa] = 2 nM) better.

We used the following objective for minimization,

$$u = \sum_{i=1}^{4} \frac{(C_i^{reduced} - C_i^{full})^{1/2}}{C_i^{constant}} \tag{5.6}$$

which is sum of squared differences of the normalized concentration profiles between the common species in the full model and the reduced model. Mean concentration of prothrombin and antithrombin were used as normalization constants. Comparison of the reduced and full model simulation for the physiological mean initial composition is shown in Figure 5.5.

5.6 Parameter Study of the Simplified Model

We show the effect of parameters by changing one at a time. Clot time depends exponentially on K_S (seen in Figure 5.6). Rate constant k_{i2} also controls the clot time Figure 5.7. For certain combinations of K_S and k_{i2} it takes more than 1200 seconds for clot initiation. Both the parameters together offer more control over dynamics of clot initiation.

Figures 5.8 and 5.9 show the effect of changes in the rate constants K_P and k_{i1} respectively. The parameters offer a wide range of thrombin generation rates. Similar to the initiation, there are certain values of K_P (for a given value of k_{i1}) and vice versa where thrombin generation is too low. These two parameters together offer control over simulating a wide range of thrombin propagation.

5.7 Prediction of Variation in Prothrombin and Antithrombin

Finally, we check the qualitative response of the model predictions for variations in initial prothrombin and antithrombin concentrations. As seen in Figure 5.10, higher values of prothrombin are able to predict more thrombin generation. This has been observed in experiments [39]. Thrombin rates during termination could

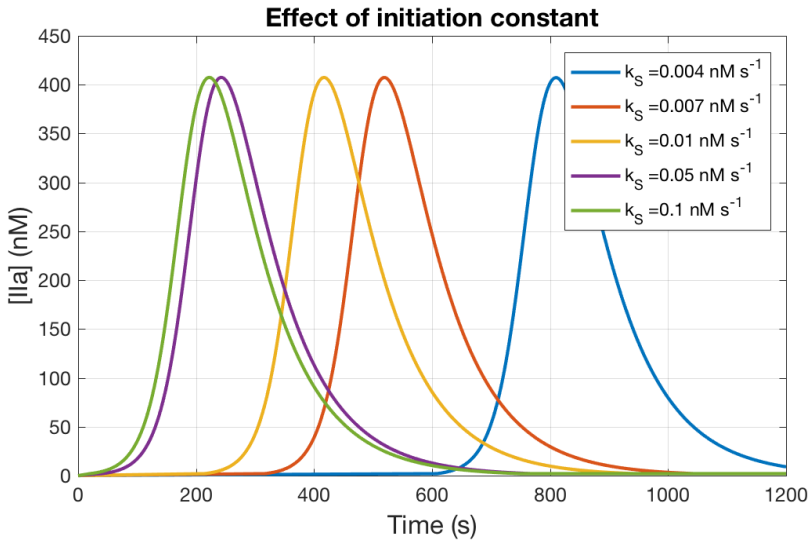

Figure 5.6: Controlling thrombin initiation using K_S. There is an exponential dependence of clot time on this parameter.

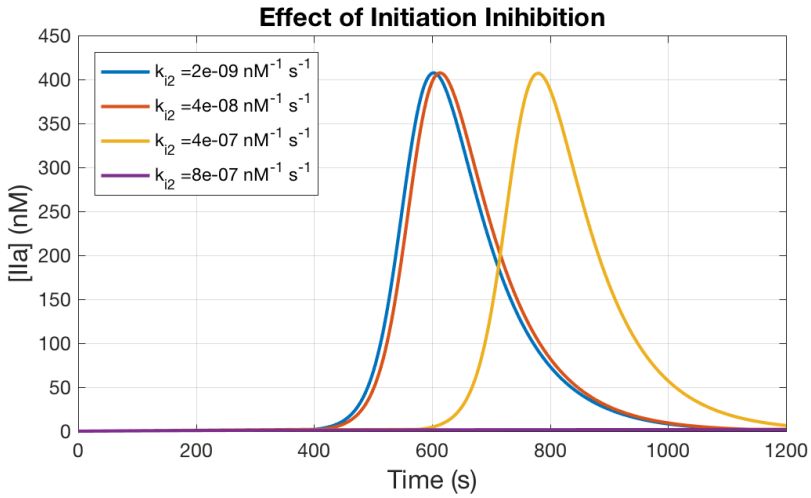

Figure 5.7: Controlling thrombin initiation using k_{i2}. This parameter along with K_S, allows for modeling a wide range of clot times and dynamics during initiation.

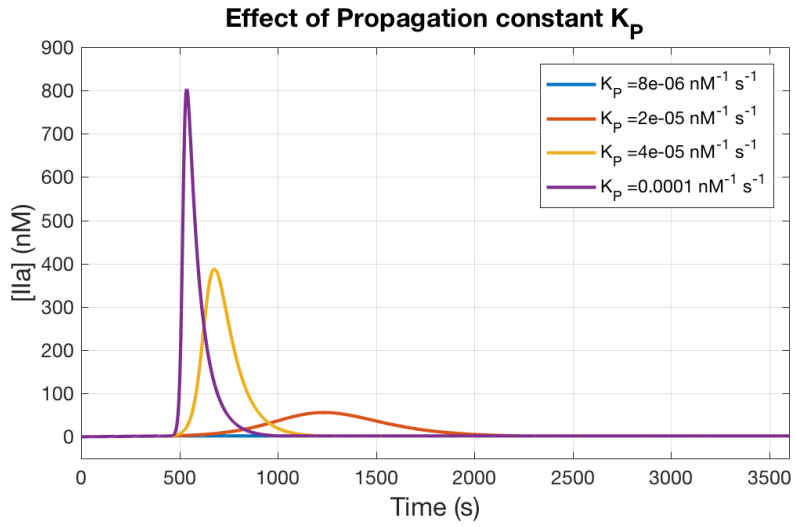

Figure 5.8: Controlling thrombin propagation using K_P. As expected, variations in the propagation rate constant is able to capture a wide range of thrombin generation rate.

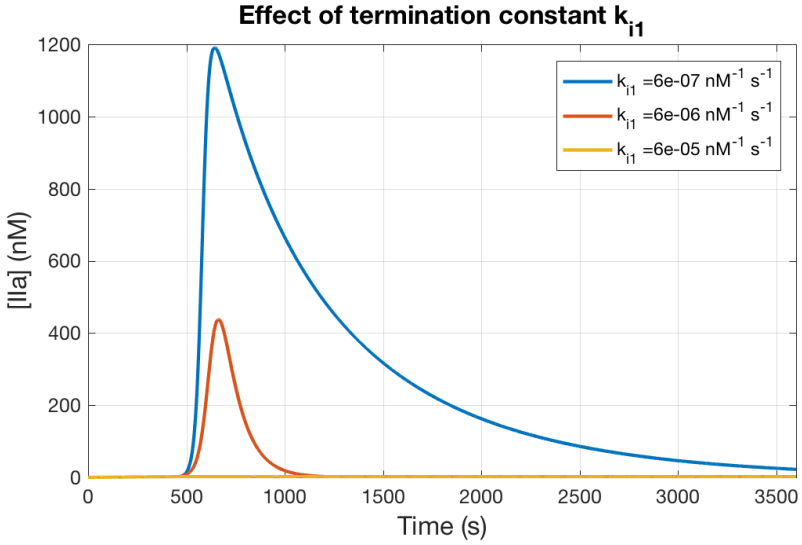

Figure 5.9: Controlling thrombin termination using k_{i1}. This parameter has more effect on the termination phase of thrombin generation.

be improved using better training data and using different reaction rates. Similarly, lower values of antithrombin are able to predict higher thrombin generation.

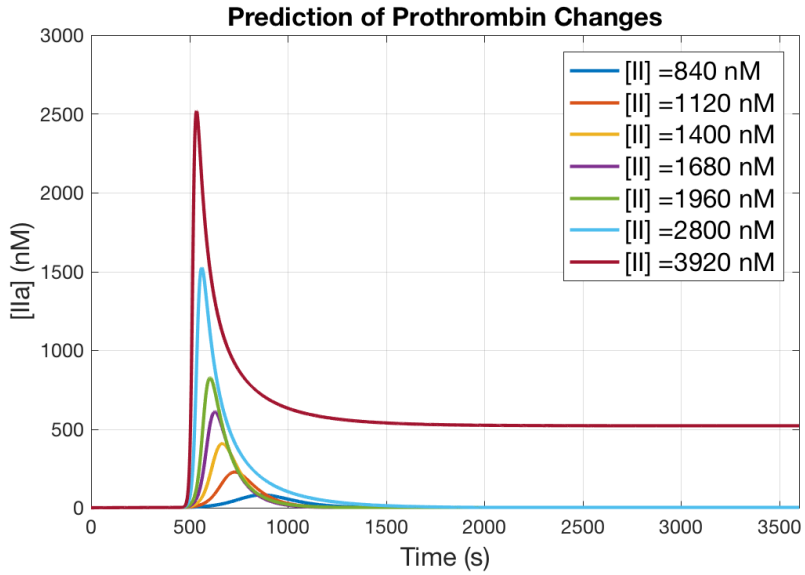

Figure 5.10: Prediction on prothrombin variation.

One of the most important predictions of this model is that thrombin termination appears to halt at non-zero values (Figure 5.11). Such sustained activity is also in observed in experiments when there is too much prothrombin compared to antithrombin [39]. Inhibitors such as APC may need to modeled in order to account for oscillations observed in such sustained activity. In this model, the reaction essentially runs out of the inhibitor [AT] when initiated with a certain plasma factor composition. In such a scenario, other phenomena like diffusion and convection will control the extent of clotting. For example, when there is less inflow of antithrombin concentration, as in the case of stasis, we would expect more clotting due to the presence of excess active thrombin. Further, active thrombin could be propagated

downstream and could cause clotting elsewhere. This observation and model prediction on sustained activity of thrombin are hypothesized to play a necessary role towards effectively studying clotting in realistic geometries.

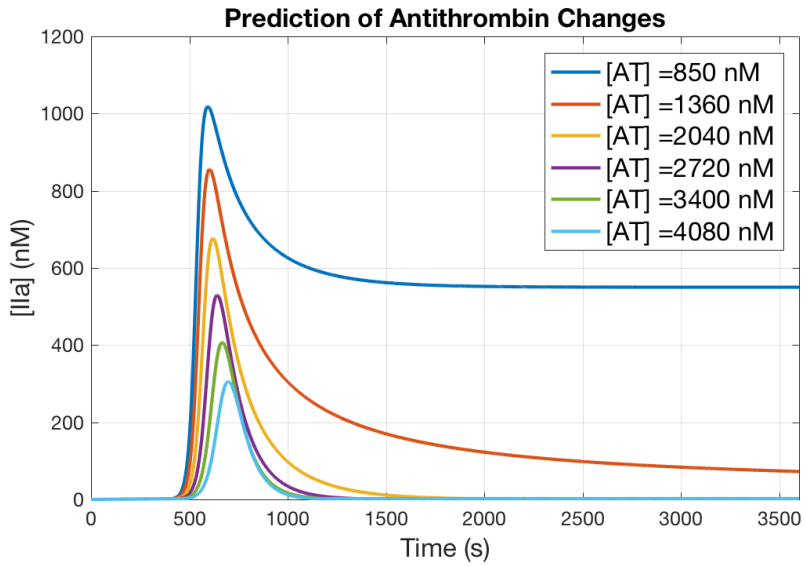

Figure 5.11: Prediction on antithrombin variation. This is the most important and significant prediction of the model. This has been observed in thrombin generation experiments [39]. Moreover, this phenomena could be blind to markers like TAT that estimate thrombin activity.

The parameters need to be altered in order to better simulate thrombin generation on a wider range of plasma factor composition. We are currently working on augmenting this model with data-driven models to predict thrombin generation in a wider range of plasma factor composition and to account for the effect of transport.

5.8 Conclusion

We proposed a simplified model for thrombin generation based on the stoichiometry of certain important chemical species. The model fits data well and different

aspects of thrombin dynamics are easily alterable. The model is able to predict certain important changes in thrombin generation due to changes in prothrombin and antithrombin concentration.

6. CONCLUSIONS AND FUTURE DIRECTIONS

"At the root of science and scientific research is the urge, the compulsion, to understand the nature of things."

– David Ruelle, *The Mathematicians Brain*

Our results include the following:

- We sampled patient-specific plasma factor composition using the maximum entropy principle. We used GMM to model dependent features and used expectation maximization to infer the GMM parameters. The classification model we built using GMM of 2 thrombin features had an accuracy of 77 %. In the future, GMM could be further used to better describe plasma factor composition.

- We used information from all protein factors in the extrinsic pathway model to classify ACS and CAD. While the conventionally used thrombin generation parameters could classify with accuracies close to 81 %, the models we built using information from all protein factors could classify with accuracies close to 88.7 %. Further, we found certain combinations of 3 protein factor activities at specific times during clotting that could classify ACS and CAD with accuracies close to 87 %. These improvements obtained using information from model simulations has the potential to screen thousands of patients with greater certainty every year.

- Finally, we proposed a simplified model for the dynamics of thrombin. The model predicts sustained thrombin activity in plasma factor compositions with low concentrations of the inhibitor antithrombin compared to prothrombin concentration.

This work could be improved by coupling with simplified platelet aggregation and fibrin dynamics models. Further, such simplified models need to be coupled with fluid flow and transport models. The likely involvement of thrombin in cancer and inflammatory diseases [103] needs to be better characterized.

We hypothesize that sustained thrombin activity is a key phenomenon to model clotting in realistic geometries. It opens up a way to study the effect of clotting in one part of the body on clotting elsewhere in the body. We also envision better risk assessment by the use of complex networks [130] to model transport in the human body. The simplified model proposed in this work is expected to catalyze any such advancements.

To conclude, in this thesis we have developed a non self-evident simplified model for patient-specific thrombin dynamics.

REFERENCES

[1] J. Mackay, G. A. Mensah, S. Mendis, and K. Greenlund, *The Atlas of Heart Disease and Stroke*. World Health Organization, 2004.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. de Ferranti, J.-P. Despres, H. J. Fullerton, V. J. Howard, *et al.*, "Heart disease and stroke statistics-2015 update: a report from the American Heart Association.," *Circulation*, vol. 131, no. 4, p. e29, 2015.

[3] S. A. Achar, S. Kundu, and W. A. Norcross, "Diagnosis of acute coronary syndrome," *Chest*, vol. 100, p. 9, 2005.

[4] V. Fuster, L. Badimon, J. J. Badimon, and J. H. Chesebro, "The pathogenesis of coronary artery disease and the acute coronary syndromes," *New England Journal of Medicine*, vol. 326, no. 4, pp. 242–250, 1992.

[5] H. A. DeVon, N. Hogan, A. L. Ochs, and M. Shapiro, "Time to treatment for acute coronary syndromes: the cost of indecision," *The Journal of Cardiovascular Nursing*, vol. 25, no. 2, p. 106, 2010.

[6] D. K. Moser, L. P. Kimble, M. J. Alberts, A. Alonzo, J. B. Croft, K. Dracup, K. R. Evenson, A. S. Go, M. M. Hand, R. U. Kothari, *et al.*, "Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the american heart association council on cardiovascular nursing and stroke council," *Journal of Cardiovascular Nursing*, vol. 22, no. 4, pp. 326–343, 2007.

[7] S. S. Johnston, S. Curkendall, D. Makenbaeva, E. Mozaffari, R. Goetzel, W. Burton, and R. Maclean, "The direct and indirect cost burden of acutecoro-

nary syndrome," *Journal of Occupational and Environmental Medicine*, vol. 53, no. 1, pp. 2–7, 2011.

[8] K. A. Fox, O. H. Dabbous, R. J. Goldberg, K. S. Pieper, K. A. Eagle, F. Van de Werf, Á. Avezum, S. G. Goodman, M. D. Flather, F. A. Anderson, *et al.*, "Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (grace)," *BMJ*, vol. 333, no. 7578, p. 1091, 2006.

[9] J. L. Mega, E. Braunwald, S. D. Wiviott, J.-P. Bassand, D. L. Bhatt, C. Bode, P. Burton, M. Cohen, N. Cook-Bruns, K. A. Fox, *et al.*, "Rivaroxaban in patients with a recent acute coronary syndrome," *New England Journal of Medicine*, vol. 366, no. 1, pp. 9–19, 2012.

[10] J. W. Eikelboom, S. R. Mehta, S. S. Anand, C. Xie, K. A. Fox, and S. Yusuf, "Adverse impact of bleeding on prognosis in patients with acute coronary syndromes," *Circulation*, vol. 114, no. 8, pp. 774–782, 2006.

[11] M. Lamberts, J. B. Olesen, M. H. Ruwald, C. M. Hansen, D. Karasoy, S. L. Kristensen, L. Køber, C. Torp-Pedersen, G. H. Gislason, and M. L. Hansen, "Bleeding after initiation of multiple antithrombotic drugs, including triple therapy, in atrial fibrillation patients following myocardial infarction and coronary intervention: a nationwide cohort study," *Circulation*, vol. 126, no. 10, pp. 1185–1193, 2012.

[12] M. A. Panteleev and H. C. Hemker, "Global/integral assays in hemostasis diagnostics: promises, successes, problems and prospects," *Thrombosis Journal*, vol. 13, no. 1, pp. 1–4, 2015.

[13] M. Anand, K. Rajagopal, and K. Rajagopal, "A model incorporating some of the mechanical and biochemical factors underlying clot formation and dissolu-

tion in flowing blood: review article," *Journal of Theoretical Medicine*, vol. 5, no. 3-4, pp. 183–218, 2003.

[14] R. Tran, D. R. Myers, J. Ciciliano, E. L. Trybus Hardy, Y. Sakurai, B. Ahn, Y. Qiu, R. G. Mannino, M. E. Fay, and W. A. Lam, "Biomechanics of haemostasis and thrombosis in health and disease: from the macro-to molecular scale," *Journal of Cellular and Molecular Medicine*, vol. 17, no. 5, pp. 579–596, 2013.

[15] S. L. Diamond, "Systems analysis of thrombus formation," *Circulation Research*, vol. 118, no. 9, pp. 1348–1362, 2016.

[16] J. D. Humphrey, "Mechanics of the arterial wall: review and directions," *Critical Reviews in Biomedical Engineering*, vol. 23, no. 1-2, pp. 1–162, 1994.

[17] C. A. Taylor and M. T. Draney, "Experimental and computational methods in cardiovascular fluid mechanics," *Annual Review of Fluid Mechanics*, vol. 36, pp. 197–231, 2004.

[18] K. G. Mann, S. Butenas, and K. Brummel, "The dynamics of thrombin formation," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 23, no. 1, pp. 17–25, 2003.

[19] M. Schenone, B. C. Furie, and B. Furie, "The blood coagulation cascade," *Current Opinion in Hematology*, vol. 11, no. 4, pp. 272–277, 2004.

[20] K. Mann, K. Brummel, and S. Butenas, "What is all that thrombin for?," *Journal of Thrombosis and Haemostasis*, vol. 1, no. 7, pp. 1504–1514, 2003.

[21] P. Morawitz, R. Hartmann, and P. F. Guenther, *The Chemistry of Blood Coagulation*. No. 314, Thomas, 1958.

[22] S. Rapaport, "Blood coagulation and its alterations in hemorrhagic and thrombotic disorders," *Western Journal of Medicine*, vol. 158, no. 2, p. 153, 1993.

[23] C. N. Bagot and R. Arya, "Virchow and his triad: a question of attribution," *British journal of haematology*, vol. 143, no. 2, pp. 180–190, 2008.

[24] B. C. Dickson, "Venous thrombosis: on the history of Virchow's triad," *University of Toronto Medical Journal*, vol. 81, no. 3, pp. 166–171, 2004.

[25] A. J. Quick, M. Stanley-Brown, and F. W. Bancroft, "A study of the coagulation defect in hemophilia and in jaundice," *The American Journal of the Medical Sciences*, vol. 190, no. 4, pp. 501–510, 1935.

[26] D. LaCroix and M. Anand, "A model for the formation, growth, and dissolution of clots in vitro. Effect of the intrinsic pathway on antithrombin III deficiency and protein C deficiency," *International Journal of Advances in Engineering Sciences and Applied Mathematics*, vol. 3, no. 1-4, pp. 93–105, 2011.

[27] A. J. Patek Jr and F. H. L. Taylor, "Hemophilia. II. Some properties of a substance obtained from normal human plasma effective in accelerating the coagulation of hemophilic blood," *Journal of Clinical Investigation*, vol. 16, no. 1, p. 113, 1937.

[28] K. G. Mann, C. M. Heldebrant, and D. N. Fass, "Multiple active forms of thrombin I. partial resolution, differential activities, and sequential formation," *Journal of Biological Chemistry*, vol. 246, no. 19, pp. 5994–6001, 1971.

[29] M. E. Nesheim, K. H. Myrmel, L. Hibbard, and K. G. Mann, "Isolation and characterization of single chain bovine factor V," *Journal of Biological Chemistry*, vol. 254, no. 2, pp. 508–517, 1979.

[30] M. Nesheim, R. Tracy, and K. Mann, "'Clotspeed', a mathematical simulation of the functional properties of prothrombinase," *Journal of Biological Chemistry*, vol. 259, no. 3, pp. 1447–1453, 1984.

[31] G. M. Willems, T. Lindhout, W. T. Hermens, and H. C. Hemker, "Simulation model for thrombin generation in plasma," *Pathophysiology of Haemostasis and Thrombosis*, vol. 21, no. 4, pp. 197–207, 1991.

[32] K. C. Jones and K. G. Mann, "A model for the tissue factor pathway to thrombin. II. a mathematical simulation," *Journal of Biological Chemistry*, vol. 269, no. 37, pp. 23367–23373, 1994.

[33] A. L. Fogelson, "A mathematical model and numerical method for studying platelet adhesion and aggregation during blood clotting," *Journal of Computational Physics*, vol. 56, no. 1, pp. 111–134, 1984.

[34] H. C. Hemker, P. Giesen, R. Al Dieri, V. Regnault, E. De Smedt, R. Wagenvoord, T. Lecompte, and S. Béguin, "Calibrated automated thrombin generation measurement in clotting plasma," *Pathophysiology of Haemostasis and Thrombosis*, vol. 33, no. 1, pp. 4–15, 2003.

[35] V. Zarnitsina, A. Pokhilko, and F. Ataullakhanov, "A mathematical model for the spatio-temporal dynamics of intrinsic pathway of blood coagulation. i. the model description," *Thrombosis Research*, vol. 84, no. 4, pp. 225–236, 1996.

[36] J. C. Fredenburgh, P. L. Gross, and J. I. Weitz, "Emerging anticoagulant strategies," *Blood*, vol. 129, no. 2, pp. 147–154, 2017.

[37] M. A. Panteleev, N. M. Dashkevich, and F. I. Ataullakhanov, "Hemostasis and thrombosis beyond biochemistry: roles of geometry, flow and diffusion," *Thrombosis Research*, vol. 136, no. 4, pp. 699–711, 2015.

[38] C. A. Taylor and J. Humphrey, "Open problems in computational vascular biomechanics: hemodynamics and arterial wall mechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 198, no. 45, pp. 3514–3523, 2009.

[39] G. A. Allen, A. S. Wolberg, J. A. Oliver, M. Hoffman, H. R. Roberts, and D. M. Monroe, "Impact of procoagulant concentration on rate, peak and total thrombin generation in a model system," *Journal of Thrombosis and Haemostasis*, vol. 2, no. 3, pp. 402–413, 2004.

[40] K. Brummel-Ziedins, C. Vossen, F. Rosendaal, K. Umezaki, and K. Mann, "The plasma hemostatic proteome: thrombin generation in healthy individuals," *Journal of Thrombosis and Haemostasis*, vol. 3, no. 7, pp. 1472–1481, 2005.

[41] T. Orfeo, S. Butenas, K. E. Brummel-Ziedins, and K. G. Mann, "The tissue factor requirement in blood coagulation," *Journal of Biological Chemistry*, vol. 280, no. 52, pp. 42887–42896, 2005.

[42] J. Crawley, S. Zanardelli, C. Chion, and D. Lane, "The central role of thrombin in hemostasis," *Journal of Thrombosis and Haemostasis*, vol. 5, no. s1, pp. 95–101, 2007.

[43] T. Baglin, "The measurement and application of thrombin generation," *British Journal of Haematology*, vol. 130, no. 5, pp. 653–661, 2005.

[44] H. C. Hemker, R. Al Dieri, E. De Smedt, S. Béguin, *et al.*, "Thrombin generation, a function test of the haemostatic-thrombotic system," *Thromb Haemost*, vol. 96, no. 5, pp. 553–61, 2006.

[45] K. Brummel-Ziedins, "Models for thrombin generation and risk of disease," *Journal of Thrombosis and Haemostasis*, vol. 11, no. s1, pp. 212–223, 2013.

[46] K. E. Brummel-Ziedins, S. J. Everse, K. G. Mann, and T. Orfeo, "Modeling thrombin generation: plasma composition based approach," *Journal of Thrombosis and Thrombolysis*, vol. 37, no. 1, pp. 32–44, 2014.

[47] K. Brummel-Ziedins, R. Pouliot, and K. Mann, "Thrombin generation: phenotypic quantitation," *Journal of Thrombosis and Haemostasis*, vol. 2, no. 2, pp. 281–288, 2004.

[48] K. E. Brummel-Ziedins, T. Orfeo, F. R. Rosendaal, A. Undas, G. E. Rivard, S. Butenas, and K. G. Mann, "Empirical and theoretical phenotypic discrimination," *Journal of Thrombosis and Haemostasis*, vol. 7, no. s1, pp. 181–186, 2009.

[49] K. Brummel-Ziedins, A. Undas, T. Orfeo, M. Gissel, S. Butenas, K. Zmudka, and K. Mann, "Thrombin generation in acute coronary syndrome and stable coronary artery disease: dependence on plasma factor composition," *Journal of Thrombosis and Haemostasis*, vol. 6, no. 1, pp. 104–110, 2008.

[50] A. Undas, K. Szułdrzyński, K. E. Brummel-Ziedins, W. Tracz, K. Zmudka, and K. G. Mann, "Systemic blood coagulation activation in acute coronary syndromes," *Blood*, vol. 113, no. 9, pp. 2070–2078, 2009.

[51] A. Undas, M. Jankowski, P. Kaczmarek, K. Sladek, and K. Brummel-Ziedins, "Thrombin generation in chronic obstructive pulmonary disease: dependence on plasma factor composition," *Thrombosis Research*, vol. 128, no. 4, pp. e24–e28, 2011.

[52] M. Jankowski, A. Undas, P. Kaczmarek, and S. Butenas, "Activated factor XI and tissue factor in chronic obstructive pulmonary disease: links with inflammation and thrombin generation," *Thrombosis Research*, vol. 127, no. 3, pp. 242–246, 2011.

[53] M. Gissel, A. Undas, A. Slowik, K. G. Mann, and K. E. Brummel-Ziedins, "Plasma factor and inhibitor composition contributes to thrombin generation dynamics in patients with acute or previous cerebrovascular events," *Thrombosis Research*, vol. 126, no. 4, pp. 262–269, 2010.

[54] A. Undas, M. Gissel, B. Kwasny-Krochin, P. Gluszko, K. G. Mann, and K. E. Brummel-Ziedins, "Thrombin generation in rheumatoid arthritis: dependence on plasma factor composition," *Thrombosis and Haemostasis*, vol. 104, no. 2, p. 224, 2010.

[55] A. Undas, "Fibrin clot properties and their modulation in thrombotic disorders," *Thromb Haemost*, vol. 112, no. 1, pp. 32–42, 2014.

[56] H. C. Hemker and S. Béguin, "Phenotyping the clotting system," *Thrombosis and Haemostasis*, vol. 84, no. 5, pp. 747–751, 2000.

[57] R. H. White, "The epidemiology of venous thromboembolism," *Circulation*, vol. 107, no. 23 suppl 1, pp. I–4, 2003.

[58] K. G. Mann, "Is there value in kinetic modeling of thrombin generation? Yes," *Journal of Thrombosis and Haemostasis*, vol. 10, no. 8, pp. 1463–1469, 2012.

[59] H. Hemker, S. Kerdelo, and R. Kremers, "Is there value in kinetic modeling of thrombin generation? No (unless...)," *Journal of Thrombosis and Haemostasis*, vol. 10, no. 8, pp. 1470–1477, 2012.

[60] M. F. Hockin, K. C. Jones, S. J. Everse, and K. G. Mann, "A model for the stoichiometric regulation of blood coagulation," *Journal of Biological Chemistry*, vol. 277, no. 21, pp. 18322–18333, 2002.

[61] S. D. Bungay, P. A. Gentry, and R. D. Gentry, "A mathematical model of lipid-mediated thrombin generation," *Mathematical Medicine and Biology*, vol. 20,

no. 1, pp. 105–129, 2003.

[62] M. Panteleev, N. Ananyeva, F. Ataullakhanov, and E. Saenko, "Mathematical models of blood coagulation and platelet adhesion: clinical applications," *Current Pharmaceutical Design*, vol. 13, no. 14, pp. 1457–1467, 2007.

[63] R. Wagenvoord, P. Hemker, and H. Hemker, "The limits of simulation of the clotting system," *Journal of Thrombosis and Haemostasis*, vol. 4, no. 6, pp. 1331–1338, 2006.

[64] S. Butenas, T. Orfeo, M. T. Gissel, K. E. Brummel, and K. G. Mann, "The significance of circulating factor IXa in blood," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 22875–22882, 2004.

[65] K. E. Brummel-Ziedins, T. Orfeo, P. W. Callas, M. Gissel, K. G. Mann, and E. G. Bovill, "The prothrombotic phenotypes in familial protein C deficiency are differentiated by computational modeling of thrombin generation," *PloS one*, 2012.

[66] C. M. Danforth, T. Orfeo, S. J. Everse, K. G. Mann, and K. E. Brummel-Ziedins, "Defining the boundaries of normal thrombin generation: investigations into hemostasis," *PloS one*, vol. 7, no. 2, p. e30385, 2012.

[67] C. M. Danforth, T. Orfeo, K. G. Mann, K. E. Brummel-Ziedins, and S. J. Everse, "The impact of uncertainty in a blood coagulation model," *Mathematical Medicine and Biology*, vol. 26, no. 4, pp. 323–336, 2009.

[68] F. I. Ataullakhanov and M. A. Panteleev, "Mathematical modeling and computer simulation in blood coagulation," *Pathophysiology of haemostasis and thrombosis*, vol. 34, no. 2-3, pp. 60–70, 2005.

[69] S. Jordan and E. Chaikof, "Simulated surface-induced thrombin generation in a flow field," *Biophysical Journal*, vol. 101, no. 2, pp. 276–286, 2011.

[70] K. B. Hansen and S. C. Shadden, "A reduced-dimensional model for near-wall transport in cardiovascular flows," *Biomechanics and Modeling in Mechanobiology*, vol. 15, no. 3, pp. 713–722, 2016.

[71] K. Papadopoulos, *Flow effect on thrombus formation in stenosed coronary arteries: a computational study*. PhD thesis, City University London, 2015.

[72] K. P. Papadopoulos, M. Gavaises, I. Pantos, D. G. Katritsis, and N. Mitroglou, "Derivation of flow related risk indices for stenosed left anterior descending coronary arteries with the use of computer simulations," *Medical Engineering & Physics*, vol. 38, no. 9, pp. 929–939, 2016.

[73] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*, vol. 1. Wiley New York, 1998.

[74] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer Series in Statistics Springer, Berlin, 2001.

[75] C. M. Bishop *et al.*, *Pattern Recognition and Machine Learning*, vol. 1. Springer New York, 2006.

[76] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[77] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[78] L. Breiman and A. Cutler, "Random forest-manual," 2004.

[79] J. Arumugam, S. T. Bukkapatnam, K. R. Narayanan, and A. R. Srinivasa, "Random forests are able to identify differences in clotting dynamics from kinetic models of thrombin generation," *PloS one*, vol. 11, no. 5, p. e0153776, 2016.

[80] S. Butenas, T. Orfeo, and K. G. Mann, "Tissue factor in coagulation," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 29, no. 12, pp. 1989–1996, 2009.

[81] K. E. Brummel-Ziedins, M. F. Whelihan, M. Gissel, K. G. Mann, and G. E. Rivard, "Thrombin generation and bleeding in haemophilia A," *Haemophilia*, vol. 15, no. 5, pp. 1118–1125, 2009.

[82] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, p. 620, 1957.

[83] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[84] S. Y. Park and A. K. Bera, "Maximum entropy autoregressive conditional heteroskedasticity model," *Journal of Econometrics*, vol. 150, no. 2, pp. 219–230, 2009.

[85] L. F. Shampine and M. W. Reichelt, "The matlab ode suite," *SIAM journal on scientific computing*, vol. 18, no. 1, pp. 1–22, 1997.

[86] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.

[87] H. C. Hemker and S. Beguin, "Thrombin generation in plasma: its assessment via the endogenous thrombin potential," *Thrombosis and Haemostasis*, vol. 74,

no. 1, pp. 134–138, 1995.

[88] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[89] C. Cagniart, E. Boyer, and S. Ilic, "Probabilistic deformable surface tracking from multiple videos," in *European conference on computer vision*, pp. 326–339, Springer, 2010.

[90] J. Schulman, A. Lee, J. Ho, and P. Abbeel, "Tracking deformable objects with point clouds," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 1130–1137, IEEE, 2013.

[91] S. Doraiswamy, K. R. Narayanan, and A. R. Srinivasa, "Finding minimum energy configurations for constrained beam buckling problems using the Viterbi algorithm," *International Journal of Solids and Structures*, vol. 49, no. 2, pp. 289–297, 2012.

[92] Z. Wang, A. Ruimi, and A. Srinivasa, "A direct minimization technique for finding minimum energy configurations for beam buckling and post-buckling problems with constraints," *International Journal of Solids and Structures*, vol. 72, pp. 165–173, 2015.

[93] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[94] C. E. Rasmussen, "The infinite gaussian mixture model.," in *NIPS*, vol. 12, pp. 554–560, 1999.

[95] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[96] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity*. CRC Press, 2015.

[97] W. K. Hastings, "Monte Carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[98] T. P. Minka, *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

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

[100] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[101] R. Luddington, "Thrombelastography/thromboelastometry," *Clinical & Laboratory Haematology*, vol. 27, no. 2, pp. 81–90, 2005.

[102] J. I. Weitz *et al.*, "Insights into the role of thrombin in the pathogenesis of recurrent ischaemia after acute coronary syndrome," *Thrombosis and Haemostasis*, vol. 112, no. 5, pp. 924–31, 2014.

[103] H. ten Cate and H. C. Hemker, "Thrombin generation and atherothrombosis: What does the evidence indicate?," *Journal of the American Heart Association*, vol. 5, no. 8, p. e003553, 2016.

[104] K. Brummel-Ziedins, T. Orfeo, M. Gissel, K. G. Mann, and F. R. Rosendaal, "Factor Xa generation by computational modeling: an additional discriminator to thrombin generation evaluation," *PloS one*, vol. 7, no. 1, p. e29178, 2012.

[105] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[106] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. CRC Press, 1984.

[107] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 415–434, 1963.

[108] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[109] J. G. Makin and S. Narayanan, "A hybrid-system model of the coagulation cascade: Simulation, sensitivity, and validation," *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 05, p. 1342004, 2013.

[110] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning 2nd Edition*. New York: Springer, 2009.

[111] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[112] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[113] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[114] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound

classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[115] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, ACM, 2006.

[116] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. E. John, L.-O. Wahlund, Y. Dodge, and T. Dierks, "Application and comparison of classification algorithms for recognition of alzheimer's disease in electrical brain activity (eeg)," *Journal of neuroscience methods*, vol. 161, no. 2, pp. 342–350, 2007.

[117] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, pp. 96–103, ACM, 2008.

[118] S. E. O'Bryant, G. Xiao, R. Barber, J. Reisch, R. Doody, T. Fairchild, P. Adams, S. Waring, and R. Diaz-Arrastia, "A serum protein–based algorithm for the detection of alzheimer disease," *Archives of Neurology*, vol. 67, no. 9, pp. 1077–1081, 2010.

[119] S. J. Kiddle, M. Sattlecker, P. Proitsi, A. Simmons, E. Westman, C. Bazenet, S. K. Nelson, S. Williams, A. Hodges, C. Johnston, *et al.*, "Candidate blood proteome markers of alzheimer's disease onset and progression: a systematic review and replication study," *Journal of Alzheimer's Disease*, vol. 38, no. 3, pp. 515–31, 2014.

[120] L. Breiman, "Out-of-bag estimation," tech. rep., Citeseer, 1996.

[121] L. Breiman *et al.*, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199–231,

2001.

[122] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[123] E. Lipets, O. Vlasova, E. Urnova, O. Margolin, A. Soloveva, O. Ostapushchenko, J. Andersen, F. Ataullakhanov, and M. Panteleev, "Circulating contact-pathway-activating microparticles together with factors IXa and XIa induce spontaneous clotting in plasma of hematology and cardiologic patients," *PloS one*, vol. 9, no. 1, 2014.

[124] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[125] I. J. Laurenzi and S. L. Diamond, "Monte carlo simulation of the heterotypic aggregation kinetics of platelets and neutrophils," *Biophysical Journal*, vol. 77, no. 3, pp. 1733–1746, 1999.

[126] K. Lo, W. S. Denney, and S. L. Diamond, "Stochastic modeling of blood coagulation initiation," *Pathophysiology of Haemostasis and Thrombosis*, vol. 34, no. 2-3, pp. 80–90, 2006.

[127] K. P. Papadopoulos, M. Gavaises, and C. Atkin, "A simplified mathematical model for thrombin generation," *Medical engineering & physics*, vol. 36, no. 2, pp. 196–204, 2014.

[128] A. Sagar and J. D. Varner, "Dynamic modeling of the human coagulation cascade using reduced order effective kinetic models," *Processes*, vol. 3, no. 1, pp. 178–203, 2015.

[129] J. N. Weiss, "The hill equation revisited: uses and misuses," *The FASEB Journal*, vol. 11, no. 11, pp. 835–841, 1997.

[130] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

APPENDIX A

CODE FOR THROMBIN GENERATION

Listing A.1: Thrombin Full Model Reaction Rates

```
function dC = fReaction42Rates2002(t,C)

    dC = zeros(34,1);

    k1 = 3.1E-03;
    k2 = 3.2E+06;
    k3 = 3.1E-03;
    k4 = 2.3E+07;
    k5 = 4.4E+05;
    k6 = 1.3E+07;
    k7 = 2.3E+04;
    k8 = 1.05E+00;
    k9 = 2.5E+07;
    k10 = 6.0E+00;
    k11 = 19.0E+00;
    k12 = 2.2E+07;
    k13 = 2.4E+00;
    k14 = 1.0E+07;
    k15 = 1.8E+00;
    k16 = 7.5E+03;
    k17 = 2.0E+07;
```

```
k18 = 5.0E-03;
k19 = 1.0E+07;
k20 = 1.0E-03;
k21 = 1.0E+08;
k22 = 8.2E+00;
k23 = 2.2E+04;
k24 = 6.0E-03;
k25 = 1.0E-03;
k26 = 2.0E+07;
k27 = 0.2E+00;
k28 = 4.0E+08;
k29 = 103.0E+00;
k30 = 1.0E+08;
k31 = 63.5E+00;
k32 = 1.5E+07;
k33 = 3.6E-04;
k34 = 9.0E+05;
k35 = 1.1E-04;
k36 = 3.2E+08;
k37 = 5.0E+07;
k38 = 1.5E+03;
k39 = 7.1E+03;
k40 = 4.9E+02;
k41 = 7.1E+03;
k42 = 2.3E+02;
k43 = 0;
k44 = 0;
```

```
dC(1,1) = -k2*C(1)*C(2) + k1*C(3) - k4*C(1)*C(4) + k3*C(5);
dC(2,1) = -k2*C(1)*C(2) + k1*C(3) - k5*C(5)*C(2) - k6*C(6)*C(2)
    - k7*C(7)*C(2);
dC(3,1) = -k1*C(3) + k2*C(1)*C(2);
dC(4,1) = -k4*C(1)*C(4) + k3*C(5) + k5*C(5)*C(2) + k6*C(6)*C(2)
    + k7*C(7)*C(2);
dC(5,1) = -k3*C(5) + k4*C(1)*C(4) - k9*C(5)*C(8) + k8*C(9)...
    -k12*C(5)*C(6) + k11*C(10) - k14*C(5)*C(11) + k13*C(12)...
    +k15*C(12) - k37*C(5)*C(27) - k42*C(5)*C(29);
dC(6,1) = -k12*C(5)*C(6) + k11*C(10) + k22*C(18) -
    k28*C(6)*C(22)...
    +k27*C(23) - k34*C(6)*C(26) + k33*C(27) - k38*C(6)*C(29)...
    +k43*C(13)*C(8);
dC(7,1) = k16*C(6)*C(14) + k32*C(25)*C(23) - k41*C(7)*C(29);
dC(8,1) = -k9*C(5)*C(8) + k8*C(9) - k21*C(17)*C(8) +
    k20*C(18)...
    +k25*C(18) - k43*C(13)*C(8);
dC(9,1) = k9*C(5)*C(8) - k10*C(9) - k8*C(9);
dC(10,1) = k10*C(9) + k12*C(5)*C(6) - k11*C(10) -
    k36*C(10)*C(26)...
    +k35*C(28);
dC(11,1) = -k14*C(5)*C(11) + k13*C(12);
dC(12,1) = k14*C(5)*C(11) - k13*C(12) -k15*C(12);
dC(13,1) = k15*C(12) - k19*C(16)*C(13) + k18*C(17) + k25*C(18)
    ...
    +k25*C(17) - k40*C(13)*C(29);
```

```
dC(14,1) = -k16*C(6)*C(14) - k30*C(23)*C(14) + k29*C(24);
dC(15,1) = -k17*C(7)*C(15);
dC(16,1) = k17*C(7)*C(15) - k19*C(16)*C(13) + k18*C(17) -
    k24*C(16)...
    +k23*C(19)*C(20);
dC(17,1) = k19*C(16)*C(13) - k18*C(17) - k21*C(17)*C(8) +
    k20*C(18)...
    +k22*C(18) - k25*C(17);
dC(18,1) = k21*C(17)*C(8) - k20*C(18) - k22*C(18) - k25*C(18);
dC(19,1) = k24*C(16) + k25*C(18) + k25*C(17) - k23*C(19)*C(20);
dC(20,1) = k24*C(16) + k25*C(18) + k25*C(17) - k23*C(19)*C(20);
dC(21,1) = -k26*C(7)*C(21) - k44*C(25)*C(21);
dC(22,1) = k26*C(7)*C(21) - k28*C(6)*C(22) + k27*C(23) +
    k44*C(25)*C(21);
dC(23,1) = k28*C(6)*C(22) - k27*C(23) - k30*C(23)*C(14) +
    k29*C(24)...
    +k31*C(24);
dC(24,1) = k30*C(23)*C(14) - k29*C(24) - k31*C(24);
dC(25,1) = k31*C(24) - k32*C(25)*C(23) - k39*C(25)*C(29);
dC(26,1) = -k34*C(6)*C(26) + k33*C(27) - k36*C(10)*C(26) +
    k35*C(28);
dC(27,1) = k34*C(6)*C(26) - k33*C(27) - k37*C(5)*C(27);
dC(28,1) = k36*C(10)*C(26) - k35*C(28) + k37*C(5)*C(27);
dC(29,1) = -k38*C(6)*C(29) - k39*C(25)*C(29) - k40*C(13)*C(29)
    ...
    -k41*C(7)*C(29) - k42*C(5)*C(29);
dC(30,1) = k38*C(6)*C(29);
```

```
    dC(31,1) = k39*C(25)*C(29);
    dC(32,1) = k40*C(13)*C(29);
    dC(33,1) = k41*C(7)*C(29);
    dC(34,1) = k42*C(5)*C(29);

end
```

APPENDIX B

CODE FOR SIMPLIFIED THROMBIN GENERATION

Listing B.1: Thrombin Simplified Model Reaction Rates

```
function [dy] = fch6_fThesisSimplfiedThrombinRate(t, y, K)

Ksurft = K.surf;
dy = zeros(4,1);

if (y(2) <= K.IIaThres)
    ksurf = Ksurft;
    kin = K.in2;
    kpropagation = 0;
elseif ( y(2) > K.IIaThres)
    ksurf = 0;
    kin = K.in1;
    kpropagation = K.propagation;
end

dy(1) = -ksurf - kpropagation*y(1)*y(2);
dy(2) = -kin*y(2)*y(3) + ksurf + kpropagation*y(1)*y(2);
dy(3) = -kin*y(2)*y(3);
dy(4) = kin*y(2)*y(3);

end
```