Molecular Modeling and Machine Learning Guided Target Fishing and Therapeutic Interventions against Cardiovascular Diseases

By

Rida Azhar

Fall 21-MSBI-NUST-00000362922

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(August 2024)

Molecular Modeling and Machine Learning Guided Target Fishing and Therapeutic Interventions against Cardiovascular Diseases

By

Rida Azhar

Fall 21-MSBI-NUST-00000362922

A thesis submitted to the National University of Sciences & Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Bioinformatics

Supervisor: Prof. Dr. Ishrat Jabeen

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(August 2024)

Annex A to NUST Letter No. 0972/102/Exams/Thesis-Cert dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms Rida Azhar Registration No. 00000362922 of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: _____

DR. ISHRAT JABEEN
Professor
School of Interdisciplinary Engineering & Sciences
NUST Sector H-12 Islamabad

Name of Supervisor: Prof. Dr. Ishrat Jabeen

Date: _____

For HOD Sci

Signature of HoD with stamp: _____

Dr. Mian Ilyas Ahmad
HoD Engineering
Professor
SINES - NUST, Sector H-12
Islamabad

Date: 16/07/2024

Countersign by

Signature (Dean/Principal): _____

Principal SINES

Date: 16/07/2024

AUTHOR'S DECLARATION

I Rida Azhar hereby state that my MS thesis titled "Molecular Modeling and Machine Learning Guided Target Fishing and Therapeutic Interventions against Cardiovascular Diseases" is my own work and has not been submitted previously by me for taking any degree from the National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Rida Azhar

Date: 16th Sep, 2024

DEDICATION

وما توفیقی الا باللہ

“My success is only by Allah”

(Al-Quran-11:88)

I dedicate this work to my parents for their unconditional love, care and support. Thank you for supporting me in my higher education.

Rida Azhar

(Fall 21-MSBI-NUST-00000362922)

ACKNOWLEDGEMENT

I owe my deepest gratitude and feel immense pleasure and privilege to pay my profound respect to my inspirational and encouraging supervisor *Dr. Ishrat Jabeen*. Her cooperation, kind interest, inspiring guidance, valuable suggestions, discussions and her patience empowered me to complete this research work. Her commitment, thorough understanding of the subject, and tenacity served as a fantastic source of inspiration throughout the whole tenure of MS.

My words of lifelong thanks go for my guidance committee members *Dr. Uzma Habib* and *Dr. Yusra Sajid Kiani* for their valuable suggestions during my research work. Special thanks to *Dr. Yusra* for always being there for me and helping me at each step. I would also like to acknowledge SINES, NUST and IT staff for providing me such a positive environment and infrastructure that made this research easy.

I have no words to express my love for my father *Azhar Mehmood* and my mother *Mussarat Bibi* who always wanted to see me graduate. I'm deeply indebted to them for everlasting support, love, care, encouragement, sacrifices and the hardships that they have faced to bring us to this position. Without their efforts this could never be possible. I also wanted to thank my siblings for all the love and care.

Furthermore, I would like to thank my research team, friends, and classmates for supporting and creating such a positive and learning environment in the lab.

I also want to pay my thanks to seniors, for sharing their knowledge and research experience which helped me during my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ADAMTS13	A Disintegrin-Like And Metalloproteinase with Thrombospondin type 1 repeats 13
AF	Atrial Fibrillation
ANN	Artificial Neural Network
BRN	Biological Regulatory Network
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
CRP	C-reactive protein
CSR	Cysteine-rich
CVD	CVD
Dis	Disintegrin-like
FVIII	Factor 8
GPIbα	Glycoprotein Ibα
HF	Heart Failure
IL	Interleukin
IS	Ischemic Stroke
MDTCS	Metalloproteinase, Disintegrin-like, Thrompospondin, Cystein-rich, Spacer
MI	Myocardial Infarction
MP	Metalloprotease
mRNA	Messenger Ribonucleic Acid

NETs	Neutrophil Extracellular Traps
RBCs	Red Blood Cells
RFC	Random Forest Classifier
S	Spacer
SVC	Support Vector Classifier
TSP	Thrombospondin
ULVWF	Ultra-Large Von Willebrand Factor
VWF	Von Willebrand Factor

ABSTRACT

Cardiovascular Disease (CVD) is a major global health issue, with thrombosis, or blood clot formation, as a key factor. A disintegrin-like and metalloproteinase with thrombospondin type 1 repeats 13 (ADAMTS13), a protease that cleaves the ultra-large von Willebrand factor (ULVWF), regulates thrombosis by preventing plug development. Despite its importance, ADAMTS13 regulation remains poorly understood. This study explores platelet plug formation, focusing on ADAMTS13's role in clot regulation. We developed a knowledge-driven biological regulatory network (BRN) and built four classification models using protein data from STRING and DISGENET to distinguish proteins linked to ADAMTS13, CVD, and thrombosis from those related only to CVD and thrombosis. The models, including support vector classifier, random forest, logistic regression, and ANN, were optimized using GridSearchCV. The logistic regression and ANN models showed strong performance, with the accuracy rates of 87.05% and 88.82%, respectively. The ANN model demonstrated a balanced performance between precision (83.82%) and recall (87.69%). Thrombin and plasmin were identified as ADAMTS13 inhibitors from BRN, offering insights into regulation and potential therapeutic targets. ADAMTS13 mRNA secondary structure was predicted using RNAfold, though reliability was limited by mRNA dynamics.

This study investigates ADAMTS13 regulation and its role in thrombosis and CVD, using computational approaches to deepen understanding of the molecular mechanisms. Future work should aim to probe the regulatory mechanistic of ADAMTS13, enhance classification models performance, and predict its full-length protein structure for insights into its functional mechanism.

Keywords: CVDs, Thrombosis, ADAMTS13, VWF, Machine Learning, Classification Models, Biological Regulatory Network

CHAPTER 1 INTRODUCTION

Chapter 1 provides an overview of the topic. It discusses the understudied disease and the molecules associated with it, defines the problem, and highlights the objectives of the study as well as the approaches adopted to achieve these objectives.

1.1. Cardiovascular Diseases

Cardiovascular Diseases (CVD) is a broad medical term encompassing various interconnected conditions, including coronary heart disease (CHD), cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, and venous thromboembolism [1]. It accounts for 17.9 million deaths each year around the world, and at least 2-3 times as many have non-fatal cardiovascular events [2]. The global mortality rate due to CVD has shown a concerning increase, rising from 27.9% in 2000 to 32.2% in 2019. This highlights the pressing need for innovative biomarkers to accurately predict the onset of CVD, enhance early detection, and track the effectiveness of therapy. Developing such biomarkers would be critical in reducing the burden of CVD and improving patient outcomes [3].

CVD is also defined as a group of diseases that affect blood vessels and the heart, including heart failure (HF), atherosclerosis, and cardiomyopathy. The physiological function of blood vessels is disrupted in many diseases (including CVD) due to inflammatory responses, oxidative stress, and abnormal expression of some molecules on the surface of endothelial cells [4]. Thrombosis is the most dreaded consequence of CVD and a leading cause of death around the world, making it a significant health care concern [5]. The regulation of thrombosis is a well-known biological function of a disintegrin-like and metalloproteinase with thrombospondin type 1 repeats 13 (ADAMTS13). The association between ADAMTS13 and certain CVD has been the

subject of numerous studies, all of which have reached various findings [6]. Von Willebrand Factor (VWF) is the only known physiological ADAMTS13 substrate; hence, understanding their life cycles is essential for diagnosing and treating disorders related to their malfunction [7].

ADAMTS13 is a plasma protease synthesized primarily in the interstitial area of hepatic stellate cells, but it is also produced in trace amounts in the endothelial cells, megakaryocytes, platelets, renal podocytes, and tubular epithelial cells. It is released as an active enzyme into the circulation from hepatic stellate cells and endothelial cells[8] in a closed conformation through the interaction between complement C1r/C1s, Uegf, Bmp1 (CUB) and spacer domains. Endothelial-derived ADAMTS13 responsible for the cleavage of VWF, preventing the accumulation of ultra-large VWF (ULVWF) multimers that can spontaneously interact with platelets, leading to microvascular thrombosis, thus contributing to the maintenance of a cell surface that is free of hyperadhesive VWF. When VWF binds to ADAMTS13, the protease changes its configuration and becomes activated. Although the amount of ADAMTS13 produced outside the liver is small, it may still be biologically important because of its role in thrombus formation, although more studies are needed to better understand this relationship [9].

1.1.1. Thrombosis

Thrombosis occurs when blood clots form inside blood vessels, either in veins or arteries that can partially or completely block the flow of blood, leading to a variety of health problems. The symptoms and complications of thrombosis depend on where the clot forms and how severe it is, and can range from mild discomfort to life-threatening conditions such as stroke or pulmonary embolism [10].

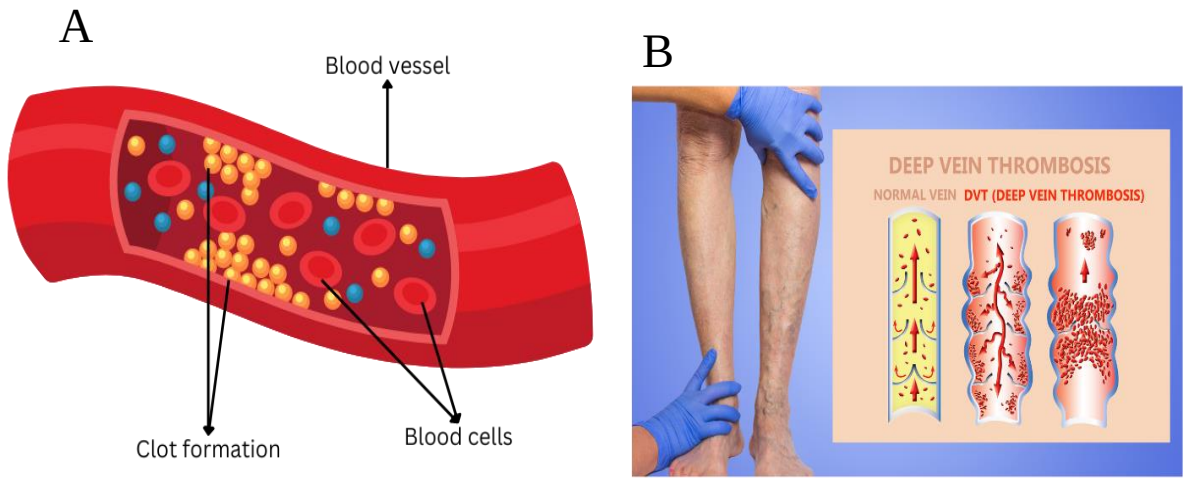


Figure 1.1: Formation of blood clot in arteries or veins resulting in decreased blood flow. (A) Clotting in blood vessels results in restricted flow of blood cells. (B) Symptoms of deep vein thrombosis is swelling of leg due to restricted flow of blood due to clotting is shown *(Scheltjens, 2023) (Abdulrahman, 2023)*

Thrombosis, while it can occur in both arterial and venous circulation, is a particularly complex disease when it occurs in the arterial system. This is because it involves inflammation of the vessel wall and the formation of an atherosclerotic plaque in the vessel, which can take years to fully develop. Once the plaque is fully developed, it becomes unstable and ruptures, triggering thrombosis and leading to acute disease. This can lead to myocardial infarction or ischemic stroke. Venous thrombosis, on the other hand, is generally caused by an imbalance between coagulation and anticoagulation and manifests itself as deep vein thrombosis or pulmonary embolism, where parts of the clot break off and block smaller veins downstream in the lung. While both types of thrombosis are serious, understanding the unique characteristics of each can help inform treatment options and improve outcomes for patients [11].

1.1.2. Pathogenesis of Thrombosis

Thrombi consist of several essential components, including fibrin, platelets, red blood cells (RBCs), leukocytes, and neutrophil extracellular traps (NETs). However, the relative importance of each component varies depending on the location of the thrombus and the underlying disease

condition. For example, the composition of a thrombus that forms in an artery may differ from that which forms in a vein, and thrombi associated with different disease pathologies may exhibit different compositions. Understanding the contributions of each component to thrombus formation in specific contexts is critical for developing effective prevention and treatment strategies [12]. One of the key components of thrombosis is platelets which are small circulatory cells that play a crucial role in coagulation. They also release various factors that aid in thrombosis and wound healing, including fibrinogen. Platelet activation and aggregation occur when collagen is exposed due to damage or injury to the endothelial layer of blood vessels [13].

1.2. Mechanism of Action of ADAMTS13

Upon vessel damage or shear stress, platelets are recruited at the injured site. The binding of platelets at the damaged site is dependent on the structural transformation of VWF. VWF is secreted in endothelial cells in a globular form that conceals the cleavage site by folding the A2 domain to protect itself from ADAMTS13 cleavage and the platelet Glycoprotein Ibα (GPIbα) binding sites in the VWF A1 domain are also buried, allowing it to circulate without platelet binding. However, following vessel injury, sub-endothelial collagen is revealed and globular VWF binds to it via its A3 domain. In response to the shear stresses generated by the flowing blood, tethered VWF undergoes a structural transformation that reveals the previously concealed GPIbα-binding sites within its A1 domains. This enables platelet capture from blood under severe shear, resulting in platelet plug development [7]. The VWF stretched conformation not only enables binding with the platelet GPIbα receptor, but also permits proteolytic attack by the zinc protease ADAMTS13, which breaks down the protein and restricts the prohaemostatic and vasoconstrictive activity of the UL-VWF by cleaving the Tyr1605-Met1606 in the A2 domain of VWF. Proteolysis mediated by ADAMTS13 transforms VWF into smaller, less active forms,

hence modulating VWF's activity. VWF's coagulation activity is diminished after cleavage, preventing clotting [14]. This mechanism is demonstrated in ***Figure 1.2***.

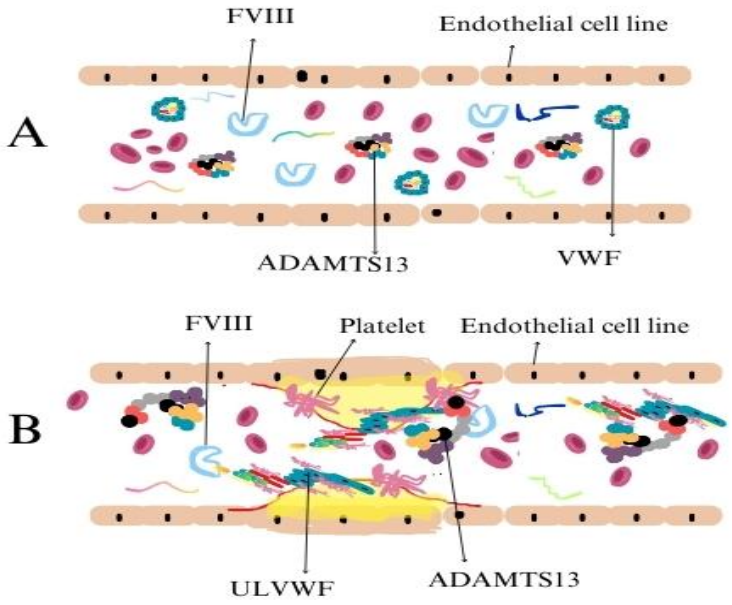


Figure 1.2: Mechanism of Action of ADAMTS13 under normal vs stressed condition (A)Normal Conditions: ADAMTS13 and VWF are present in closed conformation concealing their proteolytic domains. (B) Stress Condition: Upon vascular damage, VWF and ADAMTS13 change their conformations, VWF binds to platelets recruited at damaged site leading to ULVWF, blood flow is affected. ADAMTS13 comes in action and cleave ULVWF. FVIII enhance this cleavage.

According to research, when shear stress is applied, Factor VIII (FVIII) acts as a cofactor to speed up the processing of VWF by ADAMTS13. FVIII's propensity for binding to VWF is crucial to its rate-boosting impact. It is hypothesized that at physiological concentrations of FVIII and VWF, the selective effect of ADAMTS13 on larger VWF multimers originates from the possibility of encountering more FVIII molecules attached to the larger multimeric species [15].

ADAMTS13/VWF axis dysregulation may lead to abnormal hemostasis in the form of bleeding or thrombosis. Although this interaction is focused on shear-dependent availability of VWF or changes in ADAMTS13 circulating levels during acute or chronic disease conditions, however, ADAMTS13 proteolytic activity regulation during hemostasis remains undefined. To limit

temporal and spatial distribution of proteolytic activity in response to stimuli, extracellular proteases are often regulated by balancing rate of secretion to the rate of inhibition, but ADAMTS13 does not seem to follow these classical phenomena of protease regulation as it is secreted as an active enzyme and is resistant to natural protease inhibitors in blood [16].

1.3. Physiology of ADAMTS13

ADAMTS13 is a 37 kb gene on chromosome 9q34 that comprises 29 exons and encodes a 1427-amino acid protein [17]. It is a multidomain plasma glycoprotein with a molecular weight of 190 kDa [7] that circulates in the blood as an active enzyme at a plasma concentration of roughly 5nM after being secreted [18]. Among the 19 members of the ADAMTS protease family found in humans, ADAMTS13 has the most varied domain structure [19]. Before secretion, nascent ADAMTS13's short propeptide is proteolyzed. Secreted ADAMTS13 has metalloprotease (MP), disintegrin-like (Dis), a thrombospondin type 1(TSP1) repeat, cysteine-rich (Cys-rich), and Spacer (S) domains at the N-terminal. Seven more TSP1 repeats and two CUB domains at C-terminal fold back and interact with the core Spacer domain [7]. ADAMTS13 propeptide is not necessary for the ADAMTS13 enzyme to maintain its latency, even though propeptides are often required to maintain the latency of metalloproteases [19]. ADAMTS13 latency is governed by the noncovalent contact of the C-terminal CUB1-2 domains with the central Spacer domain, which imposes a structural constraint on the MP domain and reduces its proteolytic activity at the global level, and locally at the MP domain active site itself. Ionic interactions of the "gatekeeper trio" residues (Arg193, Asp217, and Asp252) that block the active-site cleft regulate local MP domain latency. This inhibits ADAMTS13's off-target proteolysis in circulation and probably also provides resistance to inhibition by plasma inhibitors, which explains why ADAMTS13 has a long plasma half-life (3.5 to 8 days), which is regulated by clearance instead of inhibition [20].

ADAMTS13's regulation remains poorly understood in terms of clearance and regulation. Its regulation primarily occurs at the substrate level involving VWF. Factors such as fluid shear stress, FVIII, and platelet glycoprotein 1b alpha (GP1bα)) have been identified as regulators of ADAMTS13 activity. ADAMTS13's cleavage of VWF is believed to have antithrombotic properties and may play a role in CVD [9].

ADAMTS13 is susceptible to proteolysis by coagulation proteinases. A mice model study suggest that thrombin and plasmin may inactive ADAMTS13 at site of vascular injury thus promote platelet aggregation. Thrombin proteolyzes and inactivates ADAMTS13 to promote thrombus growth. It is suggested that increased inactivation of ADAMTS13 by thrombin during the thrombotic event results in a reduction of ADAMTS13 activity. Thrombin preferably binds to thrombomodulin (TM) at the surface of uninjured endothelial cells that results in inhibition of ADAMTS13 proteolysis. TM binds Thrombin at exosite 1, this implies that exosite 1 is the possible interaction site between ADAMTS13 and thrombin. In vitro studies also confirmed that low concentration of thrombin also cleaved ADAMTS13 at multiple sites. Among these, 2 sites are significantly important, the first site appears close to the end of the protease domain and the second appears at the start of CUB domain. It is interesting to note that ADAMTS13 fragments formed after cleavage rapidly generated at cleavage sites and remain covalently associated except the 40-kDa CUB domain region. Plasmin cleaves ADAMTS13 in similar fragments as thrombin but at a faster rate. Although plasmin activates several prometalloproteinases proteolytically but it has a unique inactivating role against ADAMTS13. Since ADAMTS13 cleaves VWF, it might also suggest role of VWF in tissue repair [21].

Low levels of ADAMTS13 were observed in coronary artery disease (CAD) patients. It is assumed that consumption of ADAMTS13 antigen, which is linked to elevated VWF antigen in CAD

patients, is one probable mechanism. A statistical study found considerably greater levels of C-reactive protein (CRP) in CAD patients than those without CAD. Research indicates that inflammatory cytokine, interleukin-6 (IL-6) increases level of CRP, inhibits activity of ADAMTS13 activity under stress condition .Therefore, it can be said that low-grade inflammation may contribute to the reduction of plasma ADMATS13 levels in CAD patients [22]. Apart from IL-6, studies have also shown other inflammatory cytokines such as interferon-γ, IL-1β, IL-4 and tumor necrosis factor-α can also downregulate expression of ADAMTS13. ADAMTS13 mRNA levels are reduced by approximately 50% in hepatic stellate cells, which is primary site of expression of ADAMTS13, when exposed to interferon-γ, IL-4 or tumor necrosis factor-α [16].

ADAMTS13 deficiency, induced by gene mutations or inhibitory autoantibodies, causes UL-VWF buildup in plasma, which leads to platelet aggregation and/or thrombi amid high shear stress and microcirculatory abnormalities [23]. The contribution of ADAMTS13 in thrombosis is most evident in individuals with TTP, a disorder caused by a significant deficit of ADAMTS13. TTP patients have thrombocytopenia and microangiopathy, which can lead to stroke and myocardial infarction(MI). Beyond TTP patients, low ADAMTS13 activity and levels within the normal range are also related with a higher risk of cardiovascular complications [24].

Studies in ADAMTS13-deficient mice demonstrate that its absence results in larger infarctions, reduced regional cerebral blood flow, diminished microvascular length, and increased blood-brain barrier permeability after stroke. In humans, a moderate reduction in ADAMTS13 activity is associated with higher risks of ischemic stroke (IS) and coronary occlusion. The Rotterdam study reveals that lower ADAMTS13 activity correlates with increased risks of ischemic stroke, transient ischemic attack, and cerebrovascular events. Lower ADAMTS13 levels are found in patients with acute ischemic stroke and chronic cerebrovascular disease, while higher VWF/ADAMTS13 ratios

are linked to more severe strokes. Decreased ADAMTS13 activity is also observed in CVD patients, indicating an elevated risk of CAD and cardiovascular mortality. These findings emphasize the need for clinical trials to explore new prognostic tools and treatment options for individuals with CVD and CAD [9].

1.4. Problem Statement

The most dreaded CVD consequence and the leading cause of mortality globally is thrombosis, which poses a significant health care problem. Antithrombotic methods are efficient at stopping bleeding because they target platelets and the coagulation cascade. ADAMTS13 regulates the prothrombotic effect [25] and therefore a novel target to prevent thrombosis and reduce CVD risk. Previous studies have proved that low circulating ADAMTS13 levels that are out of proportion and high plasma VWF levels are risk factors for the onset of inflammatory and CVDs, such as MI, preeclampsia, and IS [26]. Changes in VWF and ADAMTS13 levels have also been linked to atrial fibrillation (AF) [27].

The underlying mechanism of low level of ADAMTS13 in CVDs is not discovered yet, but studies hypothesize different behavior of ADAMTS13 associated with different types of CVDs. Evaluating the role of ADAMTS13 in thrombosis and CVD by network and pathway studies and identifying the activators of ADAMTS13 can be a potential therapeutic approach to treat CVD.

1.5. Objectives

1. Knowledge-driven Biological Regulatory Network (BRN) to probe activators and inhibitors of ADAMTS13 and understand regulatory mechanism of ADAMTS13 in CVD.

2. To develop classification models to distinguish between proteins associated with ADAMTS13 and are involved in CVD and thrombosis from those only involved in CVD and Thrombosis.
3. Molecular dynamics of the ADAMTS13 mRNA structure to probe the structural integrity of ADAMTS13.

CHAPTER 2 LITERATURE REVIEW

Chapter 2 presents a comprehensive literature review on the topic "Molecular Modeling and Machine Learning Guided Target Fishing and Therapeutic Interventions against CVDs." The chapter begins with a discussion on target identification and is followed by a thorough examination of statistical and experimental evidence. It further explores the structural, conformational, and regulatory mechanisms of targets, and discusses the proposed solutions that led to the identification of gaps in existing studies and their limitations.

2.1. Target Identification

In CHAPTER 1, we have discussed thrombosis as one of the severe types of CVD and identified ADAMTS13 as a key molecule that regulates this process. It is a plasma protease that moderates the activity of VWF by cleaving VWF multimers, thus preventing the accumulation of ULVWF multimers. This prevents the spontaneous interaction of ULVWF with platelets and the subsequent occurrence of microvascular thrombosis. Therefore, ADAMTS13 is known to possess antithrombic properties [24]. Inflammatory-induced endothelial activation can release VWF into circulation, leading to ADAMTS13 consumption and a moderate reduction in its activity. This phenomenon is observed in various thrombotic disorders like myocardial infarction and ischemic stroke. Ongoing research has focused on ADAMTS13 protease, exploring potential therapeutic options for TTP patients [30]. Beyond TTP patients, several studies have shown that low ADAMTS13 activity and levels within the normal range are also related to a higher risk of cardiovascular complications [17, 24, 31-34]. Additionally, a more comprehensive understanding of ADAMTS13's role in maintaining vascular homeostasis offers potential treatments for individuals with inflammatory and thrombotic disorders, characterized by only a moderate decrease in ADAMTS13 [30].

ADAMTS13 known as primary regulator of VWF, does not follow canonical mechanism of protease regulation that are often governed by balancing activation rate and inhibition rate. It is secreted as an active enzyme and resistant to natural inhibitors. It cleaves VWF under shear stress whenever its scissile bond is exposed. Disruption of ADAMTS13-VWF axis results in abnormal homeostasis leading to bleeding or thrombosis, but the mechanism behind VWF recruiting platelets in the presence of ADAMTS13 remains unclear. The excess stoichiometry of VWF to ADAMTS13 and the mechanosensitive nature of VWF platelet-binding activity may play a role. However, this mechanism may not account for dynamic changes in early clot development, leaving ADAMTS13 potentially unregulated protease in the cardiovascular system. Downregulation of ADAMTS13 at vessel injury sites might contribute to VWF-platelet string formation and the initial clot formation stages [16].

2.1.1. Experimental Evidence of ADAMTS13 role in CVD

To evaluate role of ADAMTS13 in pathophysiology of CVD, various experiments are performed on animal models. When compared to wild-type (WT) mice, ADAMTS13 knockout (ADAMTS13-/-) animals had significantly greater infarctions after middle cerebral artery blockage [30, 35]. Another study with similar findings also noted progressively lower regional cerebral blood flow in ADAMTS13-/- mice relative to WT mice after reperfusion following an artificial middle cerebral artery closure [30, 36]. VWF is released from Weibel-Palade bodies in endothelial cells during ischemia, emphasizing the importance of ADAMTS13 function not only before but also after the event. Another animal model revealed a significant reduction in microvascular length and area, as well as impaired perfusion, lower endothelial proliferation and neovascularization, 14 days following stroke in ADAMTS13-/- mice. In addition, their blood-brain barrier permeability increased by 82%. The mechanism underlying this process may be related to

angiopoietin-1 and galectin-3 downregulation, as well as a decrease in VEGFR-2 phosphorylation in ADAMTS13-/- animals [30, 37]. These investigations found unusual findings in ADAMTS13-/- mice, whereas mice lacking VWF behaved identically to WT mice, showing that ADAMTS13's participation in stroke is dependent on its action on VWF [30]. In experimental models of stroke [38, 39] and MI [40], ADAMTS13 was induced and proved to be beneficial [16].

Chauhan, A. K. et.al had showed spontaneous formation of thrombus using intravital microscopy in venules and arterioles of Adamts13-/- mice and strong implication of natural antithrombotic activity of ADAMTS13, and suggested that recombinant human ADAMTS13 could be employed to treat TTP and possibly other thrombotic diseases [41].

2.1.2. Statistical Evidence of ADAMTS13 role in CVD

In a case-control study involving 85 healthy volunteers, 104 patients with acute ischemic stroke, and 112 patients with chronic cerebrovascular disease, lower levels of ADAMTS13 antigen were observed in those with acute ischemic stroke and chronic cerebrovascular disease compared to healthy volunteers (82.6 + 21.0% vs. 99.6 + 24.5%, $P < 0.0001$; $P < 0.03$, respectively, vs. 110.6 + 26.9%, $P < 0.0001$). Additionally, a higher VWF/ADAMTS13 ratio correlated with increased stroke severity [30, 42].

Patients with neurovascular disorders and CVD experience a moderate decrease in ADAMTS13. A study involving 27 individuals undergoing emergency cardiac catheterization for ST-elevation myocardial infarction demonstrated a significant reduction in ADAMTS13 activity (median, 75%). Notably, the study found lower ADAMTS13 activity in coronary blood compared to peripheral blood. These findings suggest that an acquired intracoronary ADAMTS13 deficiency plays a pathogenic role in acute coronary syndrome, indicating the retention of VWF at the site of acute coronary occlusion. The study also suggests that a decreased ADAMTS13/VWF ratio in

coronary flow promotes highly adhesive VWF multimers, contributing to platelet adhesion and agglutination, ultimately leading to coronary occlusion. Histologic examination of coronary thrombi from patients with acute myocardial infarction revealed co-localization of VWF with platelets, aligning with animal models showing that ADAMTS13 deficiency exacerbates VWF-dependent thrombus formation on disrupted plaques and impacts resulting infarct size [43]. The Rotterdam study investigated the correlation between ADAMTS13 activity and the risk of CAD. Among approximately 6000 individuals tracked for a median of 9.7 years, CAD occurred in 456 cases. Notably, 156 cases were in the group with ADAMTS13 activity in the lowest quartile (mean, 70.5%), while only 85 cases were in the group with ADAMTS13 activity in the highest quartile (mean, 114.3%). This indicates a 42% increased risk of developing CAD with moderately deficient ADAMTS13 activity. Subsequent analysis revealed a connection between decreased ADAMTS13 activity and a higher cardiovascular mortality rate [44, 45]. These findings should serve as the foundation for clinical trials that could lead to novel prognostic tools and therapy alternatives for patients with CAD and CVD [30].

In a study of younger (aged below 45) patients with CVD (CAD, IS, pulmonary artery disease (PAD)), those with low ADAMTS13 levels had five times the risk of CVD as those with normal levels. The association was stronger in CHD patients. People with low ADAMTS13 and high VWF are at risk CVD [46]. Changes in VWF and ADAMTS13 levels have also been linked to AF[27]. It is also proven from previous research that low circulating ADAMTS13 levels that are out of proportion and high plasma VWF levels are risk factors for the onset of inflammatory and CVDs, such as MI, preeclampsia, and IS [26].

2.2. Dynamics of ADAMTS13

ADAMTS13 gene C9ORF8, is a 37kb gene located on chromosome 9q34 containing 29 exons encodes a precursor protein of 1,427 aa [17] synthesized primarily in liver [47] and expressed in hepatic stellate cells [8]. ADAMTS13 is also synthesized in limited quantities in vascular endothelial cells, megakaryocytes, platelets, glomerular podocytes, and glial cells. However, the significance of these sources in physiological terms is yet to be determined. Given their extensive surface coverage, endothelial cells have the potential to make a substantial contribution to the plasma levels of ADAMTS13. Platelets are specifically directed to sites of vascular injury, where they undergo activation and degranulation, releasing granular contents, including VWF, which are known to be prothrombotic and proinflammatory. Consequently, the simultaneous local release of even small quantities of active ADAMTS13 protease may exert profound inhibitory effects on both thrombosis and inflammation. ADAMTS13 interacts with soluble VWF as well as anchored ULVWF. The enzyme's cleavage of ULVWF is influenced by fluid shear stress and binding of platelet GP1b and coagulation FVIII by accelerating VWF proteolysis. It is suggested that these factors alter and destabilize the VWF-A2 domain. The enzyme's structure-function analysis reveals that specific domains are crucial for substrate recognition and cleavage efficiency [47].

ADAMTS13 is likely the only protease in the cardiovascular system that lacks a regulatory mechanism. However, its role in cardiovascular system is adamantly linked to VWF. Dysregulation of ADAMTS13/VWF axis results in abnormal homeostasis in form of bleeding and thrombosis. Much of the research on this interaction has been on the shear dependent availability of VWF that is required for proteolysis by ADAMTS13, as well as the alterations in circulating levels of ADAMTS13 during acute or chronic diseases. However, it is unknown how ADAMTS13 proteolytic activity is regulated during hemostasis [16].

ADAMTS13 is secreted as an active enzyme and unlike extracellular proteases it is resistant to natural protease inhibitors found in blood [16]. It is the only member of the 19 members of the ADAMTS family having a recognized function in blood plasma. The severe thrombotic phenotype due to its hereditary or acquired impairment emphasizes the relevance of its physiological role. The other known family members, work predominantly in a tissue environment with substrate targets that are generally connected to matrix and cartilage homeostasis. Many of the ADAMTS family proteases' proteolytic substrates are unknown, however, they all have a preference for large, typically multimeric structural glycoproteins [7].

ADAMTS13 is secreted in blood in an active form and plasma concentration is ~ 1μg/ml [21]. Recent reports indicate that the half-life following plasma infusion is between 3.4 to7.9 days(82.6h-189.5h), with a median of 5.4 days(130h) [48], while recombinant human ADAMTS13 has a half-life of 2 to 3 days [49]. These observations indirectly suggest that ADAMTS13 is resistant to natural protease inhibitors found in plasma, which are known to inhibit other ADAMTS proteases [16]. There is also direct evidence of ADAMTS13's resistance to various inhibitors, including alpha-2 macroglobulin, the 4 tissue inhibitors of metalloproteinases (TIMP) isoforms, and small molecule metalloprotease inhibitors. These findings have led to the hypothesis that ADAMTS13 may exist in a latent conformation and become activated only in the presence of its substrate, VWF. However, understanding the specific biochemical and structural characteristics of ADAMTS13 that contribute to its resistance to inhibition while maintaining its activity could potentially reveal novel mechanisms of protease regulation that have not been previously defined [7, 16, 50].

Plasma protease zymogens (for example, complement, hemostatic, and fibrinolytic proteases) are traditionally proteolytically activated on demand. Cofactors are routinely used by these enzymes

to localize or increase their proteolytic action. Furthermore, the proteases within each of these enzyme systems are temporally and spatially regulated by the actions of a wide variety of plasma inhibitors. Apparently, ADAMTS13 does not follow this method of action since (i) it is secreted in a form that appears to be constitutively active, (ii) there is no "on-demand" activation phase, and (iii) it lacks a cofactor that controls its function. ADAMTS13 has proteolytic selectivity for a single location of only one physiological substrate (VWF Tyr1605-Met1606), despite its apparent constitutive activity. ADAMTS13 appears to be resistant to all plasma inhibitors, resulting in a relatively long active plasma half-life (3-7 days) for a protease, which is therefore governed by clearance (as opposed to inhibition) [7].

The regulation of proteolysis by ADAMTS13 and VWF is dependent on exosite interactions. Normally, VWF circulates in globular shape that cannot bind platelets or be easily cleaved by ADAMTS13 due to the hidden cleavage site and exosite regions in the A2 domain. Proteases are typically made in latent forms and initially inactive, and their activation is controlled spatially and temporally. Multiple mechanisms can be used to confer latency. Enzyme allostery, proteolytic activation, and/or activating cofactors are usually required for the transformation of a latent enzyme into an active form [20]. ADAMTS13 is an intricate protein with an autoinhibited MP domain [51], and its activation involves a unique mechanism. Instead of relying on on-demand proteolytic activation, ADAMTS13's proteolytic activity is controlled by its substrate, VWF. This control involves two latency mechanisms: global latency, which involves the non-covalent interaction of C-terminal CUB1-2 domains with the central spacer domain, and local latency, where ionic interactions in the MP domain's active site regulate delay. Ionic interactions of the "gatekeeper triad" residues (Arg193, Asp217, and Asp252) that conceal the active-site cleft regulate local MP domain delay. This control limits off-target proteolysis and resistance to

inhibition, explaining ADAMTS13's long plasma half-life, which is regulated by clearance rather than inhibition [20].

N-linked, O-linked and C-linked glycans constitute 20% mass of ADAMTS13. The effect of O-linked and C-linked on ADAMTS13 proteolytic function are unclear but N-linked glycosylation and O-fucosylation have significant impact on secretion and folding of ADAMTS13[16]. N-linked glycans in ADAMTS13 CUB domains may alter Spacer-CUB interaction, according to some research [20]. In silico modeling and the crystal structure shows no importance of N-glycans of CUB domains in maintaining closed conformation of ADAMTS13. However, it is known that N-glycans and post-translational modifications can stabilize interdomain or intra domain interactions. CUB domains have sialic acid–ending N-glycosylation sites at Asn1235 and Asn1354. Asn to Gln mutations boosted ADAMTS13 activity toward the FRETS-VWF73 substrate and shear-dependent VWF processing, highlighting the relevance of these N-glycans in maintaining a closed conformation. In the absence of these N-glycans, CUB domains may become structurally less stable, facilitating CUB domain conversion from a closed to an open conformation. Negatively charged sialic acid residues on N-glycans linked to Asn1235 and Asn1354 may also alter the electrostatic surface potential of the CUB domains, allowing its binding to positively charged residues in the spacer domain, promoting a closed conformation of ADAMTS13 through long-range electrostatic interactions [52]. Beta 1,3-glucosyltransferase (B3GLCT) is also involved in the modification of thrombospondin type 1 repeats (TSRs) from N-terminal to C-terminal. Properly folded TSRs contain a site for the addition of glucose to O-linked fucose (O-Fuc), which is catalysed by B3GLCT [53]. Additionally, ADAMTS13 glycosylation may be critical in immune recognition and clearance. N-glycan introduction at K608 position reduce antibodies binding to spacer domain without altering its proteolytic function [16].

Mass spectral analysis was performed to study the effect of O-fucosylation on ADAMTS13 secretion. It was observed that at least six of the TSRs in human ADAMTS13 tryptic peptides are mutated with O-fucose disaccharide. Protein O-fucosyltransferase 2(POFUT2) is the enzyme that transfers fucose to serine. When POFUT2 expression was reduced by siRNA, ADAMTS13 secretion decreased. ADAMTS13 expressed in a cell line unable to synthesize GDP-fucose had a comparable effect. Overexpression of POFUT2 did not influence wtADAMTS13 secretion, but it did increase ADAMTS13 TSR1,2 double mutant secretions. O-fucosylation is functionally important for ADAMTS13 secretion [54].

2.3. Structure of ADAMTS13

ADAMTS13 5kb mRNA encodes a precursor protein of 1,427 aa [55] with a modular structure consisting of a signal peptide, and a propeptide (P) that are cleaved prior to secretion. The N-terminal comprises of catalytic metalloproteinase (MP) domain and non-catalytic disintegrin-like (Dis), a thrombospondin type 1 (TSP1) repeat, cysteine-rich (CSR), and Spacer(S) domain collectively known as MDTCS/MDT1CS domains. Seven additional TSP1 repeats and two CUB domains found at C-terminal. The C-terminal domains fold back and interact with central spacer domain [56]. The cartoonist representation of ADAMTS13 domains is shown in ***Figure 2.1***.

ADAMTS13 itself circulates in a latent form with two levels of control over its latency: first, "globally" through a noncovalent interaction between the C-terminal CUB1-2 domains and the central Spacer domain, which restricts the proteolytic function of the MP domain, and second,

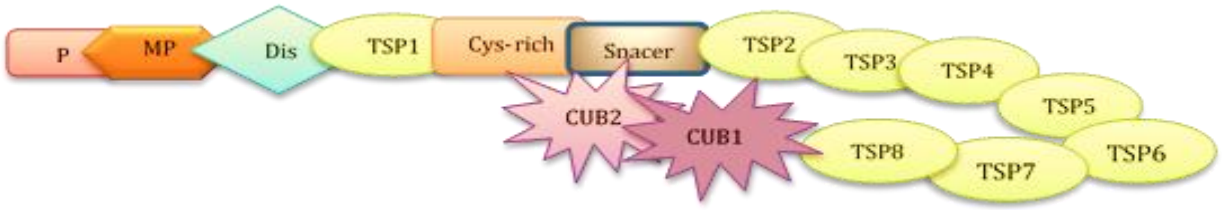


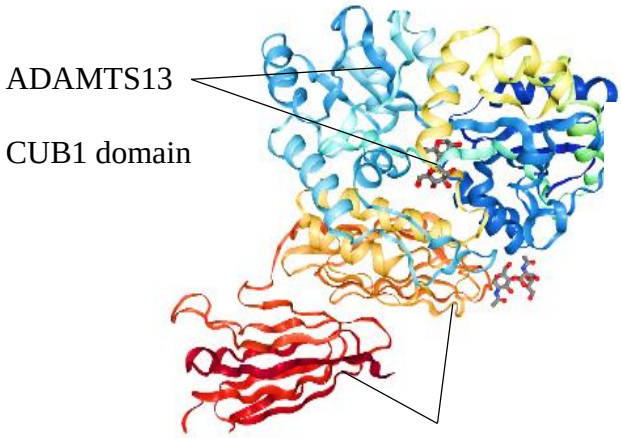
Figure 2.1 : Schematic Illustration of ADAMTS13 domains. The N-terminal contains catalytic domains and the C-terminal contains non-catalytic domains and CUB1/2 is shown to interact with central spacer domain to keep the structure intact.

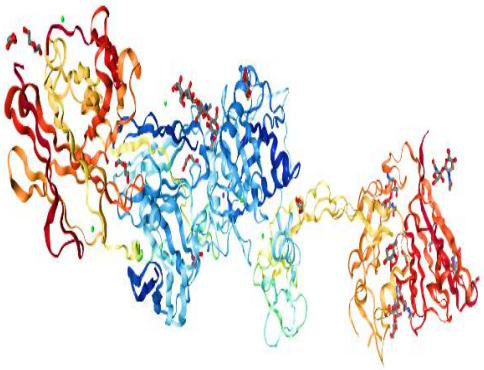
"locally" within the MP domain's active-site. The local MP domain latency is maintained by ionic interactions of the "gatekeeper triad" residues (Arg193, Asp217, and Asp252) that block the active-site cleft. This mechanism prevents non-specific proteolysis by ADAMTS13 in the bloodstream and makes it resistant to inhibition by plasma inhibitors, explaining its extended plasma half-life (3.5 to 8 days), which is regulated by clearance rather than inhibition [57].

The crystal structures of ADAMTS13 is partially resolved spanning only few domains listed in ***Table 2.1*** and shown in ***Figure 2.2***.

Table 2.1 Co-crystallized structures of non-catalytic DTCS and CUB 1/2 domains and catalytic MP domain of ADAMTS13 from Protein Databank

S.No	PDB ID	Domain	Method	Resolution (Å)	Year	Publication
1.	7B01	CUB1/2	X-RAY DIFFRACTION	2.90	2021	[20]
2.	6QIG	MP		2.80	2019	[56]
3.	3GHN	DTCS		2.80	2009	[7]
4.	3GHM	DTCS		2.60		

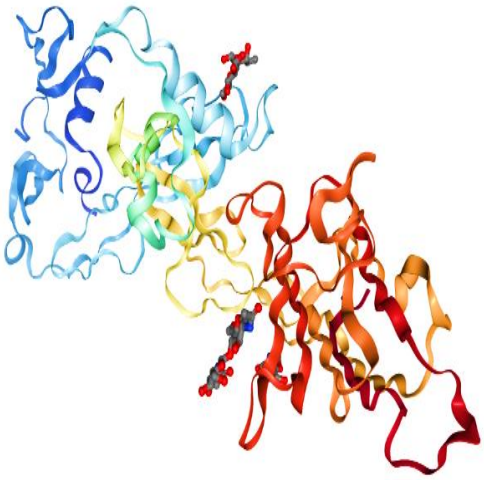




7B01: ADAMTS13 CUB1/2

6QIG: ADAMTS13 metalloprotease

3GHN: ADAMTS13 DTCS domains revealing exosites for VWF interactions (Form 2)



3GHM: ADAMTS13 DTCS domains revealing exosites for VWF interactions

Figure 2.2: Co-crystallized structures of ADAMTS13 from PDB

The homology model of MP domain with Dis domain found that MP domain is crucial for VWF cleavage which is the only substrate of ADAMTS13. The MP domain contains 3 histidine residues for coordination of a Zn^{2+} ion in the sequence HEXXHXXGXXH inside catalytic cleft which interacts with active site residue, Glu225 to facilitate substrate cleavage. Moreover, the conserved Met249 residue of MP domain turns the backbone of MP domain near active site towards the C-terminus HEXXHXXGXXH sequence (224-234 residue). The Met-turn also provides hydrophobic environment for the 3 active site histidine residues and the corresponding Zn^{2+} ion. This coordination gives rise to term "metzincins" to denote highly conserved enzyme class that contains Met-turn/ Zn^{2+} motif. Further analysis of MP domain also suggested the presence of three putative Ca^{2+} binding sites. A calcium cluster formed at site 1 and site 2 contains the well conserved residues Glu83, Asp166 Asp168, Cys281 and Asp284 in ADAMTS1, ADAMTS4 and ADAMTS5 as well, constituting a connector loop. The side chains of Asp182 and Glu212 orients the site 3 which is also conserved in above mentioned proteases. Leu183, Arg190 and Val192 contribute to the stabilization of Ca^{2+} ions. Site 3 includes a conjunction of Asp182, Glu184 and Asp187 and Glu212. One study reported that mutations at this site reduce Ca^{2+}induced ADAMTS13 activity dramatically, suggesting site 3 residues play an important role in high-affinity Ca^{2+} binding and proteolytic activity. Mutation at these residues, excluding Asp182, to alanine predicted at site 3 reported reduced substrate activity of ADAMTS13. The calcium binding loop closure is mediated by a salt bridge between Glu184 and Arg190. The MP domain is predicted

to contain a functionally important Ca^{2+} binding site adjacent to the active-site, and also an additional double Ca^{2+} binding site [58].

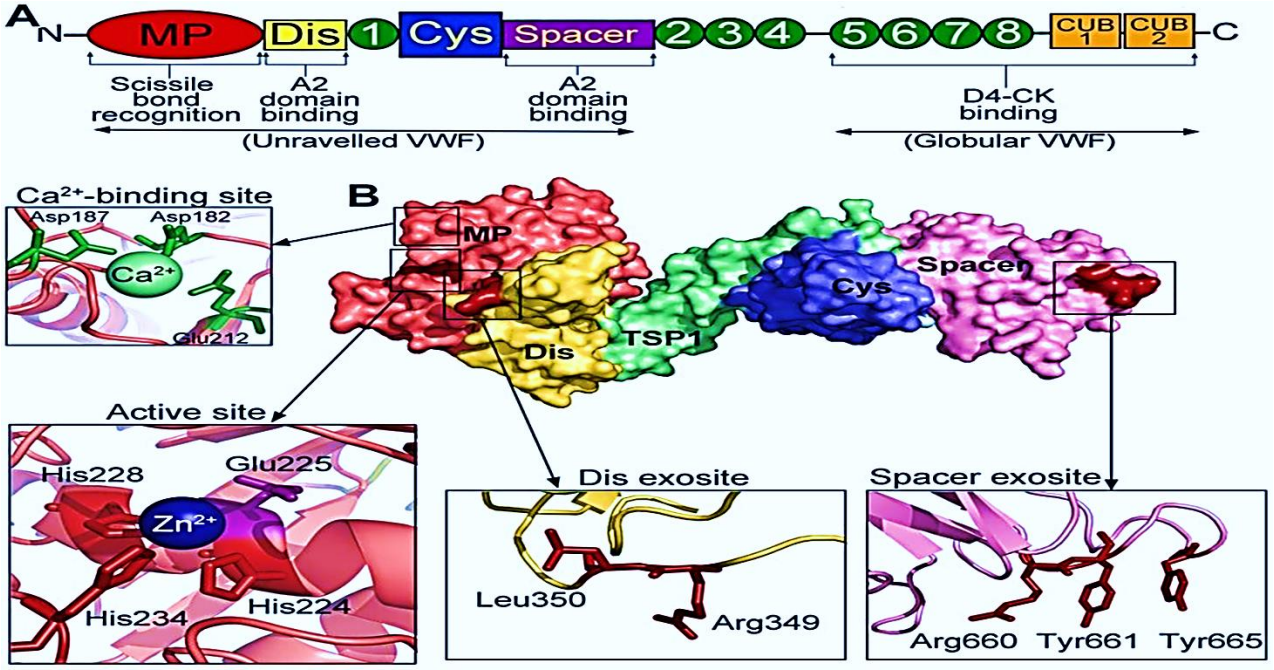


Figure 2.3: Cartoonic representation of ADAMTS13 Domains (A) Organization of Domains in ADAMTS13. (B) Surface representation of N-terminal domains(MDTCS) of ADAMTS13. Ca^{2+} binding site with interacting residues(green) and the three histidine residues in active site cleft for coordination of Zn^{2+} ion which interact with Glu225 in MP domain are shown. The linker region, known as the Dis exosite, extends behind the MP domain and situates the Dis domain on one side of the active-site cleft. Spacer domain (Lilac) containing exosite essential for cryptic exosite binding of VWF A2 domain (Crawley, 2011)

The MP domain of ADAMTS13 is insufficient for recognition and targeted breakdown of VWF, whereas an MDT1CS domain fragment achieves full VWF-cleaving activity in vitro [14]. The Dis and Cys-rich domains share a common topology consisting of α-helix, 3-4 irregular loops, and two antiparallel strands. The Dis domain (47 aa) is shorter than the Cys-rich domain (67 aa), which has a 310 α-helix between the L2C and L3C loops. Although the Dis and Cys-rich domains share a similar structure, their loop configurations and sizes differ, and they do not superpose well (RMSD

= 2.54 Å, 52 Å aligned residues). When the Dis and Cys-rich domains were superposed, the previous DTCS structure yielded an RMSD value of 1.82 Å (54 aligned residues). Loop L1D (also known as V-loop), which is likely part of the Dis domain exosite, and three loops in the Cys-rich domain are incompletely understood in the electron density, indicating a high degree of flexibility in these domains [7]. The spacer domain is highly significant for the proteolytic activity of ADAMTS13. In several studies, it is reported that in the absence of the spacer domain, activity of ADMATS13 is significantly reduced towards VWF73 and full-length VWF [59]. TSP2-8 are highly flexible and enable the interaction of spacer and CUB domains and also enable folding of CUB1-2 domains on rest of molecule [20].

DeYoung et.al. utilize the recent release of the AlphaFOLD2 structure prediction of ADAMTS13 which offers new insights into the intramolecular contacts crucial for maintaining a closed conformation. This model predicts contact points between various domains, including TSP1-8 and the CSR, CUB2, the Dis domain and the calcium-binding loop in the MP domain. These inter-domain interactions suggest that when ADAMTS13 is folded in its closed conformation, certain domains align to potentially mask substrate access to the MP domain, contributing to global latency. However, discrepancies exist between these predictions and previous experimental findings, particularly regarding the binding site for the CUB domains on the spacer domain and the domain boundaries of the TSP1 repeats linking the CUB domains to the spacer domain. Therefore, while the AlphaFOLD2 model offers valuable insights, caution is necessary in its interpretation. Nevertheless, this structure simulation may inspire new hypotheses for investigating ADAMTS13 conformational dynamics and its regulatory mechanisms [16].

The CUB acronym refers to complement C1r/C1s, Uegf (epidermal growth factor-related sea urchin protein), and Bmp1 (bone morphogenetic protein 1). Among ADAMTS proteases, only

ADAMTS13 contains CUB domains. These homologous domains consist of 107(CUB1) and 129(CUB2) residues and are likely linked by a flexible region starting from Leu1290 to Glu1298, although their orientation remains unclear due to linker flexibility. Additionally, the CUB2 domain model indicates an unstructured region beginning at Gln1409 and ending at Thr1427 [58].

2.4. Mode of Action of ADAMTS13

ADAMTS13-mediated proteolysis depends on VWF conformational changes and its own conformational activation. Kinetic investigation of wtADAMTS13 showed 2.5-fold lower activity compared to ADAMTS13 without its C-terminal tail (MDTCS) or CUB1-2 domains, showing the CUB domains limit ADAMTS13 function. Anti-CUB mAb(20E9) or VWF D4CK preincubation increased wtADAMTS13 activity by 2.5-fold (the natural binding partner for the CUB domains). The characterization of Gain of Function (GoF) ADAMTS13 provides further proof of conformational activation of ADAMTS13. There was a 2- to 2.5-fold increase in the overall rate of cleavage of the scissile bond fragments when this variation was used in combination with VWF fragments due to their higher association rate [60].

The VWF cleaving protease, cleaves ultra-large VWF multimers (UL-VWFM) into small VWF fragments; thereby lowers VWF coagulation activities [14]. VWF is secreted by endothelial cells and platelets as disulfide-linked dimers to 100-mers. Plasma VWF adopts a conformation that conceals the cleavage site by folding the A2 domain to protect itself from ADAMTS13 cleavage and the platelet GPIbα binding sites in the VWF A1 domain is also buried, allowing it to circulate without platelet binding. However, following vessel injury, sub-endothelial collagen is revealed and globular VWF binds to it via its A3 domain. In response to the shear stresses generated by the flowing blood, tethered VWF undergoes a structural transformation that reveals the previously

concealed GPIbα-binding sites within its A1 domains. This enables platelet capture from blood under severe shear, resulting in platelet plug development [7].

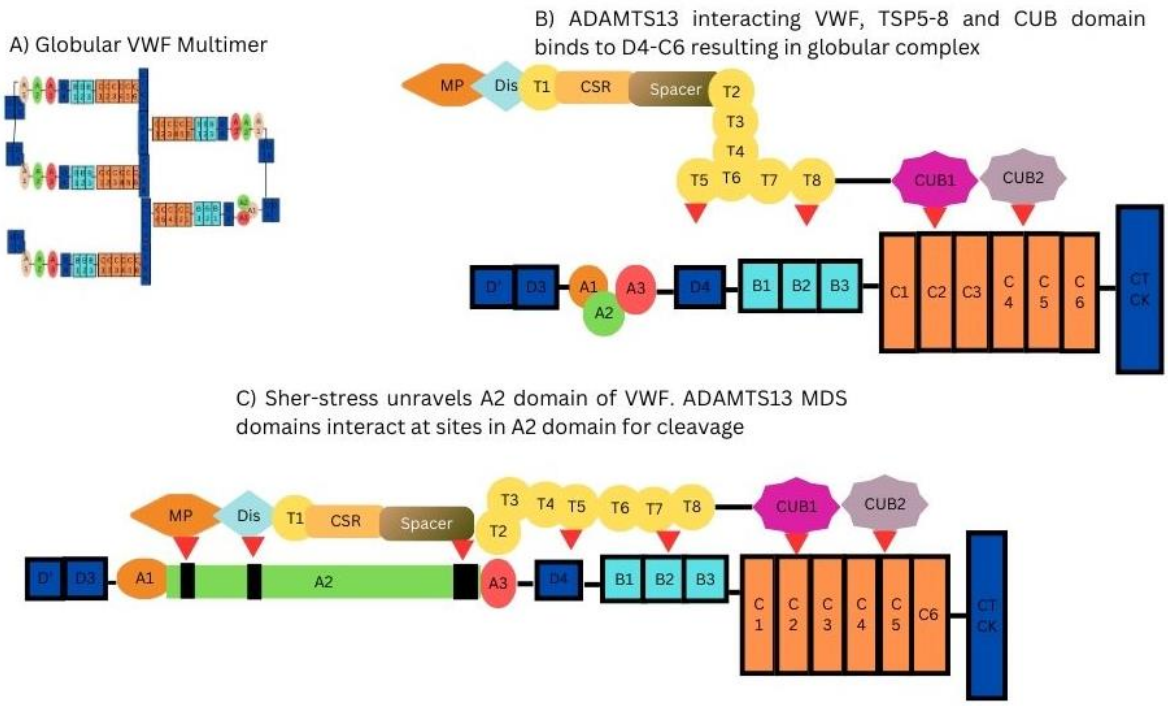


Figure 2.4 : Schematic Representation of ADAMTS13 Cleavage. (A) Globular multimeric structure of VWF. VWF monomers linked by disulfide bonds to form multimer. (B). ADAMTS13 TSP5-TSP8 and CUB domains can bind to VWF enabling formation and circulation of VWF and ADAMTS13 complex in plasma. (C). Under shear stress unravelling of VWF to expose A1 domain binding site for GPIbα and removal of molecular plug formed by the vicinal disulphide bond in the A2 domain, which causes A2 domain unfolding. This unfolding reveals cryptic exosites that enable residues in the ADAMTS13 spacer domain to bind to the unfolded A2 domain. This enables further interactions between the MP domain to occur, including an essential interaction via an S3 subsite with L1603 in VWF. Together, these interactions allow the MP to engage via S1 and S1′ subsites with the cleavage site, after which proteolysis can occur.

The VWF stretched conformation not only enables binding with the platelet GP1bα receptor, but also permits proteolytic attack by the zinc protease ADAMTS13, which breaks down the protein and restricts the prohaemostatic and vasoconstrictive activity of the UL-VWFM by cleaving the Tyr1605-Met1606 in the A2 domain of VWF [61]. Proteolysis mediated by ADAMTS13 transforms VWF into smaller, less active forms, hence modulating VWF's activity. VWF's coagulation activity is diminished after cleavage, preventing clotting [14].

According to research, when shear stress is applied, FVIII acts as a cofactor to speed up the processing of VWF by ADAMTS13. FVIII's propensity for binding to VWF is crucial to its rate-boosting impact. It is hypothesized that at physiological concentrations of FVIII and VWF, the selective effect of ADAMTS13 on larger VWF multimers originates from the possibility of encountering more FVIII molecules attached to the larger multimeric species [15].

The specific mechanism linking low levels of ADAMTS13 to CVDs remains unknown, but various studies propose different mechanisms that may be associated with specific types of CVD. Low levels of ADAMTS13 antigen and activity have been linked to the pathophysiological mechanisms of acute MI (AMI), likely because of the increased prevalence of platelet thrombi that would result from ADAMTS13 deficiency [62].

2.5. Regulation of ADAMTS13

Literature studies have not reached any conclusion regarding potential activators and inhibitors of ADAMTS13, however they suggest different regulatory mechanisms. An investigation into the impact of inflammatory cytokines on the release and cleavage of ULVWF aimed to assess potential connections between inflammation and thrombosis. The study revealed that IL-6 hindered the cleavage of ULVWF by ADAMTS13 under flowing conditions, suggesting a regulatory role for IL-6 in controlling ADAMTS13. IL8, TNF-α and to a lesser extent IL-6-sIL-6r complex induce the dose-dependent release of ULVWF from endothelial cells implying inflammatory cytokines have a role in balancing release of ULVWF and cleavage of VWF under different pathological conditions [63], however the inhibitory mechanism and its significance is uncertain under normal conditions.

ADAMTS13 expression can be decreased by inflammatory cytokines like interleukin-6 (IL-6), IL-1β, interferon-γ, IL-4, and tumor necrosis factor-α in certain cells. While hepatic stellate cells can

transition into fibroblast-like cells in response to liver injury without affecting ADAMTS13 mRNA expression. Exposure to interferon-γ, IL-4, or tumor necrosis factor-α reduces ADAMTS13 mRNA levels by approximately 50% in these cells. Given that hepatic stellate cells contribute significantly to circulating ADAMTS13 levels, these expression changes may contribute to reduced ADAMTS13 levels in patients with inflammatory conditions, although further investigation is needed. The impact of disease states on ADAMTS13 clearance mechanisms remains poorly understood, although macrophage scavenger receptor CD163 has been implicated in ADAMTS13 endocytosis in vitro. While ADAMTS13 is not directly regulated by natural protease inhibitors, its activity can be attenuated by various physiological agonists such as platelet factor 4, alpha defensins, IL-6, thrombospondin-1, and hemoglobin, which compete with ADAMTS13 for binding to VWF. Studies have shown conflicting results regarding the ability of platelet factor 4 to block ADAMTS13-mediated VWF cleavage in vitro versus in animal models, highlighting the need for further research into the local concentrations and effects of these agonists at sites of vascular injury. Similarly, although human neutrophil peptides can inhibit ADAMTS13-mediated VWF proteolysis in vitro, their physiological relevance in vivo remains uncertain due to the high concentrations required for inhibition [16].

Another study examined proteolysis as regulatory mechanism of ADAMTS13. The observation of study includes that cleavage of ADAMTS13 by thrombin, Factor Xa (FXa) and plasmin in a similar pattern but not identical. Thrombin cleaves at multiple sites preferentially near the end of MP domain and start of CUB1 domain, however the cleaved fragments are rapidly generated. In vivo, thrombin exhibits greater activity and has a higher concentration than FXa, making the proteolysis of ADAMTS13 by thrombin more likely to be of physiological significance. Plasmin, which is

well known for activating different prometalloproteinases, cleaves ADAMTS13 at a faster rate than thrombin thus has a unique inactivating role against ADAMTS13 [21].

Thrombin, a serine protease generated in response to vascular damage, plays significant role in coagulation, inflammation and cellular proliferation. In coagulation cascade, it acts as both procoagulant and anticoagulant. Following vascular injury, tissue factor (TF) and cell surface receptors are exposed, initiating a series of cascade leading to thrombin generation. Thrombin binds with Na^+ to adopt procoagulant form to convert fibrinogen to fibrin, and activates FXIII to FXIIIa to stabilize fibrin to fibrin clot in the presence of Ca^{2+}. Then, thrombin feeds back to activate FV to FVa and VWF bound FVIII to FVIIIa in order to amplify coagulation cascade and accelerating FXa and thrombin generation respectively. On platelets surface, thrombin binds to GP1ba to activate FXI to FXIa and protease activated receptors (PAR1) and GpV resulting in platelet activation. FXI activates FIX, FX and prothrombin accelerating thrombin generation level. The high levels of thrombin convert fibrinogen to fibrin in an unregulated way expanding the plug to undamaged endothelium area. Thrombin enhance platelet adhesion by cleaving ADAMTS13, thus inactivating it. The binding stability of thrombin with Na^+ change with time and thrombin reverses to free form. At the undamaged endothelium site, it binds to thrombomodulin (TM), activating thrombin-activatable fibrinolysis inhibitor(TAFI) and redirecting thrombin functioning to anticoagulant. TM also direct thrombin to activate protein C. Protein C is an anticoagulant protein which proteolytically inactivates FVa and FVIIIa. As mentioned earlier, thrombin binds with GP1ba at site of injury, but when fibrin clot expands to non-damaged endothelial sites, glycosaminoglycan cofactors binds with thrombin to inhibit it. To make this interaction irreversible, thrombin cleaves antithrombin, acting as anticoagulant. Low levels of thrombin

enhance plasmin formation which accelerate fibrin cleavage clearing fibrin and helping in injury healing [64].

Vascular endothelial damage not only results in thrombin generation but also triggers the release of proinflammatory cytokines, including tumor necrosis factor (TNFα), interleukins, and tissue plasminogen activator (tPA). Plasminogen activators such as tPA and urokinase plasminogen activator (uPA) convert inactivated plasminogen to active plasmin, with fibrin serving as a cofactor for this activation [65]. TAFI plays a role in inhibiting fibrinolysis by preventing plasminogen coactivation by fibrin [66]. Additionally, plasminogen activation is tightly regulated by plasminogen activator (PA) inhibitors, particularly PA inhibitor-1 (PAI-1), which inhibits both tPA and uPA. Plasmin activity is also directly countered by α2-antiplasmin in plasma. This intricate regulatory system highlights the broader significance of plasminogen activation in various biological processes [67].

Plasmin exhibits a wide-ranging substrate specificity beyond fibrin cleavage, targeting various elements and inhibiting their activity such as coagulation factors V, VIII, IX, X, tissue factor pathway inhibitor (TFPI), fibrinogen and ADAMTS13, however the cleavage site for ADAMTS13 is unknown. Plasmin has little tendency towards VWF that depends on various factors including shear induced unfolding presence of N-linked glycans, desialylation and modulation via ABO(H) blood group [68].

ADAMTS13 demonstrates remarkable substrate specificity, setting it apart from other thrombolytic agents like tissue plasminogen activator (tPA) and its derivatives. While tPA's activity can lead to systemic plasmin generation and subsequent bleeding due to its lack of specificity, ADAMTS13 shows promise in thrombosis treatment without causing excessive bleeding. Studies in animal models suggest that ADAMTS13 infusion can enhance ischemic stroke

recovery when combined with low doses of tPA, mitigating the risk of bleeding. However, gaps remain in understanding how ADAMTS13 activity is regulated in vivo, crucial for optimizing recombinant ADAMTS13 therapies for hereditary TTP and other thrombotic conditions [16].

Autoantibodies also play a significant role in inhibiting ADAMTS3's activity. Anti-ADAMTS13 autoantibodies in acquired TTP predominantly target the spacer domain, specifically exosite 3, hindering substrate binding and inhibiting ADAMTS13's proteolytic activity. Attempting to retain ADAMTS13's functionality while eliminating autoantibody binding, researchers modified these exosite 3 residues, exchanging Arg for Lys or vice versa. Their site-directed mutagenesis produced ADAMTS13 variants M4 and wih following mutaions respectvely: R660K/F592Y/R568K/Y661F and R660K/F592Y/R568K/Y661F/Y665F) with significantly increased specific activity toward FRETS-VWF73 and multimeric VWF, approximately 4- to 5-fold and 10- to 12-fold, respectively, compared to the wild-type. These engineered variants were also more resistant to inhibition by anti-ADAMTS13 autoantibodies from acquired TTP patients, correlating with reduced IgG binding to the variants. These findings suggest the potential for re-engineering ADAMTS13 to enhance specific activity in the presence of autoantibodies, providing therapeutic benefits for acquired TTP patients [8].

2.6. Structural dependent functioning of ADAMTS13

ADAMTS family, which is commonly known for their specificity towards large and often complex multidomain glycoproteins. ADAMTS family proteases typically have ~200 aa residues propeptide that act as molecular chaperones and provide assistance in protein folding and possess a "cysteine-switch" which aids in maintain the latency of the protease. However, ADAMTS13 has short propeptide (45 residues) that is not required for secretion or activation, therefore removed prior to secretion [8].

The MP domain of ADAMTS13 alone is not sufficient for recognition and specific cleavage of VWF, but full VWF-cleaving activity is achieved in vitro with an MDTCS or MDT1CS domain fragment [69]. Furthermore, it's been observed that antibodies obtained from patients with idiopathic TTP often hinder the activity of ADAMTS13 by attaching to the noncatalytic domains of the enzyme, particularly the proximal C-terminal domains encompassing DT1C and S (referred to as ADAMTS13-DTCS). These findings collectively underscore the significance of these noncatalytic domains in the recognition of VWF [56].

Much like numerous other clotting factors, it seems that exosite interactions play a pivotal role in controlling the enzymatic activity and substrate selectivity of ADAMTS13. The metalloprotease domain of ADAMTS13 alone exhibits minimal or no proteolytic activity when acting on VWF substrates. However, as more non-catalytic domains are progressively incorporated into the metalloprotease domain, the proteolytic activity significantly escalates. This is supported by a model of MDTCS fragments based on the crystal structure of ADAMTS13-DTCS and the metalloprotease domains of ADAMTS1, ADAMTS4, and ADAMTS5, suggesting that the metalloprotease and disintegrin-like domain (MD) function as an inseparable unit. Experimental data also confirm that the addition of the disintegrin-like domain to the metalloprotease domain substantially enhances both the efficiency and specificity of cleavage [8]. Disintegrins are 47-84 aa short polypeptides that belong to platelet aggregation inhibitor family having RGD sequence found on conserved hairpin loop which allows for high binding affinity to platelet surface protein GPIIb/IIIa that is commonly found in snake venom proteins. ADAMTS13 Dis domain do not share these structural features. It lacks tripeptide RGD, however, other conserved sequences are present. It is suggested that it might be due to structural evolution to function differently compared to other disintegrins. It enhances substrate binding of ADAMTS13 in combination with cysteine-rich and

spacer domains. The MP-Dis complex has a higher proteolytic activity. Various studies on mutation of ADAMTS13 concluded reduced or diminish proteolytic activity of ADAMTS13. The functional study of exosites on ADAMTS13 surface found that single point mutation R349D in Dis domain results in complete loss of activity [58].

Table 2.2: List of reported mutations in ADAMTS13 structure and their effects on its activity

S. No	Mutations	Effect on ADAMTS13 proteolytic activity
1.	Arg349Asp	Almost complete loss of activity
2.	Arg349Ala, Leu350Gly	10-20 fold reduction
3.	Val352Gly	4-5 fold reduction
4.	Arg349Cys,Pro353Leu	Reduce by 99.5% and 97% respectively
5.	Arg349Ala, Leu350Gly	Dramatically reduced proteolytic activity toward both peptide substrates and multimeric VWF
6.	Trp390Cys	Disrupt WXXW motif and impairs both the stability of TSR1 domain and C-mannosylation.
7.	Arg1060Trp	Diminishes the protein secretion resulting in ADAMTS13 deficiency.
8.	Arg659Ala, Arg660Ala, Tyr661Ala, and Tyr665Ala.	Impairs ADAMTS13 recognition by VWF.

9.	Arg568Lys,Phe592Tyr, Arg660Lys, Tyr661Phe, Tyr665Phe	Increased activity towards VWF.
10.	Arg349Cys and Pro353Leu	Affect enzymatic activity,
11.	Pro618Ala	Reduces secretion efficiency in cultured cells

The disintegrin-like domain plays a crucial role in presenting the VWF scissile bond to the active site and forms a significant interface with the metalloprotease domain. Without the disintegrin-like domain, proteolytic activity is undetectable, indicating that the MP domain alone lacks sufficient binding activity for VWF cleavage. Specific residues within the disintegrin-like domain, particularly those from P317 to L354, are instrumental in positioning VWF within the MP domain for proteolysis. Mutations at residues R349, L350, or V352 lead to a significant reduction in ADAMTS13's catalytic efficiency towards VWF. Residues L350 and V352 likely interact with A1612 on VWF, while an ionic interaction occurs between residue R349 in the disintegrin domain and D1614 on VWF [16]. Further investigations have highlighted the significance of specific residues, particularly Arg349 and Leu350 within the ADAMTS13 disintegrin domain, which interact with residues Asp1614 and Ala1612 in the central A2 domain of VWF. These interactions play a crucial role in aligning the Tyr1605-Met1606 bond with the active-site cleft, significantly affecting the rate constant (Km) and catalytic efficiency (kcat) during substrate proteolysis. Recent evidence suggests that VWF residues Leu1603, Tyr1605, and Asp1614 directly contact Leu198, Val195, and Arg349 in ADAMTS13, respectively. Thus, the disintegrin domain, along with other non-catalytic domains, ensures proper alignment of the scissile bond over the active center, facilitating the cleavage process [8]. The mutations and their effects are listed in ***Table 2.2***

The TSRs found large number of proteins in extracellular matrix containing a conserved tryptophan residue that are solvent exposed to facilitate ligand interaction [58]. The role of TSRs in ADAMTS13 remains unclear. The first TSP1 repeat of ADAMTS13 binds directly to VWF73 with a Kd of approximately 136 Nm. The TSP1 5–8 repeats of ADAMTS13 appear to bind native VWF through the D4 domain. Additionally, the C-terminal TSP1 repeats of ADAMTS13 interact with the endothelial cell surface receptor CD36, potentially enhancing proteolytic cleavage of UL-VWF under flow conditions. However, human and murine ADAMTS13 variants lacking the C-terminal 2–8 TSP1 repeats and CUB domains MDTCS cleave cell-bound UL-VWF and soluble VWF with similar efficacy as full-length ADAMTS13. These differing conclusions may stem from assay sensitivity and the source of recombinant ADAMTS13 variants. Further research is needed to clarify the biological function of the TSP1 repeats [8].

The cysteine-rich domain (CRD) found in various proteins serves diverse functions. ADAMTS13's CRD shares similarity with CRDs in other ADAMTS proteases, except for a distinct region between residues Ala451 and Ser501. Within this region lies exosite-2, composed of amino acid residues Gly471, Ala472, Ala473, and Val474, positioned just before the V-loop (residues 474-481). Exosite-2 forms a pocket facilitating hydrophobic interactions with specific residues in the VWF-A2 domain, which become exposed during shear stress-induced unfolding. Adjacent to Pro475, this hydrophobic pocket is significant. Pro475, a common SNP in the Japanese population, is linked to a 16% reduction in plasma ADAMTS13 activity and a two-fold decrease in substrate affinity. Studies propose that cis-proline, especially at position 475, favors interactions with aromatic side chains, particularly electron-rich ones. It is predicted that cis-proline at position 475 interacts with the aromatic side chain of Trp1644 in VWF, known to interact with the newly identified hydrophobic pocket within the CRD [58]. The CRD exhibits structural similarity with

the Dis domain and is stabilized by six disulfide bonds. Removing the CRD leads to a tenfold decrease in catalytic efficiency for VWF73 cleavage [16].

The central spacer domain is a 130 aa stretch [58], featuring a jelly roll fold of 10 β-sheets, binds to the unfolded VWF A2 domain's C-terminus. Removal of the spacer domain reduces VWF73 cleavage catalytic efficiency (kcat/KM) by 25-fold. This binding site consists mainly of a hydrophobic cluster surrounded by positively charged arginine residues R636, R660, and R568. These residues are crucial for CUB domains binding when ADAMTS13 adopts a closed conformation and are commonly associated with ADAMTS13 autoantibody epitopes in immune TTP. The mutations of these residues increases VWF73 cleavage rate by 4-fold due to disruption of intramolecular interaction with CUB domains, potentially enhancing antithrombotic activity in vivo. Anti-spacer domain antibodies that conformationally open ADAMTS13 also increase VWF73 cleavage rate, indicating the spacer domain as a primary exosite for ADAMTS13 recognition of VWF [16]. Further binding studies identified Arg568 and Phe592 as contributors to the epitope of anti-spacer domain-directed antibodies. Alanine mutation of surface-exposed amino acids Arg660, Tyr661, and Tyr665 impairs ADAMTS13 recognition by VWF. Conservative substitution of these exosite-3 residues with Arg568Lys, Phe592Tyr, Arg660Lys, Tyr661Phe, Tyr665Phe leads to a GoF variant with increased activity towards VWF. However, this variant may pose a risk of bleeding due to less efficient platelet recruitment at vascular injury sites resulting from increased proteolysis of low molecular weight VWF multimers. A comparison between variants truncated before MDTCS and after MDTC the spacer domain revealed that MDTCS retains proteolytic activity towards VWF, whereas MDTC exhibits significantly reduced VWF proteolytic activity. Exosite-3, located within the spacer domain, contains a cluster of hydrophobic residues with polar and charged side chains facing the solvent [58].

ADAMTS13 variants that lack either both the Cys-rich and spacer domains or just the spacer domain demonstrate minimal activity when it comes to peptidyl substrates. Furthermore, they exhibit almost no activity towards UL-VWF bound to cells or soluble VWF across different conditions. Within the spacer domain, removing residues Arg659-Glu664 (Δ6aa) or substituting Arg659, Arg660, and Tyr661 with alanine significantly reduces the cleavage of various VWF substrates [8].

ADAMTS13 contains eight TSR domains, with three linker regions positioned between TSR2, TSR3, TSR4, TSR5, and TSR8/CUB1 domains. TSR1 in ADAMTS13 features a WXXW motif, potentially involved in ligand binding and protein activation, represented as Trp387-X-X-Trp390. Mutating Trp387 significantly reduces ADAMTS13's binding affinity to denatured VWF and diminishes its proteolytic activity on the synthetic substrate FRETS-VWF73. O-fucosylation, a common post-translational modification of TSRs, is crucial for proper folding and cellular secretion of ADAMTS13. Absence of O-fucosylation leads to a complete reduction in secretion due to improper TSR folding. Substituting N-glycans at Asn707 of TSR2 and Asn828 of TSR4 with Gln does not impact protein secretion but decreases activity against FRETS-VWF73 under static conditions, suggesting increased flexibility in the linker regions. Lack of N-glycosylation may favor ADAMTS13's closed conformation, considered inactive. C-mannosylation of Trp387 enhances protein secretion, and cation-pi interactions between Trp387/Trp390 and Arg409/Arg407 contribute to protein stability. Mutations in these residues may compromise TSR1 stability, leading to conformational changes between MP/Dis and CRD/Spacer complexes [58].

Autoantibodies targeting CUB domains are present in both TTP patients and healthy individuals, sharing an epitope between Pro1292 and Tyr1296, which corresponds to the linker region between CUB1 and CUB2 domains. However, the connection between these antibodies and acquired TTP's

pathophysiology remains unclear. The roles of CUB domains in ADAMTS13 are not fully understood due to conflicting study results. C-terminal truncated ADAMTS13 versions, MDTCS, lacking TSR2–7 and CUB1–2 domains can still cleave VWF, while recombinant CUB-1 inhibits ULVWF cleavage. In vivo mouse experiments suggest that ADAMTS13 CUB domains modulate thrombus formation. The ADAMTS13-binding site spans VWF's D4-CK domains, interacting with ADAMTS13's C-terminal domains in a molecular zipper model. CUB domains can coordinate calcium ions, forming calcium-binding CUB (cbCUB) domains. However, human ADAMTS13 CUB domains show incomplete conservation of calcium-coordinating residues, suggesting unlikely calcium binding for the CUB1 domain [58].

The ADAMTS13 gene is linked to over 80 causative mutations for congenital TTP, including 16 within the DTCS region. These mutations are dispersed throughout ADAMTS13, suggesting structural abnormalities affecting folding and secretion. Mutants like R349C and P353L likely affect enzymatic activity, as they are crucial for exosite-1 and S3' pocket formation. P618A substitution reduces secretion efficiency in cultured cells. Acquired TTP, caused by autoantibodies targeting ADAMTS13, predominantly involves residues 657–666, part of exosite-3. Global latency of ADAMTS13 relies on noncovalent bonds between the Spacer domain and CUB1-2 domains. The expansive interdomain cleft in ADAMTS13 CUB1-2 gives rise to surface features like pockets P1 and P2. Loop L9 in CUB1 contains a unique unpaired Cys1275, potentially involved in redox regulation of VWF function. Docking simulations identified surface-exposed amino acids in CUB1-2 contributing to Spacer domain binding. Mutations in Trp1245, Trp1250, Lys1252 (CUB1), and Arg1326, Glu1387, Glu1389 (CUB2) disrupted Spacer-CUB interaction. Amino acid substitutions near these clusters also affected interaction. The specificity of these assays was validated by the R1219Q variant, which abolished 17G2 antibody binding and

subsequent inactivation, highlighting the assay's sensitivity to disruption of Spacer-CUB interaction [56].

The study aims to explain ADAMTS13's high substrate specificity despite circulating as an active enzyme. It proposes that ADAMTS13 exists in a latent form requiring allosteric activation by VWF exosite interactions. Kinetic analysis suggests that ADAMTS13's Cys-rich and Spacer domain exosites bind VWF, bringing the enzyme and unfolded VWF A2 domain close. Engagement of ADAMTS13's Dis exosite with VWF then activates the MP domain for proteolysis. Crystal structure analysis reveals an occluded active-site cleft in ADAMTS13's MP domain, suggesting a latent conformation needing allosteric change for substrate accommodation. The findings indicate VWF as an activating cofactor preceding ADAMTS13 proteolysis [7].

ADAMTS13's significance in the cardiovascular system is closely intertwined with VWF. Both ADAMTS13 and VWF are recognized as pivotal factors in various diseases and conditions involving vascular inflammation and thrombosis, such as cancer, atherosclerosis, sepsis, neurological disorders, and liver disease. While recombinant ADAMTS13 products are advancing in clinical trials for treating hereditary TTP, there's also exploration into additional therapeutic uses. Many of these conditions are linked to VWF-dependent inflammation or microvascular thrombosis, highlighting the broader implications of ADAMTS13 beyond TTP treatment [16].

2.7. Proposed Therapeutic Solutions

Level and activity of ADAMTS13 are independent concepts that need to be distinguished. ADAMTS13 activity is primarily determined by its concentration in the blood and its capacity to interact with VWF. Plasma level contains an active form of ADAMTS13 that is secreted into circulation. A two-sample Mendelian Randomization (MR) approach to verify role of ADAMTS13 in CVD study conclude that there is causal association between reduced CHD and

myocardial infarction(MI) with higher ADAMTS13 activity. The study also showed no causal relationship between ADAMTS13 level and five kind of CVDs(CHD, MI, AF,HF, and Venous Thrombo Embolism(VTE)) [6].

The specific mechanism linking low levels of ADAMTS13 to CVDs remains unknown, but various studies propose different mechanisms that may be associated with specific types of CVD. Low levels of ADAMTS13 antigen and activity have been linked to the pathophysiological mechanisms of acute MI (AMI), likely because of the increased prevalence of platelet thrombi that would result from ADAMTS13 deficiency [62]. In cases of inherited or acquired ADAMTS13 deficiency, a number of missense and silent mutations have been reported as common single-nucleotide polymorphisms (SNPs) without any evidence of thrombosis development. There may be a genetic predisposing effect to an elevated risk of thrombotic diseases, as shown by the frequencies of allele and functional characterization of some SNPs, which indicate an abnormal secretion of ADAMTS13 [70]. Schettert et al., who studied the association between ADAMTS13 genotype and CVD, found that the c.2699C>T (Ala900Val) variant was associated with an increased risk of death from an adverse cardiovascular event. The Pro475Ser variant has significantly low activity compared to wild-type ADAMTS13 (wtADAMTS13). The Pro618Ala variant has considerably reduced expression, secretion, and activity compared to wtADAMTS13. The in vivo impact of carrying these SNPs is unknown [62].

The study on conformational plasticity of ADAMTS13 reveals that it circulates in closed conformation and adopts a transient open conformation upon interaction with its substrate VWF [52]. ADAMTS13-mediated proteolysis depends on VWF conformational changes and its own conformational activation. Kinetic investigation of wtADAMTS13 showed 2.5-fold lower activity compared to ADAMTS13 without its C-terminal tail (MDTCS) or CUB1-2 domains, showing the

CUB domains limit ADAMTS13 function. Characterization of GoF ADAMTS13 provides further proof of conformational activation of ADAMTS13. There was a 2- to 2.5-fold increase in the overall rate of cleavage of the scissile bond fragments when this variation was used in combination with VWF fragments due to their higher association rate [60].

Pakistan is among top 6 countries which are highly affected by cardiovascular deaths [28]. In 2021, an epidemiological study showed that CVD mortality rate is highest in Pakistan among South Asian Nations [29]. There are many evidence-based CVD prevention and management guidelines. Significant gaps remain between evidence and practice. Effective interventions are needed to reduce the global burden of CVD [2]. The most dreaded CVD consequence and the leading cause of mortality globally is thrombosis, which poses a significant health care problem. Antithrombotic methods are efficient at stopping bleeding because they target platelets and the coagulation cascade. ADAMTS13 regulates prothrombotic effect[25] and therefore a novel target to prevent thrombosis and reduce CVD risk. Evaluating the role of ADAMTS13 in thrombosis and CVD by network and pathway studies and identifying the activators of ADAMTS13 can be a potential therapeutic approach to treat CVD.

2.8. Biological Regulatory Network

A biological regulatory network (BRN) outlines the interactions among biological entities, such as macromolecules or genes, within a system. This network is represented as a static interaction graph, where nodes symbolize the biological entities and edges represent their interactions. To illustrate the system's evolution, each entity's concentration level is denoted by a value linked to its respective node. The changes in these concentration levels over time define the system's dynamics. In the graph, inhibition is indicated by a "–" sign and activation is denoted by a "+"

sign [71, 72]. In this way, construction of BRN is significant approach to identify activators and inhibitors of a gene or protein.

2.9. Machine Learning Models

Arthur Samuel defines machine learning as the field of study that enables computers to learn from experience without being explicitly programmed. Machine learning (ML) aims to improve how machines process and handle data more effectively. In cases where data is too complex to interpret directly, ML techniques are applied to extract meaningful insights. With the growing availability of datasets, the demand for ML is increasing across various industries seeking to derive relevant information from their data. The core objective of ML is to enable systems to learn and adapt based on data. Extensive research has been conducted on how to create machines that can self-learn without explicit programming. Researchers and practitioners use various approaches to tackle problems involving large datasets. Machine learning employs different algorithms to address data-related challenges, and there is no universal algorithm that fits all scenarios. The choice of algorithm depends on the specific problem, the number of variables, and the most suitable model for the task [73].

Recent advancements in machine learning within healthcare have primarily enhanced the capabilities of physicians and analysts, aiding them in fulfilling their roles, identifying healthcare trends, and developing disease prediction models. In large medical institutions, machine learning techniques have been applied to improve the management of electronic health records, detect anomalies in blood samples, organs, and bones through medical imaging and monitoring, as well as in robot-assisted surgeries. Additionally, machine learning has played a crucial role in speeding up testing and hospital responses during the COVID-19 pandemic [74].

Recently, machine learning is extensively used for the development of new therapeutic interventions. The advancement of computational approaches and availability of large biological data is the motivation behind the use of artificial intelligence in pharmacology. It was about two decades since the inception of artificial intelligence in drug design and molecular informatics for the first time [75]. Machine/deep learning gained fame and continued to grow and now used in every aspect of drug discovery for revitalize the drugs development process. Machine learning is used in every aspect of ligand based and structured based drug design including target validation, prediction of target structure, prognostic biomarkers identification, virtual screening, optimization of hits and leads, prediction of biological activity, pharmacokinetic and toxicological (ADMET) profile of newly discovered ligands and to improve the efficiency and efficacy of the drugs [75-77].

Machine learning algorithms can be broadly categorized into Supervised Learning and Unsupervised Learning. **Supervised Learning** involves learning a function that maps input to output based on example input-output pairs. In this approach, the model is trained on labeled data, which consists of input-output pairs, to infer a function that can predict or classify unseen data. Supervised learning algorithms require external guidance, as the training dataset includes an output variable that needs to be predicted or classified. The model learns patterns from the training data and applies these patterns to the test data for making predictions or classifications. **Unsupervised Learning**, in contrast, does not rely on labeled data or predefined outcomes. Instead, algorithms in this category explore the data to uncover hidden patterns or structures without any explicit instructions. These algorithms identify interesting structures within the data, often for purposes such as clustering or feature reduction. When new data is introduced, unsupervised learning algorithms use previously learned features to recognize and categorize the data [73].

2.9.1. Classification Models

Classification Models in machine learning involve predicting a class label for a given example, using supervised learning methods. They map input variables to output variables (class labels) and can be categorized into three main types: **Binary Classification** involves classifying examples into one of two class labels, such as "true/false" or "spam/not spam." For example, distinguishing between "cancer detected" and "cancer not detected" in medical tests. **Multiclass Classification** involves classifying examples into one of several possible class labels, with more than two categories. Unlike binary classification, there is no normal/abnormal distinction. For instance, probing Attack in a dataset. **Multi-label Classification** extends multiclass classification by allowing examples to be associated with multiple labels simultaneously. Each example can belong to more than one class, such as categorizing a news article under multiple headings like "technology," "city name," and "latest news." This approach handles problems where labels are not mutually exclusive [78].

2.10. Prediction of mRNA Structure

mRNA serves as a crucial intermediary, transmitting genetic information from the DNA in the nucleus to the ribosomes in the cytoplasm, where proteins are synthesized. The functionality of mRNA hinges significantly on both its sequence and its three-dimensional (3D) structure, underscoring the importance of accurately understanding its structure. Typically, the 3D conformation of an mRNA is delineated by its secondary structure, which illustrates the base-pairing interactions within the molecule. Therefore, predicting and analyzing the secondary structure of mRNA is a fundamental step in deciphering its function. Various methods exist for probing RNA structure, including in vivo, in vitro, and in silico approaches. Among these, the in

silico method shows particular promise and cost-effectiveness, enabling the prediction of RNA secondary structures based on either sequence predictions or assessments of 3D structures [79].

CHAPTER 3 METHODOLOGY

The thesis workflow is structured into three distinct modules to achieve the objectives of the study. In the first module, Module I, a knowledge-driven BRN is constructed to elucidate the regulatory mechanism of the target. This involves a thorough literature study and understanding, followed by the use of network building tools.

The second module, Module II, focuses on the classification of proteins interacting with the protein target and their involvement in CVD. This is accomplished using Classification Modeling.

The third module, Module III, involves predicting the full-length 3D structure of the ADAMTS13 mRNA using RNAfold web server. This module is crucial for understanding protein structural features. The workflow is visually represented in ***Figure 3.1***.

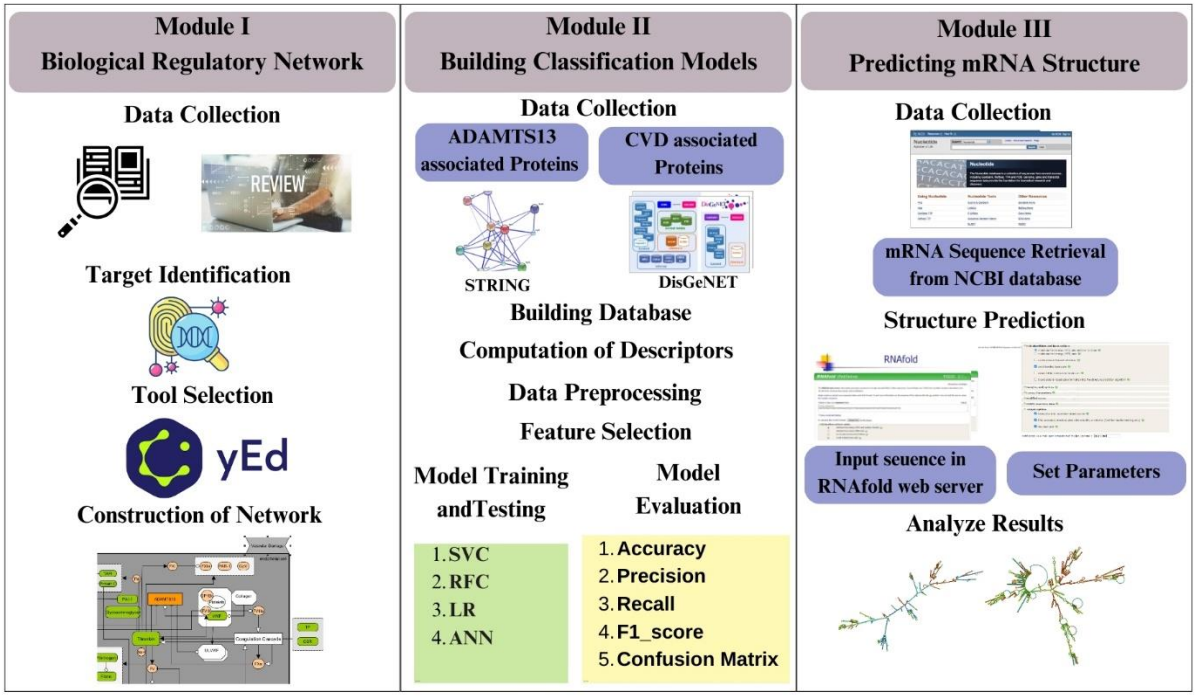


Figure 3.1: Workflow of overall methodology. This comprises Module I- Biological Regulatory Network, Module II- Building Classification Models and Module III-Prediction of mRNA structure**.**

MODULE I: CONSTRUCTION OF BIOLOGICAL REGULATORY NETWORK

3.1. Data Collection

Data collection involves retrieving data from various resources. To construct a BRN, we conducted a comprehensive review of literature focusing on ADAMTS13 structural and conformational studies, as well as ADAMTS13's role in thrombosis and CVD. The goal was to identify activators, cofactors, substrates, and inhibitors of ADAMTS13, and to understand its mechanism of action in endothelial cells. Those molecules and factors are used in constructing BRN that directly interfere with ADAMTS3 activity and indirectly have an effect on its activity and activate or inhibit those factors that affect ADAMTS13 activity. Subsequently, we constructed a regulatory network using yEd v3.23.2 [80].

MODULE II: BUILDING CLASSIFICATION MODELS

3.2. Data Retrieval

In our study, we collected data from public databases to build classification models. Firstly, we used the STRING [81] database to visualize interactions of ADAMTS13. We searched for Protein name and specified the organism as Homo Sapiens. To obtain maximum number of interactions we set 600 1st shell interactors and 600 2nd shell interactors.

From the DISGENET [82], we look for CVD category with keyword "CVDs" from which we expand our selection to similar diseases thus including "Disorder of Vein", "Venous Thrombosis" and "Vascular Diseases". The expanded diseases gene dataset summary was downloaded in csv file. Additionally, thrombosis diseases set is also explored with keyword "Thrombosis" and

explore similar diseases “Venous Thromboembolism”, “Thromboembolic stroke” expanding the selected diseases.

3.3. Database Construction

After data collection, data preprocessing is an indispensable step to remove duplicates, missing values, and noise from the data. Out of the 14,437 interactions collected from the STRING [81] database, only 613 unique protein entries were retained. The data retrieved from DisGeNET [82] is also cleaned to remove redundancy and noise.

We then built a database by shortlisting proteins involved in CVD and thrombosis and interacting directly with ADAMTS13, labeled as class_0 listed in Appendices: ***Table 0.1***, as well as those involved solely in CVD and thrombosis without interacting with ADAMTS13, labeled as class_1 listed in Appendices: ***Table 0.2***. This process resulted in 945 unique protein names, with 590 belonging to class_0 and 355 to class_1. Subsequently, we retrieved their UniProt IDs, and the sequences of the shortlisted proteins were obtained from the UniProt [83].

3.4. Descriptor Computation

After constructing the protein database, we computed protein descriptors for training our machine learning models. All protein descriptors were computed using the iFeature Web Server [84]. A FASTA file containing sequences of all proteins was submitted as an input. We selected 21 feature descriptors for each of the 945 proteins.

After computing descriptors, a CSV file was prepared including class labels for building classification models.

3.5. Building Classification Models

The classification models are built using Python libraries on Google Collaborator. We built binary classification models. We selected the following classification algorithms for study:

1. Support Vector Classifier (SVC)

2. Random Forest Classifier (RFC)

3. Logistic Regression (LR)

4. Artificial Neural Network (ANN)

Initially, the models were trained using default parameter settings(explained in section 3.9) and subsequently subjected to the test data. This was followed by fine-tuning of hyperparameters for each classifier, paving the way for reiteration of the training and testing process.

We used the following libraries to achieve our objective:

Pandas: Pandas is used for data manipulation and analysis, providing data structures like DataFrames.
Numpy: NumPy is used for numerical and array operations, essential for data handling.
Seaborn: Seaborn is a data visualization library built on top of Matplotlib
Matplotlib: Matplotlib is a popular plotting library for creating visualizations.
Scikit-learn: It is used to is used to handle missing data in datasets, feature selection, feature scaling/normalization, split data into training and testing sets for classification and regression and a cross-validation.
SciPy: SciPy is used for scientific and statistical computing

3.6. Data Preprocessing

Data preprocessing involves transforming raw data into a format that is more suitable for analysis. This process includes cleaning the data to remove errors or inconsistencies, handling missing values, and transforming the data into a format that is suitable for the specific machine learning algorithm being used. It is a crucial step in data mining and machine learning as it helps improve the quality of the data and ensures that the machine learning model can effectively learn from it.

Initially, the data was loaded from a CSV into a Pandas DataFrame to facilitate the process of data analysis and manipulation. The shape of the data was then determined, and the data types of the columns were checked to understand the attributes of the data. Next, missing values in the data were identified.

The first step in data preprocessing was feature scaling, also known as normalization, which is the standardization of independent variables (descriptors). This step is performed to ensure that all features have a similar scale so that no single feature dominates the others. It typically involves scaling the features to a range between 0 and 1 or to have a mean of 0 and a standard deviation of 1. This can be achieved using various methods, such as min-max scaling or standardization (Z-score normalization). We applied the StandardScaler submodule of the preprocessing module of the scikit-learn library, which uses Z-score normalization. The z-score normalization equation is given in $z = \frac{x-\mu}{\sigma}$ ***3.1***

$$z = \frac{x-\mu}{\sigma} \qquad \textbf{3.1}$$

Where z is the standardized value, x is the original value, μ is the mean of the feature (column), σ is the standard deviation of the feature.

The second step is Data Cleaning, which is a crucial step to identify and handle missing values, correct errors, remove inconsistencies, handle outliers, standardize the values in the correct format, and remove duplicates. This step is performed to improve the quality of data and prepare it for

analysis. To check noise in the data, a threshold of 0.5 was set and noise was detected. Our data was noisy, so noisy columns were dropped from the data frame. Furthermore, outliers were detected and removed using an isolation forest with a contamination parameter of 0.1, which indicates the expected proportion of outliers in the data. The random_state parameter was set to 42 for reproducibility.

For our dataset, we have assigned class labels. There are 2 classes:

Class_0 : Proteins interacting with ADAMTS13 and involved in CVD and thrombosis

Class_1: Proteins involved in CVD and thrombosis

3.7. Data Visualisation

Data visualization is the graphical representation of information and data in the form of charts, graphs, and maps to see and understand trends, outliers, and patterns in the data. First, we generated summary statistics of the data, such as mean, median, standard deviation, etc., to get an overview of the data. We built a distribution plot to visualize the skewness of our data before and after data preprocessing. The class distribution is visualized using a box plot.

3.8. Feature Selection

Feature selection is an important step to reduce dimensionality and improve model performance by selecting relevant features for accurate target prediction. We use sci-kit learn library module SelectKbest and tune its parameters to get optimal results. We tested f_classif and mutual_info_classif parameters with number of features ranging from 500 to 2000.

3.9. Model Training and Testing

After feature selection, we split our data into training and test sets, using a 0.2 test size, which implies that 80% of the data will be used for model training and 20% for testing.

For SVC, the parameters included the selection of an appropriate kernel based on the dimensionality of the data and the 'C' value, which is the regularization parameter crucial for preventing instance misclassification. The standard choice of kernel is the Radial Basis Function (RBF), which operates by non-linearly transforming data into a high-dimensional space, thus effectively used in situations where a non-linear association exists between class labels and attributes. The other most common choices of kernels are linear, polynomial, and least commonly, sigmoid.

For the random forest classifier, parameters include the number of trees in the forest (n_estimators), where higher numbers generally reduce overfitting. Another useful parameter is the criterion, which is a function used to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain. Further parameters include:

- max_depth: The maximum depth of the tree. Deeper trees can model more complex relationships but are more likely to overfit.
- min_samples_split: The minimum number of samples required to split an internal node. Higher values reduce overfitting.
- min_samples_leaf: The minimum number of samples required to be at a leaf node. Higher values reduce overfitting.
- max_features: The number of features to consider when looking for the best split. Can be a number or a percentage of features.
- random_state: Seed used by the random number generator.

Logistic Regression is one of the best statistical models used for binary classification tasks, where the output variable takes on only two possible values, usually represented as 0 and 1. It estimates

the probability that a given input belongs to a certain class. Some key parameters of Logistic Regression are:

- penalty: Specifies the norm used in the penalization. It can be 'l1' for L1 regularization, 'l2' for L2 regularization, or 'none' for no regularization.

- C: Inverse of regularization strength; smaller values specify stronger regularization.

- solver: Algorithm to use in the optimization problem. Common choices are 'liblinear' for small datasets, 'lbfgs' for multiclass problems, 'newton-cg', 'sag', and 'saga'.

- max_iter: Maximum number of iterations taken for the solvers to converge.

- random_state: Seed used by the random number generator.

In ANN, we used a multilayer perceptron classifier which underwent a refining process through the optimization of parameters such as the number of input neurons, the arrangement of hidden layers, the activation function, and the number of epochs. These enhancements were geared towards augmenting the performance of the neural network models depending on the dimensionality of the dataset. The output generated by the neurons in the neural network was decided because of the activation function. The most important and frequently used activation functions are Rectified Linear Unit (ReLU) and sigmoid. ReLU brings non-linearity to the results by mapping the output either as 1 or 0. It is the default activation function for neural networks. Sigmoid or logistic function generates a range of output values between 0 and 1.

The hyperparameter optimization was performed using GridSearchCV to boost model performance metrics. A comparative evaluation of model performance before and after hyperparameter tuning was undertaken to get the most optimal parameters for each classifier.

3.10. Model Evaluation

The performance of the classifiers was assessed post-training using various metrics. To identify the optimal classifier, the models were evaluated based on statistical parameters such as recall and accuracy. Subsequently, the models were deemed suitable for classifying unknown datasets. The study employed Accuracy, Confusion Matrix, Precision, Recall, F1 score, and Receiver Operating Characteristic (ROC) curve to determine the best classifier.

Classifier accuracy refers to its ability to correctly predict the class labels of instances belonging to different classes (positive and negative classes). True positives (TP) represent instances of the positive class correctly predicted by the classifier, while true negatives (TN) denote correctly predicted instances of the negative class. False positives (FP) and false negatives (FN) occur when the classifier incorrectly predicts the positive class or misclassifies instances of the positive class, respectively. These four measures underpin all the performance metrics used. Mathematically, accuracy is defined in Equation $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ ***3.2***

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{3.2}$$

The Confusion Matrix provides a visual representation for evaluating model accuracy and identifying areas for improvement. It evaluates the performance of a model by comparing actual results with predicted results. Like accuracy, it comprises four elements: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as shown in ***Figure 3.2***

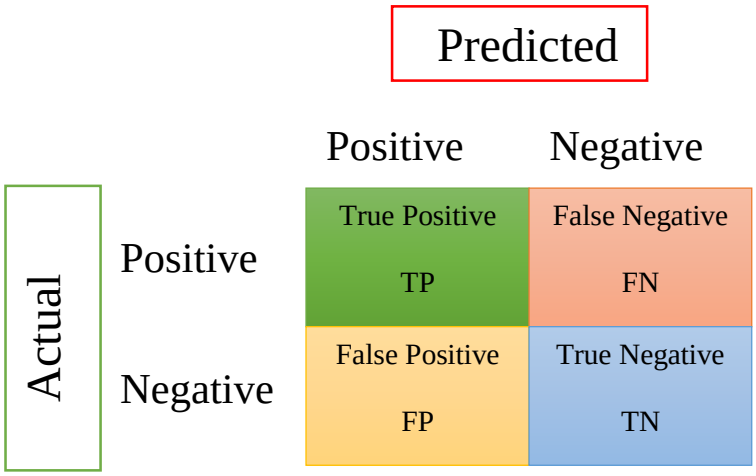


Figure 3.2 : Confusion Matrix in general

Precision and Recall both measure the true positive prediction rate of a model. Precision indicates the model's ability to accurately predict positive instances while minimizing the risk of false positives. It quantifies the proportion of true positive predictions among all positive predictions made by the model, making it crucial when the goal is to reduce false positives.

Recall, also known as Sensitivity, represents the classifier's ability to correctly classify instances of the positive class. It is the proportion of correctly classified positive instances. Improving precision often results in a decrease in recall, and vice versa. For example, a model that only predicts the most certain positive cases will have high precision but may miss many actual positive cases, leading to low recall. The mathematical representation of Precision and Recall is as given in Equation $Precision = \frac{TP}{TP + FP}$ ***3.3*** and $Recall = \frac{TP}{TP + FN}$ ***3.4*** respectively.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{3.3}$$

$$Recall = \frac{TP}{TP + FN} \quad \textbf{3.4}$$

The F1 score combines precision and recall by taking their harmonic mean. Maximizing the F1 score means optimizing both precision and recall simultaneously. Unlike the arithmetic mean, the harmonic mean is closer to the smaller number in a pair. In binary classification, a high F1 score of 1 indicates excellent precision and recall, whereas a low score suggests poor model performance. The mathematical equation for the F1 score is given in $F1\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall}$ ***3.5***

$$F1\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \qquad \textbf{3.5}$$

An receiver operating characteristic (ROC) curve is a graphical representation illustrating the performance of a classification model across all classification thresholds. This curve depicts two key parameters:

1. True Positive Rate (TPR), which is synonymous with recall.

2. False Positive Rate (FPR), defined in Equation $FPR = \frac{FP}{FP+TN}$ ***3.6***:

$$FPR = \frac{FP}{FP+TN} \qquad \textbf{3.6}$$

An ROC curve graphically represents the TPR against the FPR at various classification thresholds. Lowering the classification threshold categorizes more items as positive, thereby increasing both FP and TP.

A model that meets all these criteria would be deemed the best model because its predictions would be sufficiently reliable. Such a model could then be used as a generalized model for predicting outcomes on unknown datasets.

Module III: Prediction of mRNA Structure

To predict the secondary structure of mRNA, the first mRNA sequence was retrieved from the NCBI nucleotide database (ID: AB069698, length: 4951). The RNAfold web server, an integral part of the Vienna RNA package, was employed for this analysis. The Vienna RNA package provides a comprehensive collection of tools featuring advanced algorithms for RNA folding, comparison, and prediction of RNA-RNA interactions. Additionally, the package includes related programs that leverage these core tools, such as the barriers program, which explores RNA secondary structure energy landscapes, and LocARNA, which creates structural alignments. The most effective approach implemented is free energy minimization. The RNAfold web server is designed to predict the secondary structures of single-stranded RNA or DNA sequences. Currently, it supports up to 7,500 nucleotides (nt) for partition function calculations and up to 10,000 nt for minimum free energy (MFE) predictions. The MFE of individual sequences is determined using a dynamic programming algorithm originally developed by Zuker and Stiegler. In addition to MFE folding, equilibrium base-pairing probabilities are calculated using John McCaskill's partition function (PF) algorithm [85].

The input for the RNAfold web server can be a single RNA sequence in plain text or FASTA format, which can be pasted into the text box or uploaded as a file. By default, both the MFE and PF algorithms will be computed. A dot plot and MFE plain structure and centroid structure will be generated.

In summary, we constructed the knowledge-driven BRN to probe the activators and inhibitors of ADAMTS13 and understand regulation mechanism. Further, we employed classification models, SVC, RFC, logistic regression and ANN to distinguish ADAMTS13, CVD and thrombosis related proteins from CVD and thrombosis related proteins only. These models were selected due to their

significance to distinguish sensitive data and handle small values. Additionally, we predicted mRNA secondary structure of ADAMTS13 to get insights into its molecular dynamics.

CHAPTER 4 RESULTS AND DISCUSSION

In this chapter, we present the results of our research, detailing the findings derived from the data collected and analyzed. Our analysis aims to address the research questions outlined in the introduction and provide a comprehensive overview of the key insights. By examining these results, we can better understand the patterns, correlations, and trends that emerged from the study, offering a clear and detailed account of our findings.

MODULE I: BIOLOGICAL REGULATORY NETWORK

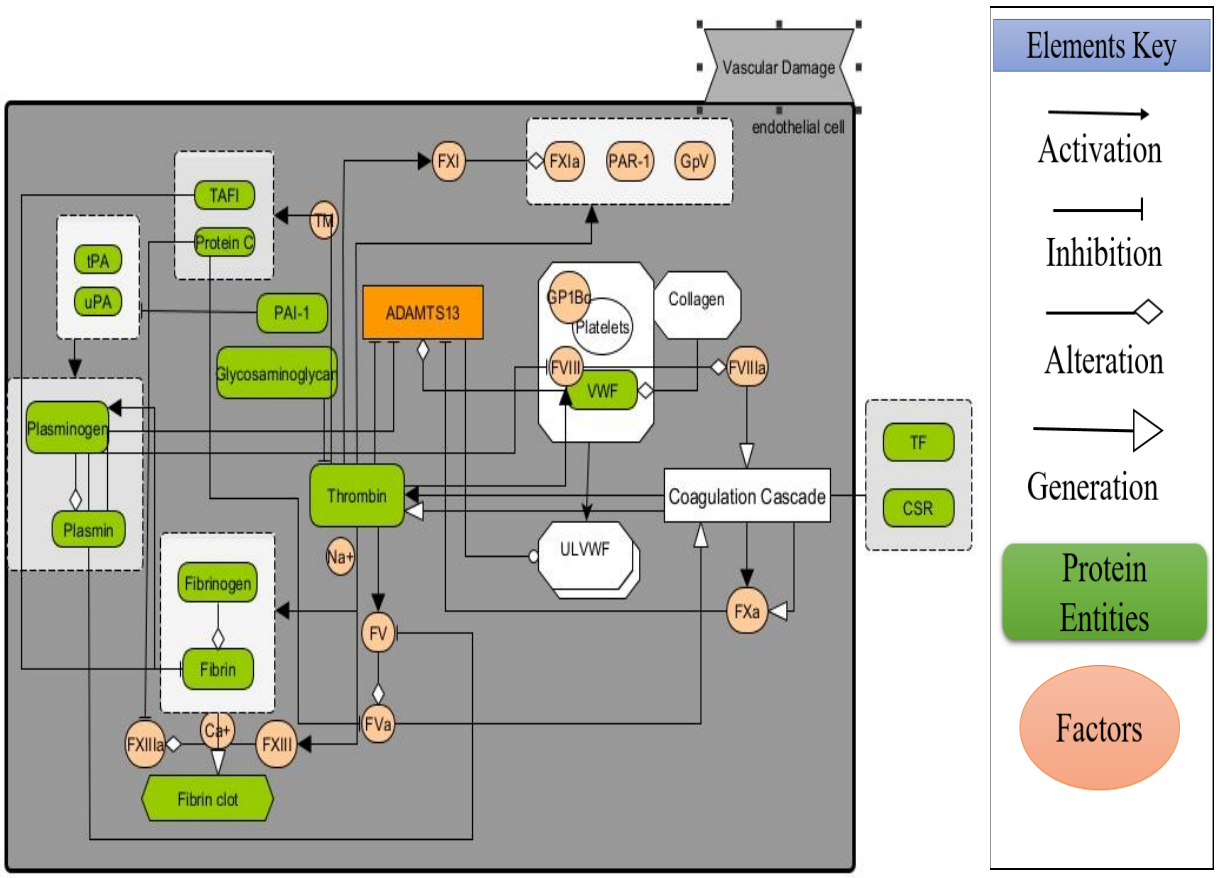


Figure 4.1: The Regulatory flow of ADAMTS13 with thrombin playing an active role in modulating different proteins to inhibit activity of ADAMTS13. The black arrows represent activation of protein. The rhombus-shaped white arrows represent modulation and white arrows represent enhancement of molecule or process and the lines with horizontal bars represent inhibition.

A BRN was constructed by data driven from a comprehensive literature search. A set of activating and inhibiting CVD proteins interacting with ADAMTS13 was collected. Those proteins that that directly or indirectly interact with ADAMTS13 were shortlisted for BRN construction. In BRN shown in **Error! Reference source not found.**, nodes represent the proteins and the edges represent different types of interaction between these proteins. The BRN was graphically represented using an interactive graphical software yEd v3.23.2 [80].

Following vessel injury, exposed sub-endothelial collagen binds to globular VWF via its A3 domain. Shear stress from blood flow triggers a structural change in VWF that reveals GPIbα-binding sites in its A1 domains, enabling platelet capture and plug formation. The stretched VWF

conformation also allows ADAMTS13, to cleave VWF at the Tyr1605-Met1606 site in the A2 domain [7, 14]. In absence or low activity of ADAMTS13, binding of platelet plug and VWF multimers results in accumulation of ULVWF multimers [30].

Vascular damage triggers TF and cell surface receptors(CSR) to initiate a series of coagulation cascade that results in thrombin generation. It binds with Na^+ to converts fibrinogen to fibrin, activates FXIII to FXIIIa and stabilize fibrin to fibrin clot in the presence of Ca^{2+}. Thrombin subsequently activates FV to FVa and VWF-bound FVIII to FVIIIa, thereby enhancing the coagulation cascade and increasing the generation of FXa and thrombin. On the surface of platelets, thrombin binds to GP1ba, which leads to the activation of FXI to FXIa and the activation of protease-activated receptors (PAR1) and GpV, resulting in platelet activation. It enhances platelet activation through various mechanisms, including cleaving ADAMTS13 to inhibit it. At undamaged sites, thrombin binds to thrombomodulin, leading to anticoagulant effects by activating TAFI and protein C, which inactivate FVa and FVIIIa. Thrombin's activity is also regulated by glycosaminoglycans and antithrombin [64].

Additionally, vascular damage triggers the release of proinflammatory cytokines and plasminogen activators like tPA and uPA, which convert plasminogen to plasmin[65]. Plasminogen activation is regulated by TAFI [66] and PA inhibitors like PAI-1, while plasmin also inhibits various factors including ADAMTS13 [67]. Plasmin's interaction with VWF is influenced by factors such as shear stress and blood group [68].

The constructed BRN highlights the role of various proteins upon vascular damage in endothelial cells. From this BRN, we analyze that thrombin has a key role in regulating different proteins, however, it triggers several proteins to enhance platelet activation and block activity of ADAMTS13 which could result in plug development. In this study we cannot suggest thrombin as

key therapeutic target because of its multifunctional role in homeostasis and coagulation cascade. This BRN also highlights inflammatory cytokines activation influencing ADAMTS13 activity. Therefore, it is suggested to study the association between coagulation, inflammation and homeostasis.

Module II: Classification Models

4.1. Data Collection and Preparation

An initial dataset of 14,437 interactions with the following features: node1, node2, node1_string_id, node2_string_id, neighborhood_on_chromosome, gene_fusion, phylogenetic_cooccurrence, homology, coexpression, experimentally_determined_interaction, database_annotated, automated_textmining, and combined_score were retrieved from STRING[81] database which after removing duplicates 613 interactions with our target ADAMTS13 were retained.

An additional dataset of 3539 proteins involved in CVD was downloaded from DisGeNET[82] which after removal of duplicates 2342 unique values were retained. After preprocessing and validation of proteins that, those proteins were shortlisted that are unique and are involved in CVD, thrombosis or interact with ADAMTS13. Based on this hypothesis, proteins are classified in two categories: 1) Involved in CVD and Thrombosis and interact with ADAMTS13. 2) Involved in CVD and Thrombosis. There were 945 unique entries out of which 590 belong to class_0 and 355 belong to class_1. We then computed 13494 descriptors of 945 proteins from iFeature Web Server[84]. The computed set of descriptors is listed in the table below:

Table 4.1: List of Descriptors, their groups and number of attributes associated with each descriptor

S.No	Name of Descriptor	No of attributes	Group of Descriptor

1.	Amino Acid Composition (AAC)	20	Group 1- Amino acid composition
2.	Composition of K-Spaced Amino Acid Pairs (CKSAAP)	2400	
3.	Dipeptide Composition (DPC)	400	
4.	Dipeptide Deviation from Expected Mean (DDE)	400	
5.	Tripeptide Composition (TPC)	8000	
6.	Grouped Amino Acid Composition (GAAC)	5	Group 2- Grouped amino acid composition
7.	Composition of K-Spaced Amino Acid Group Pairs (CKSAAGP)	150	
8.	Grouped Dipeptide Composition (GDPC)	25	
9.	Grouped Tripeptide Composition (GTPC)	125	
10.	Moran Autocorrelation	240	Group 3 - Autocorrelation
11.	Normalized Moreau-Broto Autocorrelation (NMBroto)	240	
12.	Geary Autocorrelation	240	

13.	C/T/D Composition (CTDC)	39	Group 4 - C/T/D
14.	C/T/D Distribution (CTDD)	195	
15.	C/T/D Transition (CTDT)	39	
16.	Conjoint Triad (CTriad)	343	Group 5 - Conjoint Triad
17.	K-Spaced Conjoint Triad (KSCTriad)	343	
18.	Sequence Order Coupling Number (SOCNumber)	60	Group 6 - Quasi-sequence-order
19.	Quasi-Sequence-Order (QSOrder)	100	
20.	Pseudo-Amino Acid Composition (PAAC)	50	Group 7 - Pseudo-amino acid composition
21.	Amphiphilic Pseudo-Amino Acid Composition (APAAC)	80	

4.2. Data Preprocessing

Initially, the dataset had a shape of (945, 13497), comprising 945 instances and 13,497 features. We first checked for the presence of null values and confirmed that there were no null entries in the dataset. Subsequently, we standardized the data using Z-score normalization.

Next, we examined the data distribution by plotting skewness distribution plots. To assess the presence of noise in the data, we applied a standard deviation threshold of 1. Noise was then removed using the Isolation Forest algorithm. The Isolation Forest method was selected due to its effectiveness in identifying and handling outliers in high-dimensional datasets. After removing outliers, the dataset had 850 instances and 13497 features.

After preprocessing, we plotted the data distribution once again to evaluate the effects of the applied transformations.

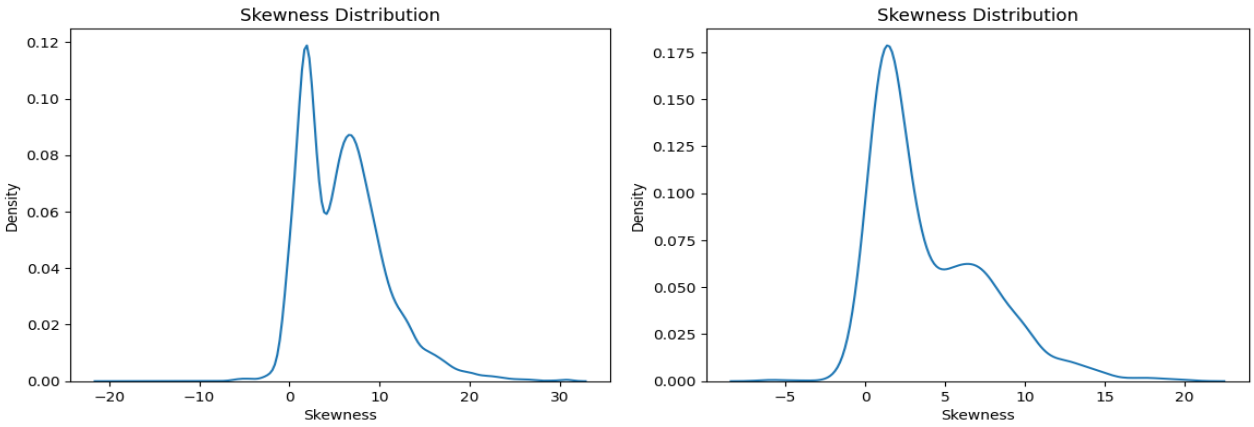


Figure 4.2: Distribution of Data before and after data preprocessing

The distribution plot in ***Figure 4.2*** shows bimodal data distribution concentrating at two peaks, 0 and 8 before data preprocessing justifying that data is influenced by 2 groups. However, after removing noise and irrelevant data, data distribution is slightly positively skewed. This might indicate presence of outliers, however outlier detection gives false results therefore, we suggest that data is concentrated towards right.

The bar chart in ***Figure 4.3*** shows class imbalance. Although 527 and 323 is not a significant class imbalance, however the dataset is small therefore the class distribution is high.

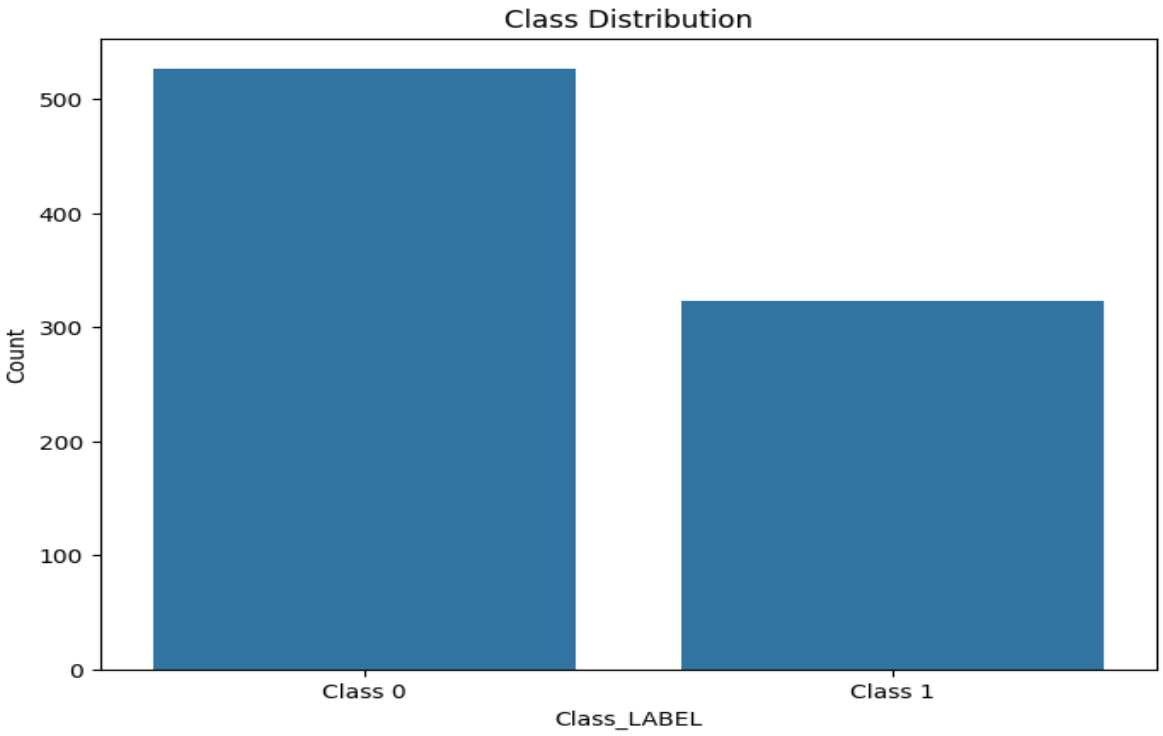


Figure 4.3: Class Distribution Bar Chart of Classification

4.3. Feature Selection and Model Outcomes

The Classification models are tested from 500-2000 number of features based on f_classif and mutual_info_classif feature selection methods. The models achieving highest accuracy of test data based on number of features selected are listed in **Error! Reference source not found.**.

Model	KBest		Random_state	Train_accuracy	Test_accuracy	AUC
ANN	f_classif	2000	21	1.0	0.8588	0.85
LR		2000	42	1.0	0.8647	0.85
SVC		1500		0.9602	0.8058	0.76
RFC		1000		1.0	0.6823	0.60
ANN	mutual_info_classif	1500	21	1.0	0.6470	0.60
LR		2000	42	1.0	0.6470	0.60
SVC		1000		0.9220	0.6529	0.53
RFC		1000		1.0	0.6352	0.56

The results of each model are discussed below:

4.3.1. Support Vector Classifier

SVC was built using default parameter (C:10, kernel: “RBF”). The results highlights accuracy of the model varies slightly at different number of features and keeping same selection parameters. However, the accuracy highly varies upon changing the parameter of feature selection. The accuracy of training dataset is ≥90% in all cases, but test accuracy of the model at various parameters is different. The model achieves highest test accuracy, that is 80.58% using KBest feature selection parameter f_classif specified to select 1500 best features. The highest AUC value achieved is 76% that implies 76% chances of model to accurately predict the correct values.

The huge different in train and test accuracy signifies that the model is not well optimized or either the model does not fit on this data. Further a 10-fold cross validation was performed which gives mean accuracy of 81.88% which is consistent with our model accuracy.

4.3.2. Random Forest Classifier

Random Forest Classifier built on default parameters with n_estimator=100, criterion='gini', max_depth='None', min_samples_leaf=1, min_samples_split=2 giving the following output at varied number of features.

The results in **Error! Reference source not found.** shows that the model learns the training data too well. The model achieves highest accuracy 68.23% and AUC 60% at f-classif and k=1000. The model is over fitted and not generalized. To avoid overfitting, we perform 10-fold cross validation which gives mean accuracy of 67.64% which is approximately similar.

4.3.3. Logistic Regression

The logistic regression model built at 'C'=1.0, 'penalty'=l2 and 'solver'=lbfgs The model is tested with various feature selection methods achieved perfect training accuracy, highlighting the model's ability to learn from the training data. The high training accuracy can be indicative of overfitting, especially if the test accuracy is significantly lower.

However, the highest test accuracy (86.47%) and AUC (85%) and 10-fold cross validation accuracy (86.11%) were observed with the f_classif method using k=2000, suggesting that this method provided the best balance between model complexity and generalization.

4.3.4. Artificial Neural Network

The performance of the artificial neural network (ANN) was evaluated using a series of metrics to assess its effectiveness in the given classification task. The ANN was trained and tested with the default specifications: 'activation'=relu, 'alpha'=0.0001, 'hidden_layer_sizes'=(100,)

During the training phase, the ANN achieved a convergence rate that was consistent with the expected behavior of the model. The training loss decreased steadily, indicating effective learning from the data. The final training accuracy reached 100%, demonstrating the model's capability to fit the training data effectively.

Upon evaluating the ANN on the test set, the model attained a test accuracy of 85.88%. This result reflects the model's ability to generalize to new, unseen data.

4.4. Model Evaluation

The models are evaluated by confusion matrix, computing precision, recall and f1_score to understand generalization of models. The confusion matrix of ***Figure 4.4***.

In SVC confusion matrix, out of 170 entries, 6 entries were predicted as FP. 27 entries were predicted as FN. The FN rate is high and lowering the accuracy of the model.

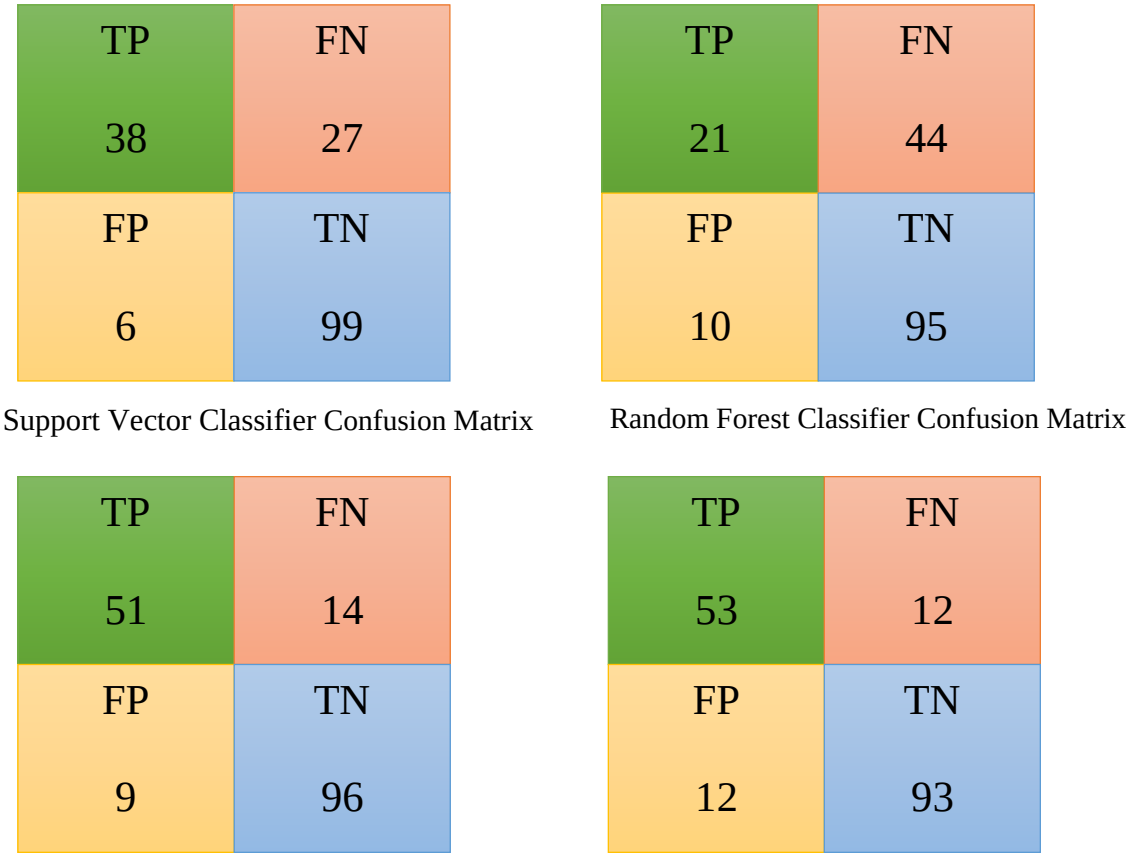


Figure 4.4: Confusion Matrix of Support Vector Classifier, Random Forest Classifier, Logistic Regression and Artificial Neural Network.

The precision score is 0.8636 and recall score is 0.5846. This implies model has low sensitivity. The F1 score is 0.6972. The overall model performance is good, however, lowering the FP rate is crucial for model.

RFC confusion matrix shows 10 entries were predicted as FP. 44 entries were predicted as FN. The FN rate is high and lowering the accuracy of the model. The precision score is 0.6774 and recall score is 0.3230. This implies model has low sensitivity. The F1 score is 0.4375. The overall model performance is good, however, lowering the FP rate is crucial for model.

The logistic regression model performance evaluation measures also report relatively better results. It has lower FN rate. The precision score is 85%, recall is 78.46% and F1 score is 81.60%. The model shows limitation in predicting positive instance.

The confusion matrix of ANN revealed the following performance metrics:

Precision: 81.53%, Recall: 81.53% and F1-score: 81.53%. Additional evaluation metrics were used to assess the performance of the ANN comprehensively:

AUC-ROC Score: 85%

Loss Function: The final loss value on the test set was 0.4305, reflecting the model's error in predicting the test data.

To ensure the robustness of the model, cross-validation was performed. The cross-validation results (87.76%) were consistent with the test results, demonstrating that the ANN is not overfitting and generalizes well to different subsets of the data.

The ANN demonstrated strong performance in terms of both training and test accuracy, with evaluation metrics indicating a well-balanced model. The achieved results suggest that the ANN is effective for the classification task at hand and provides a reliable tool for predicting proteins associated with ADAMTS13 and involved in CVD and thrombosis.

4.5. Hypertuned Models Performance

The results of GridSearch_CV highlights optimal parameters for the SVC model to achieve highest accuracy. The model gives its best performance at ‘C’=1 and at ‘sigmoid’ kernel while C is the regularization parameter. It achieves 83.52% test accuracy, 87.74% precision score and 66.15% sensitivity and 75.43% f1 score shown in ***Figure 4.5***. The model performance improves compared to performance at default parameters.

Best Hyperparameters : [‘C’: 1, ‘kernel’:sigmoid]

Test Accuracy: 0.8352

Confusion Matrix:

43	22
6	99

Precision score: 0.8774

Recall score: 0.6615

F1 score: 0.7543

Figure 4.5: Support Vector Classifier GridSearch_CV results

The RFC model is hypertuned to find optimal parameters on the training data using GridSearch_CV which gives the following results in ***Figure 4.6***.

Best Hyperparameters : [‘max_depth’: None, ‘min_samples_leaf’:1, ‘min_samples_split’:2, ‘n_estimator’: 500]

Test Accuracy: 0.6588

Confusion matrix

19	46
12	93

Precision score: 0.6129

Recall score: 0.2923

F1 score: 0.3958

Figure 4.6: Random Forest Classifier GridSearch_CV results

Logistic Regression GridSearch_CV results are shown in ***Figure 4.7***.

Best Hyperparameters : [‘C’: 0.01, ‘penalty’: ‘l2’, ‘solver’: ‘lbfgs’]

Test Accuracy: 0.8705

Confusion matrix

50	15
7	98

Precision score: 0.8771

Recall score: 0.7692

F1 score: 0.8196

Figure 4.7: Logistic Regression Classifier GridSearch_CV results

The hyperparameter optimization of the model indicates that the optimal parameters are ‘C’ = 0.01, ‘penalty’ = 'l2', and ‘solver’ = 'lbgfs' following parameter tuning. With these parameters, the test accuracy achieved is 87.05%, accompanied by relatively consistent evaluation scores.

The GridSearchCV was employed to identify the optimal hyperparameters for the Artificial Neural Network (ANN) model. The optimal hyperparameters identified by GridSearchCV are listed in ***Figure 4.8***.

Best Hyperparameters: (‘activation’: ‘logistic’, ‘alpha’:0.001, ‘hidden_layer_sizes’:(100,100))

Test Accuracy: 0.8882

Confusion matrix

57	8
11	94

Precision score: 0.8382

Recall score: 0.8769

F1 score: 0.8571

Log Loss: 0.3916

Figure 4.8: Artificial Neural Network GridSearch_CV results

The GridSearchCV process successfully identified the optimal hyperparameters for the ANN model, leading to an improved performance on the test set. The selected parameters resulted in high accuracy, balanced precision, recall, and a favorable F1-score, demonstrating the effectiveness of the model for the given classification task. The log loss metric also indicates good calibration of the predicted probabilities.

The results suggest that the chosen hyperparameters provide a robust configuration for the ANN, enhancing both its predictive accuracy and generalization capabilities. Future work may explore further fine-tuning or alternative hyperparameter configurations to potentially improve performance even further.

A comparison of the accuracy of all models, as illustrated in ***Figure 4.9***, reveals that the SVC, LR, and ANN models perform well on this dataset, whereas the RFC model overfits the data and does not exhibit strong performance. Among these, the ANN model achieves the highest overall performance metrics, although with a slight compromise in precision. The Logistic Regression model shows a trade-off in recall, and while the SVC model performs well, its performance is notably superior to that of the RFC model.

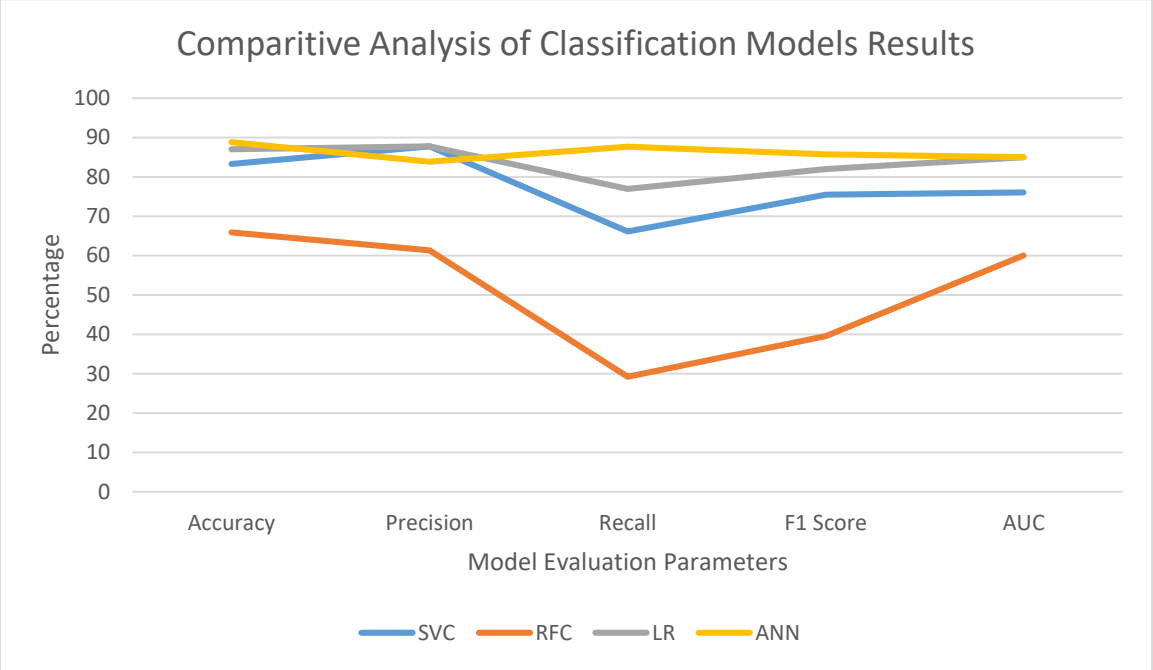


Figure 4.9: Comparison of Classification Model Performance

Module III: Prediction of mRNA Structure

4.6. Minimum Free Energy of Predicted Structure

The RNAfold server output contains the predicted MFE secondary structure in the usual dot-bracket notation. The optimal secondary structure in dot-bracket notation with a minimum free energy of -2325.90 kcal/mol.

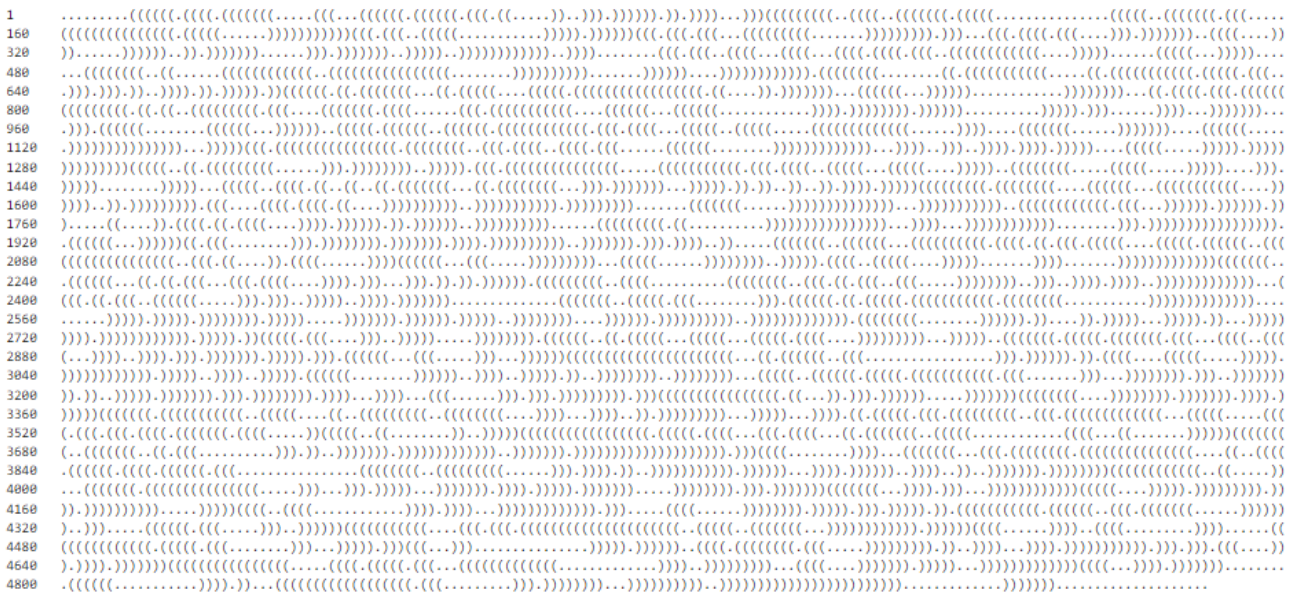


Figure 4.10: ADAMTS13 mRNA Secondary Structure in Dot Bracket notation

4.7. Partition Function Folding

The secondary structure, along with the sequence, can be submitted to the RNAeval web server, which provides a detailed thermodynamic analysis based on the loop-based energy model. When PF folding is selected, base-pairing probabilities are displayed as a Postscript dot plot. This dot plot is a two-dimensional graph where each potential base pair is represented by a box, with its size proportional to the probability of the corresponding base pair within an n × n grid. Consequently, the dot plot offers insights into all possible structures, rather than just a single optimal one [85].

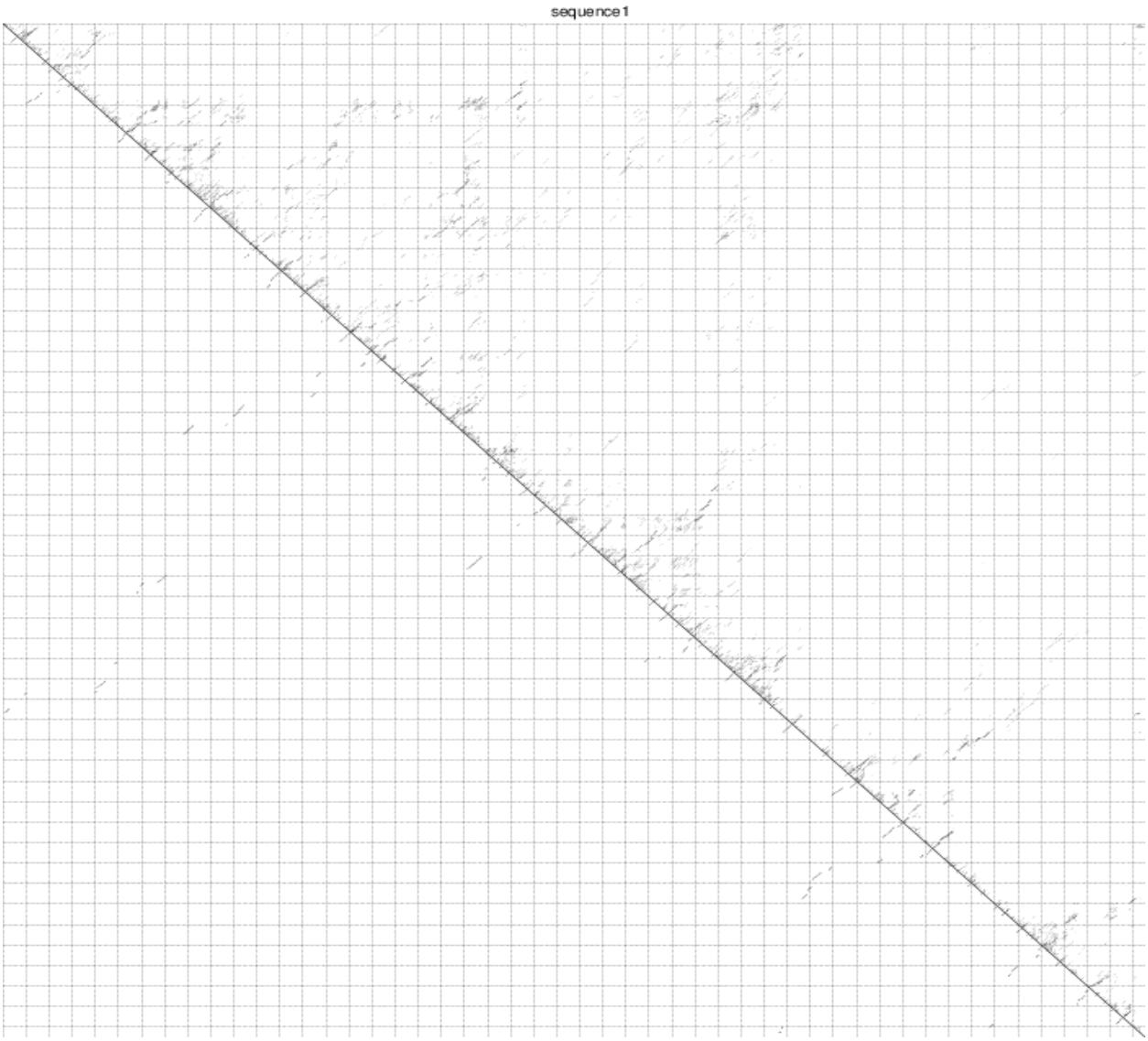


Figure 4.11: Partition Function Folding Dot Plot highlighting results from all structures

The dot plot covers the ensemble free energy, the frequency of the MFE structure, the ensemble diversity and base-pairing probabilities. Most predicted structures have same energy except few outliers. The free energy of the thermodynamic ensemble is -2394.73 kcal/mol. The free energy of the thermodynamic ensemble represents the total free energy of the ensemble of RNA secondary structures predicted for the given RNA sequence. This value is usually calculated as the sum of the free energies of all possible structures, weighted by their probability.

A more negative value typically indicates a more stable ensemble of structures. However, in this context, the free energy value seems unusually high in magnitude, suggesting that there might be a scale or unit misinterpretation.

The frequency of the MFE structure in the ensemble is 0.00 %. The MFE structure is the single RNA secondary structure with the lowest free energy, which is often used as a reference for predicting the most stable configuration of the RNA molecule.

The frequency of 0.00% indicates that, within the thermodynamic ensemble, the MFE structure is essentially absent or has an extremely low probability. This could suggest that the RNA molecule has a highly diverse set of possible structures, with the MFE structure not being significantly favored compared to other structures. This might be due to the RNA being in a dynamic state where multiple structures are almost equally probable.

Ensemble diversity quantifies the variability or spread of the structures within the thermodynamic ensemble. It reflects how diverse the secondary structures are, based on their free energies.

A high diversity value (1413.41 in this case) indicates a large number of structurally distinct conformations within the ensemble. This suggests that the RNA molecule can adopt a wide range of secondary structures with nearly equal stability. High ensemble diversity is typically associated with RNA sequences that are flexible and capable of adopting various conformations, which might be relevant for their biological function or interactions.

These findings indicate that the RNA molecule demonstrates significant structural flexibility and diversity. If the free energy value appears unusual, it is crucial to verify the results and their interpretation by consulting the RNAfold documentation or seeking support to ensure accurate analysis.

Given that RNA structure prediction is prone to errors, it is essential to supplement predicted structures with reliability information. One approach is to consider the centroid structure, which is the structure with the minimal base-pair distance to all other structures in the thermodynamic ensemble. This centroid structure can be determined by identifying all base pairs with a probability (P) greater than 0.5. Various types of reliability annotations can be derived from the results of partition function folding. A high degree of similarity between the centroid and MFE structures suggests a reliable prediction. Additionally, ensemble diversity, which measures the average base-pair distance between all structures in the Boltzmann ensemble, serves as another global indicator of prediction reliability[85].

The centroid secondary structure in dot-bracket notation with a minimum free energy of -1735.69 kcal/mol is given below.

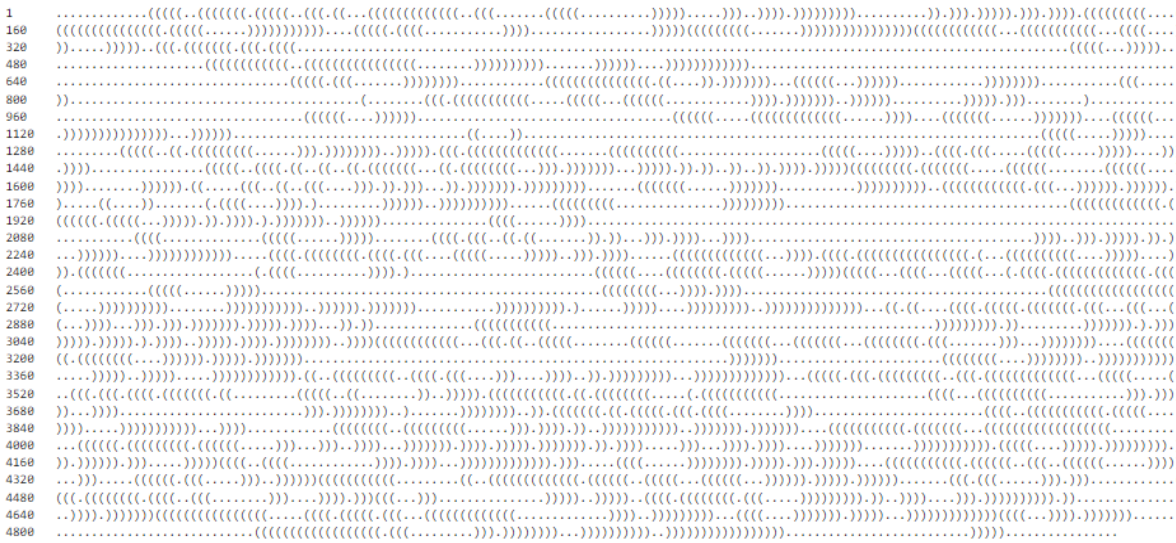


Figure 4.12: Centroid Secondary Structure in Dot Bracket Notation

Finally, we compared the graphical outputs of the Minimum Free Energy (MFE) structure and the centroid structure. The color coding in these visualizations indicates the base pairing probabilities. Our analysis reveals that the centroid structure differs significantly from the MFE structure, suggesting that the predictions may lack reliability.

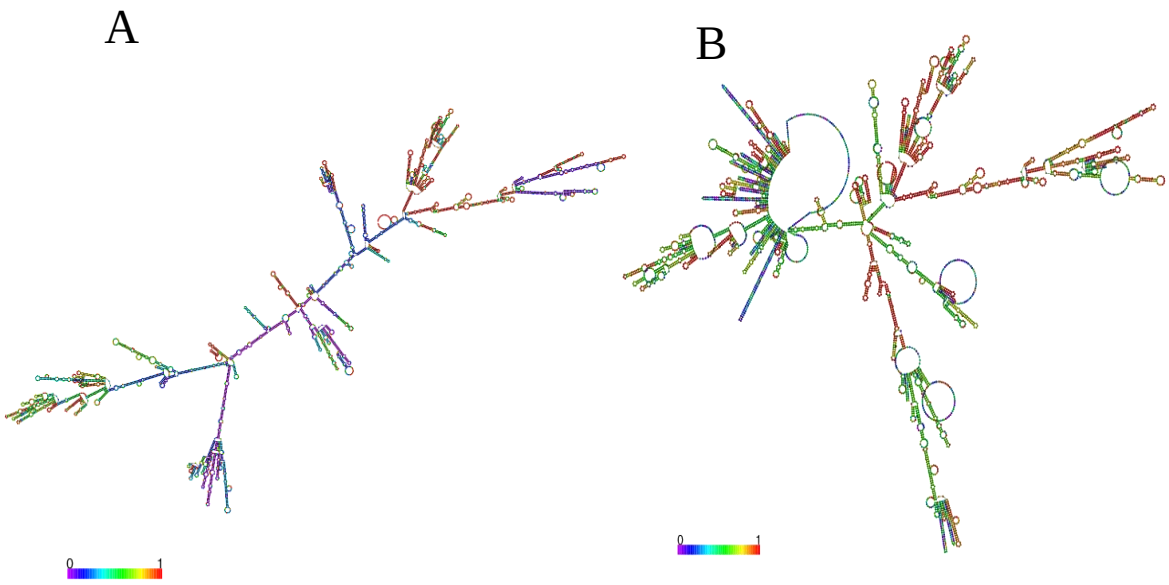


Figure 4.13: Graphical Structure of mRNA. (A) MFE structure drawing encoding base-pair probabilities. (B) Centroid structure drawing encoding base-pair probabilities

The substantial difference between the centroid structure and the MFE structure highlights a potential issue with the predictive accuracy of the RNA folding models used. While the MFE structure represents the most stable configuration predicted by the model, the centroid structure reflects a more averaged representation of the entire ensemble of predicted structures. A high degree of divergence between these structures may indicate that the MFE structure is not a reliable representative of the RNA's overall folding landscape.

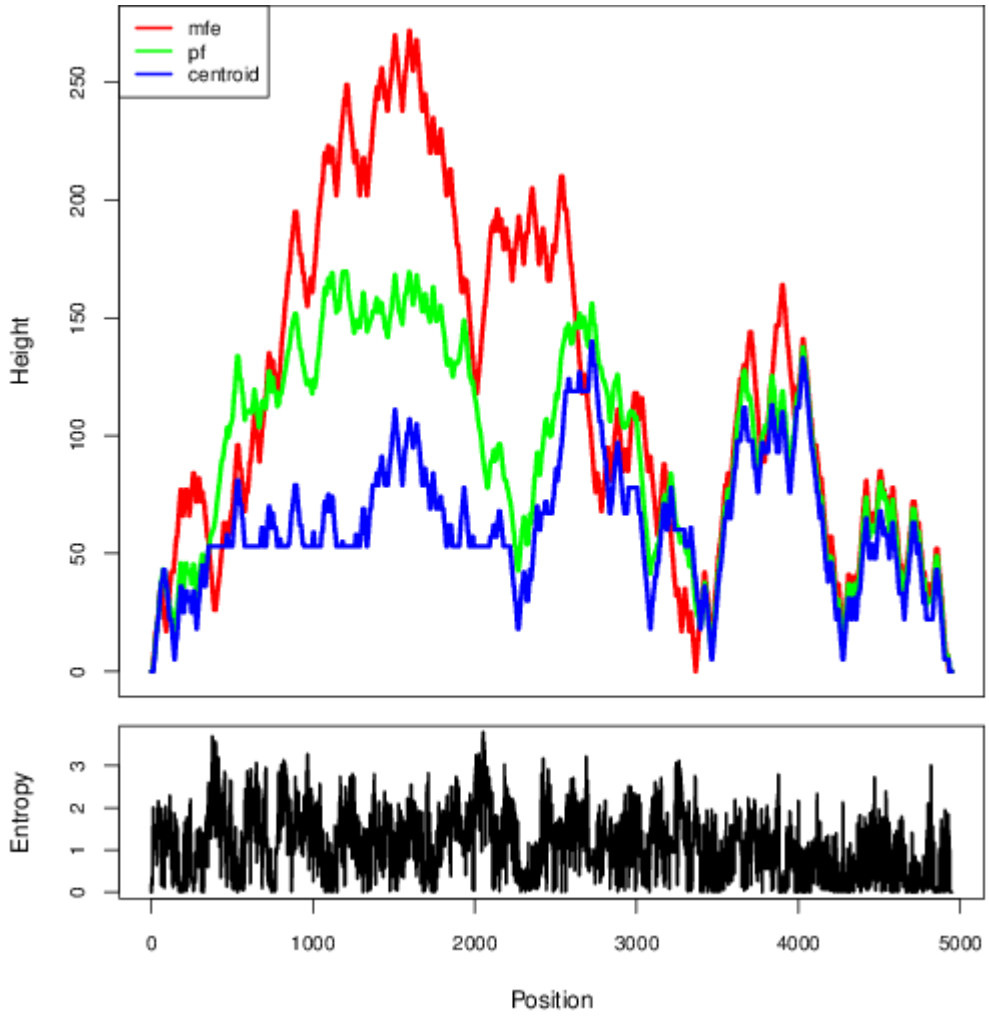


Figure 4.14: Mountain Plot depicting MFE structure, PF structure and centroid structure and entropy of each nucleotide position

We also present a mountain plot representation of the MFE structure, the thermodynamic ensemble of RNA structures, and the centroid structure. Additionally, we provide the positional entropy for each nucleotide position.

The mountain plot provides a visual representation of the folding landscape, showing how the predicted structures vary in terms of their free energies and stability. This plot can help in understanding the distribution of different structures within the ensemble and assessing their relative stabilities.

Positional entropy, which measures the variability or uncertainty at each nucleotide position, is also presented. High positional entropy suggests greater variability and less predictability in the base pairing at specific positions, further supporting the notion that the RNA may adopt multiple structurally diverse conformations.

Overall, these results emphasize the need for cautious interpretation of RNA secondary structure predictions, especially when significant discrepancies are observed between different structural models. Further validation and additional experimental data may be required to confirm the biological relevance of the predicted structures.

CONCLUSION AND FUTURE RECOMMENDATIONS

This study set to explore the regulation of ADAMTS13 in CVD via thrombosis by constructing a knowledge-driven BRN and distinguish ADAMTS13, CVD and thrombosis related proteins from those that are only involved in CVD and thrombosis by building and evaluating four classification models. Additionally, we also explore mRNA secondary structure of ADAMTS13 to study its dynamic nature.

We identified thrombin and plasmin as inhibitors of ADAMTS13, suggesting the need for further development of regulatory networks related to coagulation, inflammation, and homeostasis.

The classification models, specifically logistic regression and artificial neural networks (ANN), performed well on our dataset, achieving 85.88% and 86.47% test accuracy respectively. Further, these models evaluation parameters, precision, recall and F1_score, show model generalizability which can further be improved by increasing dataset and enhancing model parameters. Future work should aim to enhance model performance by incorporating additional datasets and fine-tuning model parameters.

The predicted mRNA structures were less reliable, likely due to the dynamic nature of mRNA. Predicting the full-length ADAMTS13 protein structure could provide more accurate insights into its functional mechanisms.

Overall, this study contributes to a better understanding of the regulatory roles of ADAMTS13 in thrombosis and CVD, paving the way for future research into targeted therapeutic strategies and the refinement of predictive models.

BIBILIORGRAPHY

1. Stewart, J., G. Manmathan, and P.J.J.c.d. Wilkinson, *Primary prevention of CVD: A review of contemporary guidance and literature.* 2017. **6**: p. 2048004016687211.
2. Nieuwlaat, R., et al., *Why are we failing to implement effective therapies in CVD?* 2013. **34**(17): p. 1262-1269.
3. Pluta, K., et al., *Platelet–leucocyte aggregates as novel biomarkers in CVDs.* 2022. **11**(2): p. 224.
4. Haybar, H., et al., *Endothelial cells: from dysfunction mechanism to pharmacological effect in CVD.* 2019. **19**: p. 13-22.
5. Stark, K. and S.J.N.R.C. Massberg, *Interplay between inflammation and thrombosis in cardiovascular pathology.* 2021. **18**(9): p. 666-682.
6. Ye, Z. and J. Zheng, *Verification of the Role of ADAMTS13 in the CVD Using Two-Sample Mendelian Randomization.* Frontiers in genetics, 2021. **12**: p. 1055.
7. Petri, A., et al., *Crystal structure and substrate-induced activation of ADAMTS13.* Nature Communications, 2019. **10**(1): p. 3781.
8. Zheng, X.L., *Structure-function and regulation of ADAMTS-13 protease.* 2013(1538-7836 (Electronic)).
9. Masias, C. and S.R. Cataland, *The role of ADAMTS13 testing in the diagnosis and management of thrombotic microangiopathies and thrombosis.* 2018(1528-0020 (Electronic)).
10. Ashorobi, D., M.A. Ameer, and R. Fernandez, *Thrombosis*, in *StatPearls*. 2023, © 2023, StatPearls Publishing LLC.: Treasure Island FL.
11. La Corte, A.L.C., H. Philippou, and R.A.S. Ariëns, *Chapter 3 - Role of Fibrin Structure in Thrombosis and Vascular Disease*, in *Advances in Protein Chemistry and Structural Biology*, R. Donev, Editor. 2011, Academic Press. p. 75-127.
12. Alkarithi, G., et al., *Thrombus Structural Composition in CVD.* 2021. **41**(9): p. 2370-2383.
13. Koupenova, M., et al., *Thrombosis and platelets: an update.* European Heart Journal, 2017. **38**(11): p. 785-791.
14. Akyol, O., S. Akyol, and C.-H. Chen, *Update on ADAMTS13 and VWF in cardiovascular and hematological disorders.* Clinica Chimica Acta, 2016. **463**: p. 109-118.
15. Cao, W., et al., *Factor VIII accelerates proteolytic cleavage of von Willebrand factor by ADAMTS13.* Proceedings of the National Academy of Sciences, 2008. **105**(21): p. 7416-7421.
16. DeYoung, V., et al., *Mechanisms of ADAMTS13 regulation.* 2022. **20**(12): p. 2722-2732.
17. Sonneveld, M.A., M.P. de Maat, and F.W.J.B.r. Leebeek, *Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis.* 2014. **28**(4): p. 167-178.
18. Gardner, M.D., *The essential role of calcium in ADAMTS13 function.* 2009.
19. Miyata, T., K. Kokame, and F. Banno, *Measurement of ADAMTS13 activity and inhibitors.* Current opinion in hematology, 2005. **12**(5): p. 384-389.
20. Kim, H.J., et al., *Crystal structure of ADAMTS13 CUB domains reveals their role in global latency.* Science Advances, 2021. **7**(16): p. eabg4403.

21. Crawley, J.T., et al., *Proteolytic inactivation of ADAMTS13 by thrombin and plasmin.* 2005. **105**(3): p. 1085-1093.
22. Miura, M., et al., *Prognostic value of plasma von Willebrand factor-cleaving protease (ADAMTS13) antigen levels in patients with coronary artery disease.* Thrombosis and haemostasis, 2010. **103**(03): p. 623-629.
23. Feng, Y., et al., *ADAMTS13: more than a regulator of thrombosis.* International Journal of Hematology, 2016. **104**(5): p. 534-539.
24. de Vries, P.S., et al., *Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity.* Blood, 2015. **125**(25): p. 3949-55.
25. Stark, K. and S. Massberg, *Interplay between inflammation and thrombosis in cardiovascular pathology.* Nature Reviews Cardiology, 2021. **18**(9): p. 666-682.
26. Ahmed, A., S. Ahmed, and G. Rådegran, *Plasma ADAMTS13 and von Willebrand factor in diagnosis and prediction of prognosis in pulmonary arterial hypertension.* Pulm Circ, 2021. **11**(4): p. 20458940211041500.
27. Reardon, B., L. Pasalic, and E.J. Favaloro, *The Intriguing Relationships of von Willebrand Factor, ADAMTS13 and Cardiac Disease.* J Cardiovasc Dev Dis, 2021. **8**(9).
28. Rehman, S., et al., *CVD (CVD): assessment, prediction and policy implications.* 2021. **21**(1): p. 1-14.
29. Joseph, P., et al., *CVD, mortality, and their associations with modifiable risk factors in a multi-national South Asia cohort: a PURE substudy.* 2022. **43**(30): p. 2831-2840.
30. Masias, C. and S.R. Cataland, *The role of ADAMTS13 testing in the diagnosis and management of thrombotic microangiopathies and thrombosis.* Blood, 2018. **132**(9): p. 903-910.
31. Chion, C.K., et al., *ADAMTS13 and von Willebrand factor and the risk of myocardial infarction in men.* 2007. **109**(5): p. 1998-2000.
32. Crawley, J.T., et al., *Evidence that high von Willebrand factor and low ADAMTS-13 levels independently increase the risk of a non-fatal heart attack.* 2008. **6**(4): p. 583-588.
33. Bongers, T., et al., *Lower levels of ADAMTS13 are associated with CVD in young patients.* 2009. **207**(1): p. 250-254.
34. Andersson, H.M., et al., *High VWF, low ADAMTS13, and oral contraceptives increase the risk of ischemic stroke and myocardial infarction in young women.* 2012. **119**(6): p. 1555-1560.
35. Zhao, B.-Q., et al., *von Willebrand factor–cleaving protease ADAMTS13 reduces ischemic brain injury in experimental stroke.* 2009. **114**(15): p. 3329-3334.
36. Fujioka, M., et al., *ADAMTS13 gene deletion aggravates ischemic brain damage: a possible neuroprotective role of ADAMTS13 by ameliorating postischemic hypoperfusion.* Blood, The Journal of the American Society of Hematology, 2010. **115**(8): p. 1650-1653.
37. Xu, H., et al., *ADAMTS13 controls vascular remodeling by modifying VWF reactivity during stroke recovery.* Blood, The Journal of the American Society of Hematology, 2017. **130**(1): p. 11-22.
38. Denorme, F., et al., *ADAMTS13-mediated thrombolysis of t-PA–resistant occlusions in ischemic stroke in mice.* Blood, The Journal of the American Society of Hematology, 2016. **127**(19): p. 2337-2345.
39. South, K., et al., *Enhanced activity of an ADAMTS-13 variant (R568K/F592Y/R660K/Y661F/Y665F) against platelet agglutination in vitro and in a*

murine model of acute ischemic stroke. Journal of Thrombosis and Haemostasis, 2018. **16**(11): p. 2289-2299.
40. De Meyer, S.F., et al., *Protective anti-inflammatory effect of ADAMTS13 on myocardial ischemia/reperfusion injury in mice.* Blood, The Journal of the American Society of Hematology, 2012. **120**(26): p. 5217-5223.
41. Chauhan, A.K., et al., *Systemic antithrombotic effects of ADAMTS13.* J Exp Med, 2006. **203**(3): p. 767-76.
42. Denorme, F., et al., *Reduced ADAMTS13 levels in patients with acute and chronic cerebrovascular disease.* PLoS One, 2017. **12**(6): p. e0179258.
43. Pedrazzini, G., et al., *Acquired intracoronary ADAMTS13 deficiency and VWF retention at sites of critical coronary stenosis in patients with STEMI.* Blood, The Journal of the American Society of Hematology, 2016. **127**(23): p. 2934-2936.
44. Sonneveld, M.A.H., et al., *Von Willebrand factor, ADAMTS13, and the risk of mortality: the Rotterdam study.* Arteriosclerosis, thrombosis, and vascular biology, 2016. **36**(12): p. 2446-2451.
45. Sonneveld, M.A.H., et al., *Low ADAMTS-13 activity and the risk of coronary heart disease–a prospective cohort study: the Rotterdam Study.* Journal of Thrombosis and Haemostasis, 2016. **14**(11): p. 2114-2120.
46. Rangarajan, S., C. Kessler, and L. Aledort, *The Clinical Implications of ADAMTS13 Function: The Perspectives of Haemostaseologists.* Thrombosis Research, 2013. **132**(4): p. 403-407.
47. Zheng, X.L., *ADAMTS13 and von Willebrand factor in thrombotic thrombocytopenic purpura.* Annu Rev Med, 2015. **66**: p. 211-25.
48. Taylor, A., et al., *Pharmacokinetics of plasma infusion in congenital thrombotic thrombocytopenic purpura.* 2019. **17**(1): p. 88-98.
49. Scully, M., et al., *Recombinant ADAMTS-13: first-in-human pharmacokinetics and safety in congenital thrombotic thrombocytopenic purpura.* 2017. **130**(19): p. 2055-2063.
50. Kretz, C.A., et al., *High throughput protease profiling comprehensively defines active site specificity for thrombin and ADAMTS13.* Scientific Reports, 2018. **8**(1): p. 2788.
51. Zhu, J., et al., *Exploring the "minimal" structure of a functional ADAMTS13 by mutagenesis and small-angle X-ray scattering.* Blood, 2019. **133**(17): p. 1909-1918.
52. Ercig, B., et al., *Conformational plasticity of ADAMTS13 in hemostasis and autoimmunity.* Journal of Biological Chemistry, 2021. **297**(4).
53. Zhang, A., et al., *Peters plus syndrome mutations affect the function and stability of human β1,3-glucosyltransferase.* Journal of Biological Chemistry, 2021. **297**(1): p. 100843.
54. Ricketts, L.M., et al., *O-fucosylation is required for ADAMTS13 secretion.* J Biol Chem, 2007. **282**(23): p. 17014-23.
55. Lancellotti, S. and R. De Cristofaro, *Chapter 3 - Structure and Proteolytic Properties of ADAMTS13, A Metalloprotease Involved in the Pathogenesis of Thrombotic Microangiopathies*, in *Progress in Molecular Biology and Translational Science*, E. Di Cera, Editor. 2011, Academic Press. p. 105-144.
56. Akiyama, M., et al., *Crystal structures of the noncatalytic domains of ADAMTS13 reveal multiple discontinuous exosites for von Willebrand factor.* Proceedings of the National Academy of Sciences, 2009. **106**(46): p. 19274-19279.
57. Adams, R., et al., *Binding sites in membrane proteins – Diversity, druggability and prospects.* European Journal of Cell Biology, 2012. **91**(4): p. 326-339.

58. Ercig, B., et al., *Insights into 3D Structure of ADAMTS13: A Stepping Stone towards Novel Therapeutic Treatment of Thrombotic Thrombocytopenic Purpura.* Thromb Haemost, 2018. **118**(01): p. 028-041.
59. Shin, Y., et al., *Proteolytic inactivation of ADAMTS13 by plasmin in human plasma: risk of thrombotic thrombocytopenic purpura.* The Journal of Biochemistry, 2017. **163**(5): p. 381-389.
60. South, K., et al., *Conformational activation of ADAMTS13.* 2014. **111**(52): p. 18578-18583.
61. Lancellotti, S., et al., *Mechanochemistry of von Willebrand factor*. 2019. **10**(1): p. 194-208.
62. Tseng, S.C. and C. Kimchi-Sarfaty, *SNPs in ADAMTS13.* Pharmacogenomics, 2011. **12**(8): p. 1147-1160.
63. Bernardo, A., et al., *Effects of inflammatory cytokines on the release and cleavage of the endothelial cell–derived ultralarge von Willebrand factor multimers under flow.* Blood, 2004. **104**(1): p. 100-106.
64. Al-Amer, O.M., *The role of thrombin in haemostasis.* 2022. **33**(3): p. 145-148.
65. Heissig, B., et al., *The multifaceted role of plasminogen in inflammation.* Cellular Signalling, 2020. **75**: p. 109761.
66. Miszta, A., et al., *Assessing Plasmin Generation in Health and Disease.* Int J Mol Sci, 2021. **22**(5).
67. Fay, W.P., N. Garg, and M. Sunkar, *Vascular Functions of the Plasminogen Activation System.* 2007. **27**(6): p. 1231-1237.
68. van der Vorm, L.N., et al., *Effects of Plasmin on von Willebrand Factor and Platelets: A Narrative Review.* TH Open, 2018. **02**(02): p. e218-e228.
69. Zander, C.B., W. Cao, and X.L. Zheng, *ADAMTS13 and von Willebrand factor interactions.* Curr Opin Hematol, 2015. **22**(5): p. 452-9.
70. Schettert, I.T., et al., *Association between ADAMTS13 polymorphisms and risk of cardiovascular events in chronic coronary disease.* Thrombosis Research, 2010. **125**(1): p. 61-66.
71. Richard, A., J.-P. Comet, and G. Bernot, *Formal methods for modeling biological regulatory networks*, in *Modern formal methods and applications*. 2006, Springer. p. 83-122.
72. Bernot, G., et al., *Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic.* Journal of Theoretical Biology, 2004. **229**(3): p. 339-347.
73. Mahesh, B.J.I.J.o.S. and R. . *Machine learning algorithms-a review.* 2020. **9**(1): p. 381-386.
74. Habehh, H. and S. Gohel, *Machine Learning in Healthcare.* Curr Genomics, 2021. **22**(4): p. 291-300.
75. Gawehn, E., J.A. Hiss, and G.J.M.i. Schneider, *Deep learning in drug discovery.* 2016. **35**(1): p. 3-14.
76. Vamathevan, J., et al., *Applications of machine learning in drug discovery and development.* 2019. **18**(6): p. 463-477.
77. Patel, L., et al., *Machine learning methods in drug discovery.* 2020. **25**(22): p. 5277.
78. Sarker, I.H., *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science, 2021. **2**(3): p. 160.
79. Therapeutics, m., *In Silico mRNA Structure Prediction.* 2023.

80. Editor, y.G., 2023.
81. Szklarczyk, D., et al., *The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic Acids Res, 2023. **51**(D1): p. D638-D646.
82. Piñero, J., et al., *The DisGeNET knowledge platform for disease genomics: 2019 update.* Nucleic Acids Research, 2020. **48**(D1): p. D845-D855.
83. The UniProt, C., *UniProt: the Universal Protein Knowledgebase in 2023.* Nucleic Acids Research, 2023. **51**(D1): p. D523-D531.
84. Chen, Z., et al., *iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences.* Bioinformatics, 2018. **34**(14): p. 2499-2502.
85. Gruber, A.R., et al., *The Vienna RNA websuite.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W70-4.

APPENDICES

Table 0.1: List of Protein and their UniProtIDs belonging to Class_0-ADAMTS13 associated and involved in CVD and Thrombosis Proteins of Classification Models

S.No	Protein Name	UniProtID	S.No	Protein Name	UniProtID
1	SIRT1 hCG_32306	A0A024QZQ1	2	FLNA FLN FLN1	P21333
3	UBE2D1	A0A087WW00	4	NF1	P21359
5	PPAN-P2RY11 hCG_2039996	A0A0B4J1V8	6	FPR1	P21462
7	SERPINI1	A0A0S2Z455	8	KITLG MGF SCF	P21583
9	PROS1	A0A0S2Z4L3	10	C5AR1 C5AR C5R1	P21730
11	SELENOP	A0A182DWH7	12	ERBB3 HER3	P21860
13	APP	A0A218KGR2	14	MATN1 CMP CRTM	P21941
15	HSPG2	A0A2R8YH07	16	FUT4 ELFT FCT3A	P22083

17	PIK3R1	A0A2X0SFG1	18	CDH3 CDHP	P22223
19	SPEN	A0A669KB49	20	IGFBP4 IBP4	P22692
21	RPSA2 RPSA RPSAP58	A0A8I5KQE6	22	GNLY LAG2 NKG5 TLA519	P22749
23	C3	A0A8Q3SI05	24	MMP8 CLG1	P22894
25	PTK2	A0A8Q3WLM4	26	FBLN1 PP213	P23142
27	NCSTN	A0A8V8TNV7	28	GLDC GCSP	P23378
29	PSEN1	A0A8V8TQ36	30	RPS3 OK/SW-cl.26	P23396
31	MED19 LCMR1	A0JLT2	32	JAK1 JAK1A JAK1B	P23458
33	NEURL1B NEURL3	A8MQ27	34	BDNF	P23560
35	CD55	B1AP13	36	MMP11 STMY3	P24347
37	TFPI	C9JKV3	38	IGFBP5 IBP5	P24593

39	F2	C9JV37	40	TNC HXB	P24821
41	AHSG	C9JV77	42	CCNC	P24863
43	RBPJ	D6RF98	44	APC DP2.5	P25054
45	SERPINA1	E9KL23	46	CD24 CD24A	P25063
47	F2	E9PIT3	48	F2R CF2R PAR1 TR	P25116
49	B2M	F5H6I0	50	RPS12	P25398
51	SPEN	F6WRY4	52	YY1 INO80S	P25490
53	KRT8	F8W1U3	54	ITGA3 MSK18	P26006
55	SERPINA6	G3V4V7	56	ITGB7	P26010
57	ACAN	H0YMF1	58	ITGB8	P26012
59	GNB2L1	I3QNU9	60	PTX3 TNFAIP5 TSG14	P26022

61	SNAPC5	I7GKX5	62	MSN	P26038
63	USP22	J3QRV6	64	DNMT1 AIM CXXC9 DNMT	P26358
65	FLT1	L7RSL3	66	PON1 PON	P27169
67	SNAP23	O00161	68	CBLIF GIF IFMH	P27352
69	STXBP3	O00186	70	DPP4 ADCP2 CD26	P27487
71	TLR4	O00206	72	COL8A1 C3orf7	P27658
73	F2RL2 PAR3	O00254	74	CAD	P27708
75	TAF4 TAF2C TAF2C1 TAF4A TAFII130 TAFII135	O00268	76	CALR CRTC	P27797
77	PIK3R2	O00459	78	CANX	P27824
79	AGRN AGRIN	O00468	80	CFP PFC	P27918

81	BIN1 AMPHL	O00499	82	LOX	P28300
83	PES1	O00541	84	MAPK1 ERK2 PRKM1 PRKM2	P28482
85	DLL1 UNQ146/PRO172	O00548	86	EPHA2 ECK	P29317
87	MFNG	O00587	88	SHC1 SHC SHCA	P29353
89	FCN1 FCNM	O00602	90	NOS3	P29474
91	CXCL11 ITAC SCYB11 SCYB9B	O14625	92	PEBP1 PBP PEBP	P30086
93	DVL1	O14640	94	SDC4	P31431
95	ADAM10 KUZ MADM	O14672	96	STX2 EPIM STX2A STX2B STX2C	P32856
97	PTGES MGST1L1 MPGES1 PGES PIG12	O14684	98	CDH5	P33151
99	ITGB1BP1 ICAP1	O14713	100	SHMT1	P34896
101	RIOK3 SUDD	O14730	102	SHMT2	P34897

103	NHERF1 NHERF SLC9A3R1	O14745	104	GPC1	P35052
105	NRP1 NRP VEGF165R	O14786	106	PHB1 PHB	P35232
107	APOL1 APOL	O14791	108	ADM AM	P35318
109	ADAMTS3 KIAA0366	O15072	110	PTGS2 COX2 PHG2	P35354
111	SMAD7 MADH7 MADH8	O15105	112	THBS2 TSP2	P35442
113	AXIN1 AXIN	O15169	114	FBN1 FBN	P35555
115	SURF4 SURF-4	O15260	116	IRS1	P35568
117	FABP7 BLBP FABPB MRG	O15540	118	NUP214 CAIN CAN KIAA0023	P35658
119	DHX15 DBP1 DDX15	O43143	120	KDR FLK1 VEGFR2	P35968
121	KLF4 EZF GKLF	O43474	122	TCF7 TCF1	P36402
123	MED7 ARC34 CRSP9	O43513	124	RPL4 RPL1	P36578

125	BCL2L11 BIM	O43521	126	TGFBR1 ALK5 SKR4	P36897
127	SPRY2	O43597	128	CFHR2 CFHL2 FHR2 HFL3	P36980
129	MED14 ARC150 CRSP2 CXorf4 DRIP150 EXLM1 RGR1 TRAP170	O60244	130	TGFBR2	P37173
131	KDM1A AOF2 KDM1 KIAA0601 LSD1	O60341	132	PPARG NR1C3	P37231
133	CUBN IFCR	O60494	134	BRCA1 RNF53	P38398
135	TLR2 TIL4	O60603	136	GGCX GC	P38435
137	JAK2	O60674	138	ITGAE	P38570
139	CTNND1 KIAA0384	O60716	140	LIPA	P38571
141	HPGDS GSTS PGDS PTGDS2	O60760	142	RPS19	P39019
143	TBL1X TBL1	O60907	144	MMP12 HME	P39900

145	KERA SLRR2B	O60938	146	GP5	P40197
147	ADAMTS4 KIAA0688 UNQ769/PRO1563	O75173	148	THPO MGDF	P40225
149	NCOR1 KIAA1047	O75376	150	MPL TPOR	P40238
151	SAP30	O75446	152	STAT3 APRF	P40763
153	MED24 ARC100 CRSP4 DRIP100 KIAA0130 THRAP4 TRAP100	O75448	154	LEP OB OBS	P41159
155	TADA2A TADA2L KL04P	O75478	156	PTGDS PDS	P41222
157	SUPT3H SPT3	O75486	158	STAT1	P42224
159	GPC4 UNQ474/PRO937	O75487	160	STAT5A STAT5	P42229
161	TADA3 ADA3 TADA3L	O75528	162	PIK3CA	P42336
163	TAF5L PAF65B	O75529	164	PIK3CB PIK3C1	P42338

165	MED6 ARC33	O75586	166	RPS27 MPS1	P42677
167	FOXH1 FAST1 FAST2	O75593	168	WAS IMD2	P42768
169	SURF6 SURF-6	O75683	170	CDKN2A CDKN2 MTS1	P42771
171	NCR1 LY94	O76036	172	MTHFR	P42898
173	NEURL1 NEURL NEURL1A RNF67	O76050	174	MMP13	P45452
175	MTA2 MTA1L1 PID	O94776	176	CRK	P46108
177	SUPT7L KIAA0764	O94864	178	CRKL	P46109
179	GOSR1 GS28	O95249	180	CDKN1B KIP1 p27	P46527
181	MED26 ARC70 CRSP7	O95402	182	NOTCH1 TAN1	P46531
183	ADAMTS2 PCINP PCPNI	O95450	184	RPS9	P46781
185	F8 F8C	P00451	186	RPS5	P46782

187	F13A1 F13A	P00488	188	RPS10	P46783
189	EGFR ERBB ERBB1 HER1	P00533	190	IQGAP1 KIAA0051	P46940
191	LALBA LYZL7	P00709	192	CXCL12 SDF1 SDF1A SDF1B	P48061
193	F2	P00734	194	IFNAR2 IFNABR IFNARB	P48551
195	C1R	P00736	196	CCN3 IGFBP9 NOV NOVH	P48745
197	HP	P00738	198	PXN	P49023
199	HPR	P00739	200	LMAN1 ERGIC53 F5F8D	P49257
201	F9	P00740	202	CDK8	P49336
203	F10	P00742	204	CXCR3 GPR9	P49682
205	CFD DF PFD	P00746	206	COMP	P49747
207	PLG	P00747	208	NUMB C14orf41	P49757

209	F12	P00748	210	PSEN2 AD4 PS2 PSNL2 STM2	P49810
211	PLAU	P00749	212	GSK3B	P49841
213	PLAT	P00750	214	TAF6 TAF2E TAFII70	P49848
215	CFB BF BFD	P00751	216	MMP14	P50281
217	ADA ADA1	P00813	218	FABP6 ILBP ILLBP	P51161
219	SERPINC1 AT3 PRO0309	P01008	220	SMARCA4 BAF190A BRG1 SNF2B SNF2L4	P51532
221	C5 CPAMD4	P01031	222	IRAK1 IRAK	P51617
223	TIMP1 CLGI TIMP	P01033	224	STAT5B	P51692
225	KNG1 BDK KNG	P01042	226	CDR1	P51861
227	FOS G0S7	P01100	228	LUM LDC SLRR2D	P51884
229	MYC BHLHE39	P01106	230	PRELP SLRR2A	P51888

231	LDLR	P01130	232	DGKE DAGK5	P52429
233	EGF	P01133	234	VAV2	P52735
235	TGFA	P01135	236	NUP98 ADAR2	P52948
237	TGFB1 TGFB	P01137	238	MYOM2	P54296
239	NGF NGFB	P01138	240	MFAP2 MAGP1	P55001
241	POMC	P01189	242	PLTP	P55058
243	GCG	P01275	244	MTTP MTP	P55157
245	INS	P01308	246	ADK	P55263
247	EPO	P01588	248	FOXA1 HNF3A TCF3A	P55317
249	MB	P02144	250	FCGRT FCRN	P55899
251	COL1A1	P02452	252	ITGA1	P56199

253	COL2A1	P02458	254	BACE1 BACE KIAA1149	P56817
255	COL3A1	P02461	256	LOXL3 LOXL	P58215
257	APOA1	P02647	258	ANTXR2 CMG2	P58335
259	APOE	P02649	260	ADAMTS12 UNQ1918/PRO4389	P58397
261	APOA2	P02652	262	ADAMTS20	P59510
263	APOC2 APC2	P02655	264	CD81 TAPA1 TSPAN28	P60033
265	APOC3	P02656	266	PTEN MMAC1 TEP1	P60484
267	FGA	P02671	268	RPS20	P60866
269	FGB	P02675	270	SNAP25 SNAP	P60880
271	FGG PRO2061	P02679	272	S100A10 ANX2LG CAL1L CLP11	P60903
273	MBP	P02686	274	CDC42	P60953

275	CRP PTX1	P02741	276	CXCR4	P61073
277	C1QA	P02745	278	MAX BHLHD4	P61244
279	C1QB	P02746	280	RPS3A FTE1 MFTL	P61247
281	C1QC C1QG	P02747	282	RHOA ARH12 ARHA RHO12	P61586
283	C9	P02748	284	SEC61A1 SEC61A	P61619
285	APOH B2G1	P02749	286	STXBP1 UNC18A	P61764
287	FN1 FN	P02751	288	B2M CDABP0092 HDCMA22P	P61769
289	AMBP HCP ITIL	P02760	290	RPS7	P62081
291	TTR PALB	P02766	292	RPS8 OK/SW-cl.83	P62241
293	ALB GIG20 GIG42 PRO0903 PRO1708 PRO2044 PRO2619 PRO2675 UNQ696/PRO1341	P02768	294	RPS16	P62249

295	AFP HPAFP	P02771	296	RPS14 PRO2640	P62263
297	GC	P02774	298	RPS23	P62266
299	PF4 CXCL4 SCYB4	P02776	300	RPS18 D6S218E	P62269
301	CXCL10 INP10 SCYB10	P02778	302	RPS29	P62273
303	LTF GIG12 LF	P02788	304	RPS13	P62277
305	HPX	P02790	306	RPS11	P62280
307	ESR1 ESR NR3A1	P03372	308	RPS4X CCG2 RPS4 SCAR	P62701
309	F11	P03951	310	RHOB ARH6 ARHB	P62745
311	KLKB1 KLK3	P03952	312	RPS6 OK/SW-cl.2	P62753
313	MMP1 CLG	P03956	314	RAP1A KREV1	P62834
315	VTN	P04004	316	RPS15 RIG	P62841

317	RAF1 RAF	P04049	318	RPS24	P62847
319	PROC	P04070	320	RPS25	P62851
321	APOB	P04114	322	RPS28	P62857
323	PRNP ALTPRP PRIP PRP	P04156	324	FAU	P62861
325	HRG	P04196	326	RPS27A UBA80 UBCEP1	P62979
327	VWF F8VWF	P04275	328	GRB2 ASH	P62993
329	SHBG	P04278	330	RAC1 TC25 MIG5	P63000
331	GAPDH GAPD CDABP0047 OK/SW-cl.12	P04406	332	VAMP2 SYB2	P63027
333	ERBB2 HER2 MLN19 NEU NGL	P04626	334	SKP1 EMC19 OCP2 SKP1A TCEB1L	P63208
335	TP53 P53	P04637	336	RPS21	P63220

337	IGF1 IBP1	P05019	338	HBB	P68871
339	ITGB3 GP3A	P05106	340	HBA1; HBA2	P69905
341	ITGB2 CD18 MFI7	P05107	342	NOP14 C4orf9 NOL14 RES4-25	P78316
343	SERPINB2 PAI2 PLANH2	P05120	344	SIRPA BIT MFR MYD1 PTPNS1 SHPS1 SIRP	P78324
345	SERPINE1 PAI1 PLANH1	P05121	346	OLR1 CLEC8A LOX1	P78380
347	SERPINA5 PCI PLANH3 PROCI	P05154	348	JAG1 JAGL1	P78504
349	SERPING1 C1IN C1NH	P05155	350	PRKDC HYRC HYRC1	P78527
351	F13B	P05160	352	ADAM17 CSVP TACE	P78536
353	EDN1	P05305	354	ADAMTSL3 KIAA1233	P82987
355	ICAM1	P05362	356	SMAD3 MADH3	P84022
357	JUN	P05412	358	RHOG ARHG	P84095

359	CLEC3B TNA	P05452	360	VLDLR	P98155
361	SERPIND1 HCF2	P05546	362	LRP2	P98164
363	ITGB1 FNRB MDF2 MSK12	P05556	364	CDR2 PCD17	Q01850
365	COL5A2	P05997	366	MEF2A MEF2	Q02078
367	INSR	P06213	368	RHAG RH50	Q02094
369	LCK	P06239	370	COL7A1	Q02388
371	FYN	P06241	372	MAP2K1 MEK1 PRKMK1	Q02750
373	RB1	P06400	374	CFHR3 CFHL3 FHR3	Q02985
375	PGR NR3C3	P06401	376	CAV1 CAV	Q03135
377	C2	P06681	378	PLAUR MO3 UPAR	Q03405

379	APOA4	P06727	380	CFHR1 CFHL CFHL1 CFHL1P CFHR1P FHR1 HFL1 HFL2	Q03591
381	ENO1 ENO1L1 MBPB1 MPB1	P06733	382	FOLH1 FOLH NAALAD1 PSM PSMA GIG27	Q04609
383	ITGAV MSK8 VNRA VTNR	P06756	384	NOTCH2	Q04721
385	LPL LIPD	P06858	386	PTPN11 PTP2C SHPTP2	Q06124
387	SERPINE2 PI7 PN1	P07093	388	FMOD FM SLRR2E	Q06828
389	FABP1 FABPL	P07148	390	CXCL9 CMK MIG SCYB9	Q07325
391	THBD	P07204	392	KHDRBS1 SAM68	Q07666
393	P4HB ERBA2L PDI PDIA1 PO4DB	P07237	394	BCL2L1 BCL2L BCLX	Q07817
395	ANXA2 ANX2 ANX2L4 CAL1H LPC2D	P07355	396	SOS1	Q07889
397	C8A	P07357	398	SOS2	Q07890

399	C8B	P07358	400	LRP1 A2MR APR	Q07954
401	GP1BA	P07359	402	LOXL1 LOXL	Q08397
403	C8G	P07360	404	CD47 MER6	Q08722
405	EPRS1 EPRS GLNS PARS QARS QPRS PIG32	P07814	406	RBL2 RB2	Q08999
407	UMOD	P07911	408	RBBP4 RBAP48	Q09028
409	THBS1 TSP TSP1	P07996	410	EP300 P300	Q09472
411	SP1 TSFP1	P08047	412	FOXO1 FKHR FOXO1A	Q12778
413	IGF1R	P08069	414	CNTN1	Q12860
415	COL1A2	P08123	416	TP53BP1	Q12888
417	NGFR TNFRSF16	P08138	418	TRAF2 TRAP3	Q12933

419	MMP2 CLG4A	P08253	420	TAF10 TAF2A TAF2H TAFII30	Q12962
421	MMP3 STMY1	P08254	422	RALGDS KIAA1308 RGF	Q12967
423	ITGA2B GP2B ITGAB	P08514	424	PLA2G7 PAFAH	Q13093
425	LPA	P08519	426	MYBPH	Q13203
427	CFH HF HF1 HF2	P08603	428	SKP2 FBXL1	Q13309
429	ITGA5 FNRA	P08648	430	ATM	Q13315
431	SERPINF2 AAP PLI	P08697	432	MFAP5 MAGP2	Q13361
433	RPS17 RPS17L	P08708	434	CTBP1 CTBP	Q13363
435	IGFBP1 IBP1	P08833	436	GAB1	Q13480
437	CD63 MLA1 TSPAN30	P08962	438	SMAD4 DPC4 MADH4	Q13485
439	MMP7 MPSL1 PUMP1	P09237	440	SNAPC2 SNAP45	Q13487

441	MMP10 STMY2	P09238	442	MED21 SRB7 SURB7	Q13503
443	HMGB1 HMG1	P09429	444	RIPK1 RIP RIP1	Q13546
445	SPARC ON	P09486	446	HDAC1 RPD3L1	Q13547
447	PDGFRB PDGFR PDGFR1	P09619	448	SNW1 SKIIP SKIP	Q13573
449	C1S	P09871	450	SEMA5A SEMAF	Q13591
451	CSF3 C17orf33 GCSF	P09919	452	KRR1 HRB2	Q13601
453	FURIN FUR PACE PCSK3	P09958	454	LAMB3 LAMNB1	Q13751
455	C4A CO4 CPAMD2	P0C0L4	456	LAMC2 LAMB2T LAMNB2	Q13753
457	C4B CO4 CPAMD3; C4B_2	P0C0L5	458	APOF	Q13790
459	GLI3	P10071	460	BYSL ENP1	Q13895
461	MYB	P10242	462	RUNX2 AML3 CBFA1 OSF2 PEBP2A	Q13950

463	TXN TRDX TRX TRX1	P10599	464	LRP8 APOER2	Q14114
465	C7	P10643	466	SELPLG	Q14242
467	KIT SCFR	P10721	468	HES1 BHLHB39 HL HRY	Q14469
469	CD28	P10747	470	ITIH4 IHRP ITIHL1 PK120 PRO1851	Q14624
471	CLU APOJ CLI KUB1 AAG4	P10909	472	BMS1 BMS1L KIAA0187	Q14692
473	HSPA5 GRP78	P11021	474	GNMT	Q14749
475	LIPC HTGL	P11150	476	LTBP1	Q14766
477	ITGAM CD11B CR3A	P11215	478	CHD4	Q14839
479	CETP	P11597	480	SLC10A1 NTCP GIG29	Q14973
481	COL11A1 COLL6	P12107	482	SNX17 KIAA0064	Q15036
483	F5	P12259	484	KARS1 KARS KIAA0070	Q15046

485	FCGR2A CD32 FCG2 FCGR2A1 IGFR2	P12318	486	RRS1 KIAA0112 RRR	Q15050
487	SKI	P12755	488	WDR43 KIAA0007 UTP5	Q15061
489	SKIL SNO	P12757	490	POSTN OSF2	Q15063
491	SRC SRC1	P12931	492	AGER RAGE	Q15109
493	XRCC6 G22P1	P12956	494	PDCD1 PD1	Q15116
495	GP1BB	P13224	496	PTGES3 P23 TEBP	Q15185
497	CCL5 D17S136E SCYA5	P13501	498	ERBB4 HER4	Q15303
499	NCAM1 NCAM	P13591	500	ANGPT1 KIAA0003	Q15389
501	VCAN CSPG2	P13611	502	FCN2 FCNL	Q15485
503	C6	P13671	504	MED22 SURF5	Q15528
505	F3	P13726	506	TAF5 TAF2D	Q15542

507	SELL LNHR LYAM1	P14151	508	TRADD	Q15628
509	GP9	P14770	510	MED1 ARC205 CRSP1 CRSP200 DRIP205 DRIP230 PBP PPARBP PPARGBP RB18A TRAP220 TRIP2	Q15648
511	MMP9 CLG4B	P14780	512	SMAD2 MADH2 MADR2	Q15796
513	POU2F1 OCT1 OTF1	P14859	514	EZH2 KMT6	Q15910
515	BRAF BRAF1 RAFB1	P15056	516	E2F4	Q16254
517	RAC2	P15153	518	TAF12 TAF15 TAF2J TAFII20	Q16514
519	CPN1 ACBP	P15169	520	SNAPC1 SNAP43	Q16533
521	MYOD1 BHLHC1 MYF3 MYOD	P15172	522	C3AR1 AZ3B C3R1 HNFAG09	Q16581
523	IFNGR1	P15260	524	NTRK2 TRKB	Q16620
525	CD19	P15391	526	MAPKAPK3	Q16644

527	CD46 MCP MIC10	P15529	528	HIF1A BHLHE78 MOP1 PASD8	Q16665
529	VEGFA VEGF	P15692	530	DUSP5 VH3	Q16690
531	CD1D	P15813	532	LAMA3 LAMNA	Q16787
533	RPS2 RPS4	P15880	534	DUSP6 MKP3 PYST1	Q16828
535	GATA1 ERYF1 GF1	P15976	536	AGT	Q53YY1
537	CD44 LHR MDU2 MDU3 MIC4	P16070	538	RRP12 KIAA0690	Q5JTH9
539	ACAN AGC1 CSPG1 MSK16	P16112	540	SNAPC4 SNAP190	Q5SXM2
541	PECAM1	P16284	542	PDZK1 CAP70 NHERF3 PDZD1	Q5T2W1
543	CD36 GP3B GP4	P16671	544	CAMSAP1	Q5T5Y3
545	IFNAR1 IFNAR	P17181	546	HES5 BHLHB38	Q5TA89
547	ITGA2 CD49B	P17301	548	RNF123 KPC1 FP1477	Q5XPI4

549	GJA1 GJAL	P17302	550	MED27 CRSP34 CRSP8	Q6P2C8
551	CEBPB TCF5 PP9092	P17676	552	RHBDF2 IRHOM2 RHBDL5 RHBDL6	Q6PJF5
553	XBP1 TREB5 XBP2	P17861	554	APOA5 RAP3 UNQ411/PRO773	Q6Q788
555	CR1 C3BR	P17927	556	RICTOR KIAA1999	Q6R327
557	IGFBP3 IBP3	P17936	558	TBPL2 TBP2 TRF3	Q6SJ96
559	PTPN1 PTP1B	P18031	560	ADAMTSL4 TSRC1 PP1396 UNQ2803/PRO34012	Q6UY14
561	ITGB5	P18084	562	ADAMTSL5 THSD6	Q6ZMM2
563	EGR1 KROX24 ZNF225	P18146	564	THSD4 UNQ9334/PRO34005	Q6ZMP0
565	VCL	P18206	566	WBSCR22	Q75ME3
567	ITGB6	P18564	568	ADAMTS13 C9orf8 UNQ6102/PRO20085	Q76LX8

569	SDC1 SDC	P18827	570	NRARP	Q7Z6K4
571	VCAM1	P19320	572	ADAMTSL2 KIAA0605	Q86TH1
573	TNFRSF1A TNFAR TNFR1	P19438	574	TADA2B ADA2B	Q86TJ2
575	ITIH2 IGHEP2	P19823	576	CD163 M130	Q86VB7
577	CR2 C3DR	P20023	578	DTX1	Q86Y01
579	TBP GTF2D1 TF2D TFIID	P20226	580	COL23A1	Q86Y22
581	ITGAX CD11C	P20702	582	TICAM1 PRVTIRB TRIF	Q8IUC6
583	OGN OIF SLRR3A	P20774	584	GPIHBP1 HBP1	Q8IV16
585	COL5A1	P20908	586	SBSPON C8orf84 RPESP	Q8IVN8
587	RASA1 GAP RASA	P20936	588	FTSJ3 SB92	Q8IY81
589	COL27A1 KIAA1870	Q8IZC6	590	CHPF CSS2	Q8IZ52

Table 0.2: List of Protein and their UniProtIDs belonging to Class_1-Proteins involved in CVD and Thrombosis of Classification Models

S. No	Protein Name	UniProtID	S. No	Protein Name	UniProtID
1	TR11B	O00300	2	MAML1 KIAA0200	Q92569
3	CXCR6	O00574	4	AKAP1	Q92585
5	PODXL	O00592	6	ESR2	Q92667
7	TNF11	O14788	8	TNR	Q92731
9	PMM2	O15305	10	KAT2A GCN5 GCN5L2	Q92752
11	GPR42	O15529	12	KAT2B PCAF	Q92830
13	MEFV	O15553	14	SNAPC3 SNAP50	Q92831
15	CD5L	O43866	16	KAT5 HTATIP TIP60	Q92966
17	PLIN1	O60240	18	MED12 ARC240 CAGH45 HOPAKIAA0192 TNRC11 TRAP230	Q92993
19	PLPL9	O60733	20	TSR2	Q93074
21	AOC2	O75106	22	FBXW7 FBW7 FBX30 SEL10	Q969E8
23	NR1I2	O75469	24	APH1A PSF CGI-78 UNQ579/PRO1141	Q969H0
25	DDAH1	O94760	26	TADA1 TADA1L	Q96BI3
27	ABCA1	O95477	28	NMD3 CGI-07	Q96BN2
29	TSN15	O95858	30	RMC1	Q96D46

31	COX1	P00395	32	KLRG1 CLEC15A MAFA MAFAL	Q96DM3
33	SODC	P00441	34	LTV1 C6orf93	Q96E93
35	ABL1	P00519	36	BCAN BEHAB CSPG7 UNQ2525/PRO6018	Q96GA3
37	RENI	P00797	38	MED30 THRAP6 TRAP25	Q96GW7
39	AACT	P01011	40	CPB2	Q96HR3
41	A2MG	P01023	42	ITCH	Q96IY4
43	CYTC	P01034	44	LOXL4 LOXC	Q96J02
45	NEU2	P01185	46	MAML3 KIAA1816	Q96JB6
47	TNFA	P01375	48	APLD1	Q96JK9
49	IFNG	P01579	50	CCDC42 CCDC42A	Q96LR9
51	IL1A	P01583	52	AK8 C9orf98	Q96M95
53	IL1B	P01584	54	NECTIN4 LNIR PRR4 PVRL4	Q96MA6
55	HLAB	P01889	56	NLRP3	Q96NY8
57	CO4A1	P02462	58	1A1L1	Q96P20
59	LMNA	P02545	60	F2RL3 PAR4	Q96QU6
61	PA21B	P04054	62	MED15 ARC105 CTG7A PCQAP TIG1 TNRC7	Q96RI0
63	ANXA1	P04083	64	HBEGF DTR DTS HEGFL	Q96RN5
65	CSF2	P04141	66	SORT1	Q99075
67	SODM	P04179	68	PITX2 ARP1 RGS RIEG RIEG1	Q99523
69	MAS	P04201	70	MTR	Q99697

71	CY24B	P04839	72	CIB1 CIB KIP PRKDCIP	Q99707
73	ARGI1	P05089	74	EBNA1BP2 EBP2	Q99828
75	S10A8	P05109	76	FOXC2	Q99848
77	IL4	P05112	78	CITED2 MRG1	Q99958
79	IL6	P05231	80	OMD SLRR2C UNQ190/PRO216	Q99967
81	S10A9	P06702	82	GDF15	Q99983
83	CEAM5	P06731	84	RIOK1 RIO1	Q99988
85	LIPG	P07098	86	NOL10	Q9BRS2
87	PGS2	P07585	88	UBAC1 GBDR1 KPC2 UBADC1	Q9BSC4
89	MDR1	P08183	90	MED10 L6 TRG17 TRG20	Q9BSL1
91	CD14	P08571	92	SPON2 DIL1 UNQ435/PRO866	Q9BTT4
93	MET	P08581	94	MED18	Q9BUD6
95	CP3A4	P08684	96	VAMP8	Q9BUE0
97	ANXA5	P08758	98	NOC4L	Q9BV40
99	IL6RA	P08887	100	RIOK2 RIO2	Q9BVI4
101	PAEP	P09466	102	AMN UNQ513/PRO1028	Q9BVS4
103	HMOX1	P09601	104	SRRT ARS2ASR2	Q9BXJ7
105	ROA1	P09651	106	CFHR5 CFHL5 FHR5	Q9BXP5
107	LOX5	P09917	108	MAK16 RBM13	Q9BXR6
109	SAA1	P0DJI8	110	ACE2	Q9BXY0
111	IL8	P10145	112	NIFK MKI67IP NOPP34	Q9BYF1
113	COX8A	P10176	114	GTPBP4 CRFG NOG1	Q9BYG3

115	ANDR	P10275	116	TRYD	Q9BZE4
117	OSTP	P10451	118	FTO	Q9BZJ3
119	KAP0	P10644	120	IL22	Q9C0B1
121	CH60	P10809	122	NAT10 ALP KIAA1709	Q9GZX6
123	GTR1	P11166	124	SLC25A18 GC2	Q9H0A0
125	BCR	P11274	126	MED28 EG1 FKSG20	Q9H1K4
127	PYC	P11498	128	CXL16	Q9H204
129	CP2C9	P11712	130	CHRD UNQ217/PRO243	Q9H2A7
131	ACE	P12821	132	ADAMTS10	Q9H2X0
133	MYH7	P12883	134	P63	Q9H324
135	CCL2	P13500	136	GPT	Q9H3D4
137	MIF	P14174	138	MMADHC C2orf25 CL25022 HSPC161 My011	Q9H3H5
139	HGF	P14210	140	TXNIP VDUP1	Q9H3L0
141	PA2GA	P14555	142	POFUT1 FUT12 KIAA0180	Q9H3M7
143	IL1R1	P14778	144	TBB1	Q9H488
145	ARSA	P15289	146	TFB2M NS5ATP5	Q9H4B7
147	LYAM3	P16109	148	ANTXR1 ATR TEM8	Q9H5Q4
149	BGAT	P16442	150	PTGES2 C9orf15 PGES2	Q9H6X2
151	JUNB	P17275	152	MED20 TRFP	Q9H7Z7
153	EGLN	P17813	154	RPF1 BXDC5	Q9H944
155	LEG3	P17931	156	POLR1B	Q9H9Y2
157	ADA2B	P18089	158	PLGRKT C9orf46 AD025 MDS030	Q9H9Y6
159	ATF3	P18847	160	TAF9B TAF9L	Q9HBL7

161	EPOR	P19235	162	LPAR2	Q9HBM6
163	NFKB1	P19838	164	SPON1 KIAA0762 VSGP	Q9HBW0
165	CXCL2	P19875	166	SMURF1 KIAA1625	Q9HCB6
167	IL11	P20809	168	GP6	Q9HCE7
169	CP2B6	P20813	170	TCF7L1 TCF3	Q9HCN6
171	CP3A5	P20815	172	RETN	Q9HCS4
173	PGS1	P21810	174	MMP25 MMP20 MMPL1 MT6MMP	Q9HD89
175	COMT	P21964	176	IL23A	Q9NPA2
177	IL10	P22301	178	NOX4	Q9NPF7
179	PGH1	P23219	180	MED4 ARC36 DRIP36 VDRIP HSPC126	Q9NPH5
181	TNFL4	P23510	182	PPAN BXDC3 SSF1	Q9NPJ6
183	THAS	P24557	184	HEYL BHLHB33 HRT3	Q9NQ55
185	EDNRA	P25101	186	ACSA	Q9NQ87
187	ACKR3	P25106	188	SPHK2	Q9NR19
189	TNR5	P25942	190	PDGFC SCDGF UNQ174/PRO200	Q9NRA0
191	APEX1	P27695	192	S17A5	Q9NRA1
193	NOS1	P29475	194	PNO1	Q9NRA2
195	CD40L	P29965	196	THSD1 TMTSP UNQ3010/PRO9769	Q9NRX1
197	BKRB2	P30411	198	FGF21	Q9NS62
199	SSR4	P31391	200	CTNNBIP1 ICAT	Q9NSA1
201	SC5A2	P31639	202	SIR3	Q9NSA3

203	AKT1	P31749	204	MED17 ARC77 CRSP6 DRIP77 DRIP80 TRAP80	Q9NTG7
205	BRS3	P32247	206	PARVA MXRA2	Q9NVC6
207	CP2CJ	P33261	208	DDX18 cPERP-D	Q9NVD7
209	MPI	P34949	210	RBM28	Q9NVP1
211	CTNNB1	P35222	212	MED9 MED25	Q9NW13
213	ADA1A	P35348	214	MED29 IXL	Q9NWA0
215	PE2R4	P35408	216	PGPI	Q9NX70
217	CBS	P35520	218	AATF CHE1 DED HSPC277	Q9NXJ5
219	CH3L1	P36222	220	DDX56 DDX21 NOH61	Q9NY61
221	PEDF	P36955	222	PSENEN PEN2 MDS033	Q9NY93
223	ACVL1	P37023	224	IL37 FIL1Z IL1F7 IL1H4 IL1RP1	Q9NZ42
225	IL6RB	P40189	226	EIF2AK3 PEK PERK	Q9NZH6
227	VHL	P40337	228	MED11 HSPC296	Q9NZJ5
229	CASR	P41180	230	PIPOX LPIPOX PSO	Q9P086
231	S19A1	P41440	232	SEMA5B KIAA1445 SEMAG UNQ5867/PRO34001	Q9P0Z9
233	ECE1	P42892	234	ADAMTS9 KIAA1312	Q9P283
235	PE2R3	P43115	236	PPARGC1A LEM6 PGC1 PGC1A PPARGC1	Q9P2N4
237	PI2R	P43119	238	MTRR	Q9UBK2
239	XDH	P47989	240	HEY2 BHLHB32 CHF1 GRL HERP HERP1 HRT2	Q9UBK8

241	LEPR	P48357	242	STK39 SPAK	Q9UBP5
243	GSH1	P48506	244	CIDEB	Q9UEW8
245	IDHP	P48735	246	ADAMTS1 KIAA1346 METH1	Q9UHD4
247	SERPH	P50454	248	MED13 ARC250 KIAA0593 THRAP1 TRAP240	Q9UHI8
249	VASP	P50552	250	TMEFF2 HPP1 TENB2 TPEF UNQ178/PRO204	Q9UHV7
251	AAPK2	P54646	252	MSRA	Q9UIK5
253	CP4F2	P78329	254	LEF1	Q9UJ68
255	CCL7	P80098	256	GDF2	Q9UJU2
257	NGAL	P80188	258	PO2F3	Q9UK05
259	TSG6	P98066	260	RCOR1 KIAA0071 RCOR	Q9UKI9
261	NAAA	Q02083	262	ADAMTS7	Q9UKL0
263	KPCD	Q05655	264	ADAMTS6	Q9UKP4
265	TNR9	Q07011	266	SARDH DMGDHL1	Q9UKP5
267	CD69	Q07108	268	MED23 ARC130 CRSP3 DRIP130 KIAA1216 SUR2	Q9UL12
269	PDE4D	Q08499	270	NOB1 ART4 NOB1P PSMD8BP1 MSTP158	Q9ULK4
271	AAPK1	Q13131	272	MMP17 MT4MMP	Q9ULX3
273	IL18	Q14116	274	PADI4	Q9ULZ9
275	GAS6	Q14393	276	ADAMTS5 ADAMTS11 ADMP2	Q9UM07
277	PDE3A	Q14432	278	PROCR EPCR	Q9UNA0

279	HABP2	Q14520	280	DIMT1 DIMT1L HUSSY-05	Q9UNN8
281	STAT4	Q14765	282	ADAMTS8 METH2	Q9UNQ2
283	MYPC3	Q14896	284	TRIM33 KIAA1113 RFG7 TIF1G	Q9UP79
285	PON2	Q15165	286	USP22 KIAA1063 USP3L	Q9UPN9
287	ADIPO	Q15848	288	THSD7A KIAA0960	Q9UPT9
289	IL17	Q16552	290	CNTN6	Q9UPZ6
291	SMUG1	Q53HV7	292	SH2B3	Q9UQ52
293	Q6KC15	Q6KC15	294	JAG2	Q9UQQ2
295	TET2	Q6N021	296	HPSE	Q9Y219
297	Q6NXQ3	Q6NXQ3	298	VSIG4 CRIg Z39IG UNQ317/PRO362	Q9Y251
299	CTL2	Q8IWA5	300	TLR6	Q9Y279
301	MAML2 KIAA1819	Q8IZL2	302	POFUT2 C21orf80 FUT13 KIAA0958	Q9Y2C9
303	PELP1 HMX3 MNAR	Q8IZL8	304	RCL1 RNAC RPC2 RPCL1 RTC2 HSPC338	Q9Y2G5
305	VKORL	Q8N0U8	306	PTN22	Q9Y2P8
307	GPC2	Q8N158	308	MED16 DRIP92 THRAP5	Q9Y2R2
309	NLGN1 KIAA1070	Q8N2Q7	310	UTP11 UTP11L CGI-94 HDCMB12P	Q9Y2X0
311	ADAMTSL1 ADAMTSR1 C9orf94 UNQ528/PRO1071	Q8N6G6	312	RRP15 KIAA0507 CGI-115	Q9Y3A2

313	POGLUT1 C3orf9 CLP46 KTELC1 MDSRP MDS010 UNQ490/PRO1006	Q8NBL1	314	MED31 SOH1 CGI-125	Q9Y3B9
315	PCSK9 NARC1 PSEC0052	Q8NBP7	316	AK6 CINAP AD-004 CGI-137	Q9Y3C7
317	PAG15	Q8NCC3	318	AAKB1	Q9Y3D8
319	SUPT20H C13orf19 FAM48A FP757	Q8NEM7	320	TRRAP PAF400	Q9Y478
321	AGBL3 CCP3	Q8NEM8	322	IRS2	Q9Y4A5
323	LFNG	Q8NES3	324	LOXL2	Q9Y4H2
325	FLCN	Q8NFG4	326	MMACHC	Q9Y4K0
327	DNER	Q8NFT8	328	HEY1 BHLHB31 CHF2 HERP2 HESR1 HRT1	Q9Y4U1
329	MCFD2 SDNSF	Q8NI22	330	MMP24 MT5MMP	Q9Y5J3
331	ADAMTS17	Q8TE56	332	KLF2	Q9Y5R2
333	ADAMTS16 KIAA2029	Q8TE57	334	NCOR2 CTG26	Q9Y5W3
335	ADAMTS15	Q8TE58	336	GPC6 UNQ369/PRO705	Q9Y618
337	ADAMTS19	Q8TE59	338	RFNG	Q9Y625
339	ADAMTS18 ADAMTS21	Q8TE60	340	SAR1B SARA2 SARB	Q9Y644
341	SCARB1 CD36L1 CLA1	Q8WTV0	342	TAF6L PAF65A	Q9Y6B6
343	FRS2	Q8WU20	344	SLCO1B1 LST1 OATP1B1 OATP2 OATPC SLC21A6	Q9Y6J9
345	APH1B PSFL UNQ688/PRO1328	Q8WW43	346	NCOA3 AIB1 BHLHE42 RAC3 TRAM1	Q9Y6L6
347	MUC16	Q8WXI7	348	NUMBL	Q9Y6Q9

349	ADAMTS14	Q8WXS8	350	SCIN	Q9Y6R0
351	ST8SIA2 SIAT8B STX	Q92186	352	DDX49	Q9Y6U3
353	CFHR4 CFHL4 FHR4	Q92496	354	J3KPQ8	J3KPQ8
355	PIK3R3	Q92569			