Using advanced computational methods to model the binding of antibody complexes: a case study from the coagulation cascade

Martin Rosellen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Research Department of Structural and Molecular Biology

University College London

June 8, 2021

I, Martin Rosellen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Haemophilia A is a congenital bleeding disorder affecting one in 5,000 to 10,000 males. To prevent symptomatic disease, injections of recombinant factor VIII (FVIII) are administered to compensate for insufficient levels of this essential clotting factor. Patients suffering from a severe form of haemophilia A are at increased risk of forming neutralising antibodies — known as inhibitors — against therapeutic FVIII. A better understanding of the binding characteristics of inhibitors may aid the selection of optimal haemophilia A therapies, lead to the development of new therapeutics that are less antigenic, and support future initiatives in personalised and precision medicine. With this goal in mind, Classical Molecular Dynamics (CMD) in conjunction with Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) free energy calculations, together with enhanced sampling techniques, have been used to investigate interactions and the dynamics of binding site residues of the human inhibitory antibody BO2C11 bound to the C2-domain of factor VIII. In parallel, recombinant bacterial expressions of the C2-domain were initiated with the aim to explore structural changes induced by mutations that abrogate binding as described previously in surface plasmon resonance experiments. Computational binding affinity predictions were generally shown to be in good agreement with experimental findings. Additionally, binding site dynamics were investigated in detail using customized visualization techniques and an interpretable machine learning approach. Nevertheless, CMD simulations were insufficient for gaining insights into structural changes induced by mutations that were determined experimentally to be non-binding, and for exploring the underlying differences between the bound and unbound structures of the FVIII-C2 domain. To this end, Acceler-

ated Molecular Dynamics (AMD) and Umbrella Sampling (US) simulations proved to be appropriate additions to investigate the conformational changes and energetic differences associated with the binding of BO2C11.

Impact Statement

Today's understanding of biomolecules is that function is determined by structure and dynamics rather than structure alone. This is especially relevant for larger biomolecules like proteins which modulate biological processes by conformational changes and/or by occupying different conformational states. It is therefore of interest to comprehend how molecular dynamics are able to promote or hinder certain functions. This knowledge could inspire the development of new or improved biomolecules for patient treatment or for what is more the regulation of biological processes in general.

Computational methods, such as Molecular dynamics simulations, have proven useful in this context and have been used here to investigate the binding of an antibody that inhibits the function of replacement Factor VIII - a therapeutic administered to patients possessing insufficient levels of this essential blood protein which is the case in the blood disorder haemophilia A. The development of inhibitory antibodies that neutralise replacement Factor VIII represents the most significant challenge to effective haemophilia A treatment and generates costs of hundreds of thousands of dollars per annum per patient. A better understanding of the binding characteristics of inhibitory antibodies may aid the selection of optimal haemophilia A therapies, lead to the development of new therapeutics that are less antigenic, and support future initiatives in personalised and precision medicine.

With the advent of gene therapy, the rise of autoimmune diseases and allergies and the apparent need for quicker development cycles for vaccines, which is currently highlighted by the coronavirus pandemic, the characterization of the interplay between immune system and biomolecules is getting ever more attention.

Existing experimental methods capable of providing a detailed characterisation of antibody binding, such as site-directed mutagenesis assays, are typically expensive, time-consuming and challenging. Computational methods were able to rank experimentally determined effects of mutations on binding affinity with reasonable success and provided detailed insights of FVIII C2-domain epitope dynamics which may guide the design of muteins that evade antibody inhibition. Conducted on commodity computer hardware, computational methods constitute a convenient and promising additional research avenue.

It is however difficult to analyse and interpret extensive molecular dynamics simulations exhibiting hundreds of thousands of atoms and time steps. With existing methods, the development of a rationale based on atomic motion is a labour-intensive and error-prone process. Because of this, a new approach has been assembled driven by the use of a readily available interpretable machine learning technique that can easily be analysed and interpreted. By that, a potentially impactful site for antibody affinity (residue T2253) was brought to attention which was not described as such in experiments as well as in binding affinity calculations. Insights as such may guide experimenters in making well informed and more target-oriented alterations to biomolecules.

The proposed interpretable machine learning approach is especially well suited for the extensive spatio-temporal nature of molecular dynamics simulations but is generally applicable to the interpretation of highly dimensional datasets. It may very well see a number of applications in all kinds of research areas.

Acknowledgements

While I am writing this, I am sitting in the living room of my mom's house and I am enjoying the view of a jolly bird that is hopping around in the lush green of the lawn. My childhood home has been a place of peace and support for me by my girlfriend Anna-Lena and my mother Irene who created a daily life that incorporated my needs in the last episode of my PhD. For that, I like to thank them with all my heart. This goes especially for Anna-Lena, who was with me in good and bad times throughout and who always made me aware of my intrinsic motivation that ultimately helped me to outgrow myself in this phase of my life.

The latter point also wouldn't have been possible without the trust in my abilities by my supervisory team. I like to thank Flemming Hansen for challenging my ideas, for testing my beliefs and for sharing his wisdom. He enabled me to look at my project from a variety of angles, in particular by the time that I was able to spent in the wet-lab. This experience is especially valuable to me and became a cherished memory not least because of group members like Harrison O'Brien, Vaibhav Shukla, Lucas Siemons and Angelo Fugueiredo who I could ask anything anytime and who I met for extracurricular activities.

If I would have written these acknowledgements in the order of gratitude that I experience, my secondary supervisor Adrian Shepherd with his team members David Moss and William Lees would be mentioned near the very top. The freedom to develop and try out my ideas was something that was received very well and promoted keeping an open mind which resulted in the innovations developed in the course of this work.

Last but not least, a big thank you to the LIDo DTP for making this PhD pos-

sible and for building a supportive community of fellow students and consultants.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Haemophilia A

Haemophilia A is an X-linked recessive bleeding disorder that has been known since the second century and was later recognized by John Otto in 1803 who described haemophilia in family pedigrees and laid the cornerstone for modern research. There are reported cases of acquired haemophilia A with an incidence of 1 in a million per year [1]. Much more common is hereditary haemophilia A, which affects one in 5,000 males, but is still comparatively rare and therefore classified as an orphan disease, with a total number of about 180,000 cases worldwide in 2018 [2]. Since it is inherited in an X-linked recessive manner, haemophilic women are extremely rare. Costs per haemophilia A patient per year easily exceed 100,000 pounds, which is almost exclusively due to the consumption of replacement factor VIII [3]. Expenditures further increase in patients that form inhibitory antibodies against therapeutic factor VIII [2, 4].

1.1.1 Haemophiliacs experience prolonged bleeding episodes

Insufficient levels or complete absence of functional blood protein factor VIII (FVIII) are the cause of haemophilia A, which results in prolonged bleeding episodes after injury or spontaneous joint and soft tissue bleeds. Special precautions have to be taken to prepare patients for surgery. For patients with very low levels of FVIII, joint bleeds account for 80% of incidents that can result in irreversible damages to bones and cartilages [2]. Previously, it was therefore commonly rec-

ommended that haemophiliacs refrain from physical activity resulting in reduced mobility which besides acute and chronic pain causes a higher level of experienced morbidity [5]. However, care has improved significantly over the last decade and physical activity is now recommended for haemophiliacs [6].

1.1.2 Factor VIII is crucial for blood coagulation

FVIII is a large (280 kDa) glycoprotein that performs crucial functions in the blood coagulation cascade. Dysfunctional FVIII stems from one or multiple mutations in the genes responsible for the B-domain truncated FVIII amino acid sequence. The chromosomal location of the FVIII gene is the q arm of the X-chromosome at position 28, and is inherited in an X-linked recessive pattern [7]. Expression of FVIII takes place in the liver and possibly other regions bearing endothelial cells, such as the lung but the literature has yet to come to a final conclusion in this matter [8, 9].

Upon release into the blood stream FVIII binds tightly in non-covalent fashion to von Willebrand factor (VWF) whereas FVIII not bound to VWF gets rapidly removed from the blood stream [10]. Initially, thrombin stemming from the point of vascular injury cleaves FVIII which results in dissociation from VWF and activated FVIII (FVIIIa) that is able to function as a cofactor to activated factor IX (FIXa) on phospholipid surfaces found on blood platelets (figure 1.1, 1.2). The role of FVIII as a cofactor with no enzymatic activity represents an exception in the coagulation cascade, that is only shared with its homolog factor V. By far the majority of proteins involved in blood coagulation are proteases that are synthesized as zymogens that require activation.

The complex of FVIIIa with FIXa accelerates the activation of factor X 10^5 fold which in turn leads to more thrombin that activates even more FVIII and FIX. This positive feedback loop greatly increase the availability of thrombin that is needed in vast quantities for cleaving fibrinogen into fibrin and the activation of factor XIII. Factor XIII is a transglutaminase that catalyses the formation of isopeptide bonds between fibrin lysines and glutamines, cross linking and stabilizing the fibrin to finally form the blood clot. A derogation of the positive feedback loop

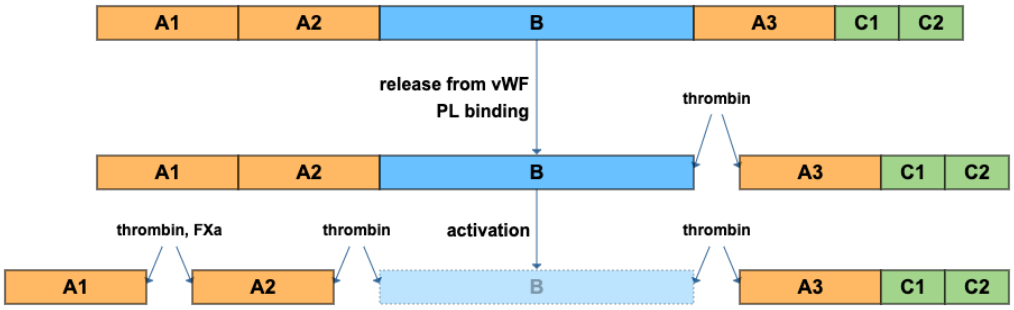


Figure 1.1: The proteolytic activation of FVIII: Thrombin is cleaving the light chain (A3, C1, C2) which reduces the affinity to VWF that in turn is competed off by the higher affinity to phospholipid membranes (PL). To function as a cofactor to FIXa the heavy chain (A1, A2) of FVIII has to be cleaved which results in a release of the B-domain. FVIIIa is the non-covalently bound complex of the heavy chain and light chain of FVIII.

by low levels of functional FVIII therefore leads to prolonged bleeding episodes [7, 11, 12].

Inactivation of FVIIIa is facilitated by protein C that cleaves the A1 and A2 domains which are then transported to intracellular degradation pathways mainly by members of the low-density lipoprotein (LDL) receptor family [13, 14].

An overview of the function of FVIII in the coagulation cascade is given in figure 1.3.

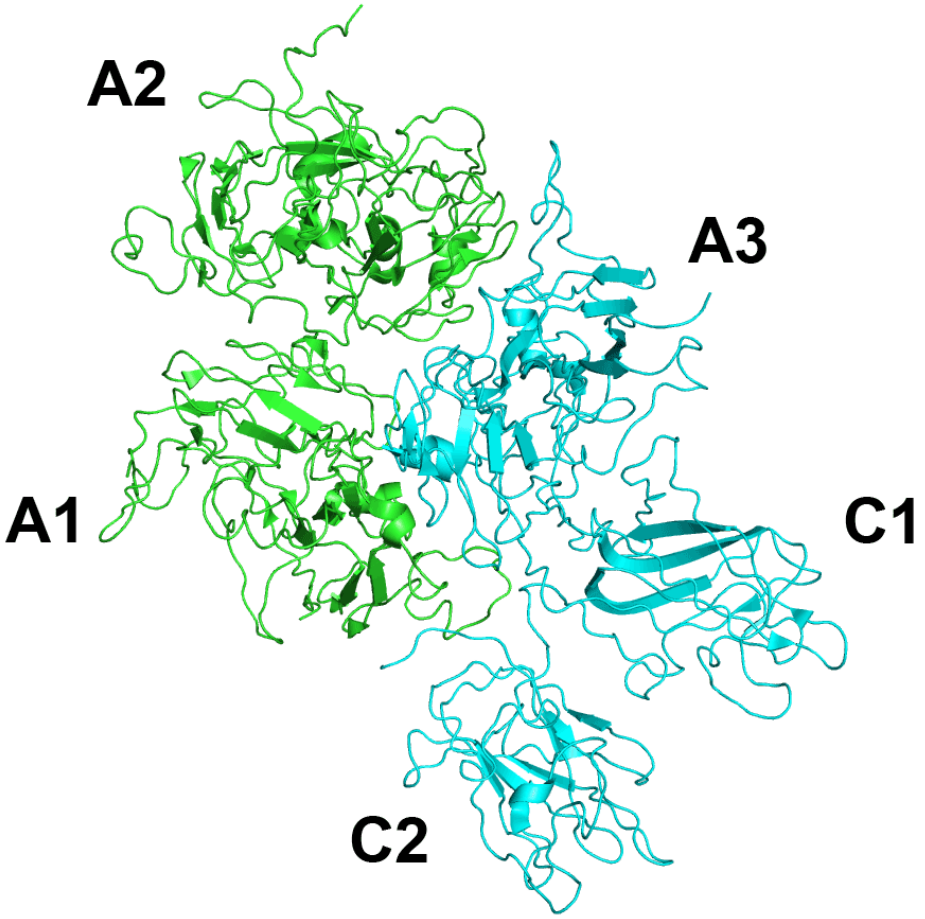


Figure 1.2: Structure of activated factor VIII (FVIIIa); green: Heavy chain; blue: light-chain (reproduced from [15])

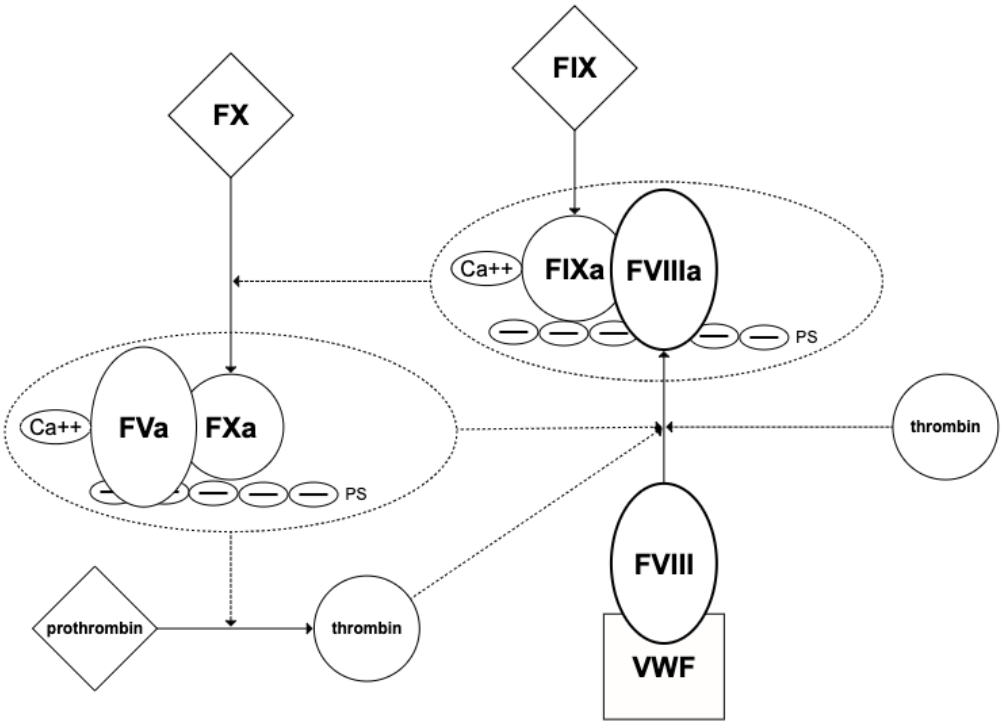


Figure 1.3: The FVIII positive feedback loop: FVIII dissociates and gets activated by thrombin and by FXa. It forms a complex with FIXa on the phospholipid surface of platelets (PS) in a calcium dependent manner. This complex accelerates the activation of FX by 10^5 fold. Activated FX in complex with FVa further activates more FVIII and thrombin. Dashed lines: proteolytic activity; dashed ovals: complexes (adapted from [16])

1.1.3 Haemophilia A therapy depends on the severity level

Before the 20th century, blood transfusion was the only means of treating prolonged bleeding which was often fatal for patients since transfused blood did not provide enough clotting factor. The application of raw blood or plasma was then more and more replaced by specific therapeutics mitigating the effect of the missing factor in the coagulation cascade which lead to the usage of clotting factor VIII derived from plasma in 1955 for the treatment of haemophilia A patients. With it came a great increase in life-quality and life-expectancy for haemophiliacs but this was sadly reversed when it became apparent that in many cases replacement clotting factor was contaminated by blood-borne viruses, mainly HIV and hepatitis C [17]. Different techniques were developed to deactivate viruses in plasma but with the

Percentage of functional FVIII (relative to normal)	classification and implications
<1	Severe: From early life on regular bleeds into joints, internal organs, muscles; if untreated joint deformations, crippling
1-5	Moderate: spontaneous bleeds possible, bleeding after slight injuries
>5	Mild: Bleeding during surgery and after major injuries only

Table 1.1: Overview of severity classification (reproduced from [20])

ability to produce recombinant factor VIII since 1992 this became the standard of care in developed countries.

In practice, three different levels of severity are distinguished depending on the percentage of functional coagulation factor. For an overview see table 1.1. Mild haemophilia A can be treated on-demand by administering 1-Diamino-8-D-arginine (DDAVP) which raises the release of FVIII from endothelial cells. However a prophylactic treatment seems to be a better choice for all patient groups. Long-term patient health as well as quality of life are superior over on demand solutions, especially for children, since insufficient levels of FVIII increase incidences of crippling haemarthroses in this patient group [18, 19]. For haemophilia A there is an increasing number of prophylactic treatment options available. The NHS recommends prophylactic treatment with Emicizumab, a bispecific antibody mimicking the function of FVIIIa by forming a complex with both FIX and FX [21, 22]. A very promising direction is gene therapy that has been successfully applied to cure haemophilia B and should soon be available for haemophilia A patients [23, 24], although this option might not be initially available for patients that are at risk of developing inhibitors. An overview and outlook of therapies can be found in Hoffbrand's excellent guide on haematology [7].

1.1.4 Inhibitory antibodies to factor VIII complicate therapy

Today, the most significant complication associated with the treatment of haemophilia A is the development of inhibitory antibodies (inhibitors) that neutralise the pro-coagulant function of the replacement FVIII products [25]. Inhibitor formation occurs in around 30% of individuals with severe haemophilia A and is accompanied by a reduced quality of life and life expectancy as well as increased costs per patient. With the proper treatment these drawbacks can be mitigated [26, 27].

1.1.4.1 Inhibitors mask functional surfaces on the FVIII C2-domain

Inhibitors form as a result of a MHC class II cellular immune cascade which involves proliferation of CD4+ T lymphocytes that recognize antigens on different domains of the FVIII molecule. This mechanism is also taking place in immune systems of non-haemophiliacs but is hindered by natural anti-idiotypic antibodies. Current research suggests that the C2-domain of FVIII in its inactive form is a prime target for inhibitory antibodies [28]. The inhibition at this location is induced by masking functional epitopes for VWF and phospholipid membranes binding as will be discussed in detail in section 1.1.4.4. Different potential risk factors have been reported that promote inhibitor formation spanning from the preparation of therapeutics to the state of patients [29, 30]. Combined with the varying epitopes that get recognized it is challenging to consolidate the different research results, which are often based on small cohorts because of the rareness of haemophilia A, to find an immunologic answer to prevent inhibitor development. Further, inhibitor prevention gets complicated by the different mechanisms at play in congenital and acquired haemophilia, and differences between patients who create some FVIII (functional or non-functional) and those who create none [31].

1.1.4.2 Immune tolerance induction is commonly used to treat patients with inhibitors

The treatment of patients with inhibitors is a major challenge; the most effective therapy differs between patients and is hard to predict, and new therapeutic options

are needed for those patients who respond poorly to existing treatment regimens [32]. To prevent inhibitor formation in the first place, replacement FVIII risks are considered as well as patient history and ancestry [18, 33]. Immune tolerance induction (ITI) by administration of high doses of recombinant factor VIII is the standard therapeutic approach today and recommended by the National Health service UK and US authorities [21, 34]. There are endeavours to provide a more targeted ITI, choosing case-specific products and doses, to eventually prepare inhibitor patients for gene therapy and also reduce costs [35].

1.1.4.3 A range of FVIII replacement products is available today

A field of interest in haemophilia A research is the development of therapeutic FVIII products with reduced antigenicity. Porcine FVIII, which has 24 amino-acid differences to the human FVIII C2-domain alone is associated with reduced antigenicity [36] and exhibits low cross-reactivity from anti-human FVIII antibodies [37], has been used therapeutically for decades, initially as a plasma-derived product and more recently in recombinant form [38, 39]. Various hybrid human/porcine FVIII molecules have been developed [40, 41] and there is continuing interest in understanding how non-porcine FVIII orthologs may aid the development of improved FVIII products [42]. A complementary strategy for reducing antigenicity involves mutating residues in key B-cell epitopes of human FVIII [43, 44, 45]. A newer strain of research are bio-engineered factor VIII products that are found to be less immunogenic, e.g. the fusion protein rFVIIIFc in which the B-domain of FVIII is replaced by the Fc domain of human IgG1 [46] or by binding the light and heavy chain of FVIII covalently to form a recombinant B-domain-truncated factor VIII single chain [47]. Other research is focusing on developing so called 'bypassing agents' that, for example, imitate the function of activated factor VIII by forming a complex between factor IXa and factor X [22]. An overview of current endeavors can be found in the summary of Pipe and co-workers [48]. Given the growing emphasis on personalised and precision medicine, it seems reasonable to expect that target-oriented and highly specialized approaches will become increasingly relevant [49]. Further, recent breakthroughs in gene-therapy for haemophilia B [50]

and haemophilia A [51] add another emerging treatment option. A better understanding of inhibitors and their epitopes is likely to be important both in the design of new therapeutics and in aiding the selection of the best treatment regimen for a given individual [27, 48, 52].

1.1.4.4 Multiple studies identify areas of antigenic residues on the FVIII C2-domain

Functional epitopes of the FVIII protein are found mostly in the A2 and C2 domains which are responsible for binding to von Willebrand factor, phospholipid membranes and FIXa[28]. Inhibitors mask these regions and induce the clearance of inactive FVIII from the blood stream, inhibit the formation of the tenase complex or the activation of FVIII by thrombin or FXa and cause the disruption of the positive feedback loop described in section 1.1.2.

The 15 kDa C2-domain of FVIII has been described as a prime target for inhibitors by Prescott and co-workers [53]. Research efforts followed with the goal to characterize mechanisms of inhibition and locate antibody epitopes.

The first detailed insights into C2-domain specific antibodies was provided by Shima and co-workers [54]. They narrowed down antigenic regions for two inhibitors to residues 2170-2327 and 2248-2312 (figure 1.8); bearing in mind that the C2-domain of the FVIII sequence runs from Ser2173 through to Tyr2332. It has been shown that the recognition of antibodies in these regions inhibit the non-covalent complex both with VWF and with phospholipid membranes.

Subsequently, Scandella and co-workers investigated six human inhibitors that were found to bind to the isolated FVIII C2-domain [55]. An overlap of antibodies with C2-domain surfaces responsible for binding to phospholipid membranes was determined as the cause for rendering FVIII dysfunctional. They further located the core region of antigenic recognition to residues 2248 to 2312 (figure 1.8).

Detailed research of another C2-domain specific antibody, the IgG4 human monoclonal inhibitor BO2C11, has been carried out by Jacquemin and co-workers [56]. They discovered that inhibition takes place through hindrance of bond formation with VWF as well as with phospholipid membranes. Surface plasmon reso-

nance (SPR) measurements indicate that BO2C11 has a very low off rate compared to VWF [56]. This suggests that after dissociation from VWF the association of BO2C11 with FVIII is virtually non-reversible. BO2C11 was further found to recognize an antigenic region spanning from residue 2173 to 2332 (figure 1.8).

In the same year, Healey and co-workers mapped a hot-spot for inhibitory antibodies against therapeutic (alloantibodies) and patients' own FVIII (autoantibodies) on the C2-domain [57]. Their study design made use of the different binding behaviour of inhibitors between human and porcine FVIII. Introduction of the porcine sequence Glu2181-Val2243 to human FVIII made it significantly less antigenic for a series of 6 C2-domain specific inhibitors.

The C2-domain as a prime target for inhibitors was also confirmed in a clinical study of Laub and co-workers [58]. The study reports that the majority of FVIII inhibitors bound to the C2-domain in a study of patients treated with injections of recombinant FVIII in Germany with similar results reported for Belgian patients [59].

The first structure of the C2-domain was then published by Pratt and co-workers with a resolution of 1.5 $\AA$ [60]. The amino-acid sequence of the FVIII C2-domain crystal-structure (PDB 1d7p) differs from wild-type recombinant FVIII owing to the introduction of substitution S2296C for crystallographic purposes (mercury derivatization). The topology of surface residues indicates that the binding of phospholipid membranes involves the hydrophobic region consisting of two β-hairpins (M2199/F2200 and L2251/L2252) and a loop region containing V2223. At least 4 basic residues - R2215, R2220, R2320, K2249 - are found at or in the vicinity of that region. The docking to phospholipid membranes is believed to involve a combination of hydrophobic and electrostatic contributions.

This work was followed by solution of the structure of the BO2C11 fab fragment bound to the FVIII C2-domain by Spiegel and co-workers (PDB 1iqd) which also contained the mutation S2296C [61]. The structure suggests that negatively charged aspartic acids in the antigen recognizing site form salt links with residues R2220 and R2215. These two positively charged residues are also proposed to

interact with negatively charged surfaces on the binding of FVIII to phospholipid membranes. For the β-hairpins M2199/F2200 and L2251/L2252, polar, hydrophobic as well as van der Waals interactions have been determined to contribute to binding affinity. The authors also point out that, compared to the apo conformation of the FVIII C2-domain reported by Pratt, the holo (bound) conformation contains a twist of about 90° in the β-hairpin M2199/F2200.

Research carried out by Barrow and co-workers [62] was inspired by differences in the amino-acid sequence of murine, porcine, canine and human FVIII, and investigated phospholipid membrane binding and titers of 5 C2-domain specific polyclonal human antibodies, including the human antibody BO2C11, by using a one-stage clotting-assay. They incubated their B-domain deleted (BDD)-FVIII wild-type and mutants with increasing levels of inhibitors until 50 percent procoagulant activity was lost. This measurement, called the Bethesda unit, reports the potency of the antibody in hindering function but could also be seen as a proxy for binding affinity. The fact that increasing levels of inhibitor influence function gives evidence that the impairment is attributable to antigenicity and structure rather than structure alone. Some of their mutants, such as F2200L, remain functional even with greater quantities of BO2C11 which suggest that antigenicity was greatly decreased, whereas M2199I loses its coagulant capabilities with lower levels of BO2C11 than for the wild-type (table 1.5). It is therefore reasonable to assume that M2199I has higher affinity to BO2C11 than the wild-type.

Meeks and co-workers produced 30 (29 mouse, 1 human) monoclonal antibodies (MABs) and let them bind competitively to FVIII in an enzyme-linked immunosorbent assay (ELISA) and further report FVIII inhibition titers using the Bethesda assay for selected antibodies [63]. Their focus was to identify antibody epitopes on the FVIII C2-domain and to analyse the effect of introduced mutations. By letting antibodies compete over binding they could form 5 groups of antibodies that belong to 3 core and 2 overlapping epitope regions on the C2-domain (figure 1.6). It showed that the human antibody BO2C11 was located in group AB along with 6 mouse antibodies, making the epitopic region belonging to group

AB the second most common target in this study. For antibody I109 belonging to the same antigen-recognizing group as BO2C11 they detected that the double mutant M2199I/F2200L on the β-hairpin produces a loss in antigenicity in a Bethesda assay. Whether the double mutant alters the affinity of the FVIII C2-domain to phospholipid membranes has not been validated in this study. Still, since I109's proposed inhibition mechanism includes the masking of the binding site of phospholipid membranes, this finding adds to the observation by Spiegel *et al.* [61] that positively charged surfaces in the vicinity of the M2199/F2200 β-hairpin are important for protein function.

Table 1.3 shows the results of the thermodynamic van't Hoff analysis for 9 mutants and the wild-type carried out by Lin *et al.* [44]. It showed that for the wild-type and most mutants the entropic contribution $T\Delta S$ is much larger than the enthalpic contribution ΔH. Hence the binding is predominantly entropically driven. After the identification of M2199 and F2196 mutants as promising in terms of destabilizing the binding to BO2C11 and further retaining the ability to bind to phospholipid membranes proven in a one stage clotting assay, they produced the BDD wild-type (WT), BDD-F2196K and BDD-M2199A and report K_D values for binding to VWF and BO2C11 (table 1.4). They identified that M2199A and F2196K have comparable binding affinities to VWF as the WT and decreased affinity to BO2C11.
BO2C11 blocks FVIII binding to von Willebrand factor and phospholipid membranes [56] and binds to a region of the C2-domain targeted by multiple human inhibitors [57, 64, 65], including three that share the VH1-24 germline with BO2C11 [64].

So far, computational studies involving the FVIII C2-domain have been carried out to determine differences between the zymogenic and activated form [66], to describe membrane binding mechanisms [67, 68] and for a virtual screening of small ligands [69].

FVIII-C2 variant	$K_D=k_d/k_a$ (pM)
WT-FVIII-C2	9
F2196A	147
T2197A	10
N2198A	56
M2199A	$k_a > 3.0 \times 10^7$
F2200A	240
R2215A	150
R2220A	NB
R2220Q	NB
Q2222A	5
V2223M	6
S2250A	22
L2251A	6
L2252A	6
T2253A	10
Q2311A	3
H2315A	5
Q2316A	25

Table 1.2: K_D values of the WT and mutants of the BO2C11 complex with the FVIII C2-domain determined by surface plasmon resonance measurements (reproduced from Lin and co-workers[44]). The association constant of M2199A was over the physical limit of the instrument. The cause of non-binding of mutations R2220A and R2220Q has not been determined in the literature yet; Nguyen *et al.* propose a large-scale change in conformation [70]. Reported errors were below 2.5 pM except for N2198A with 14 pM due to an unusually high k_d value.

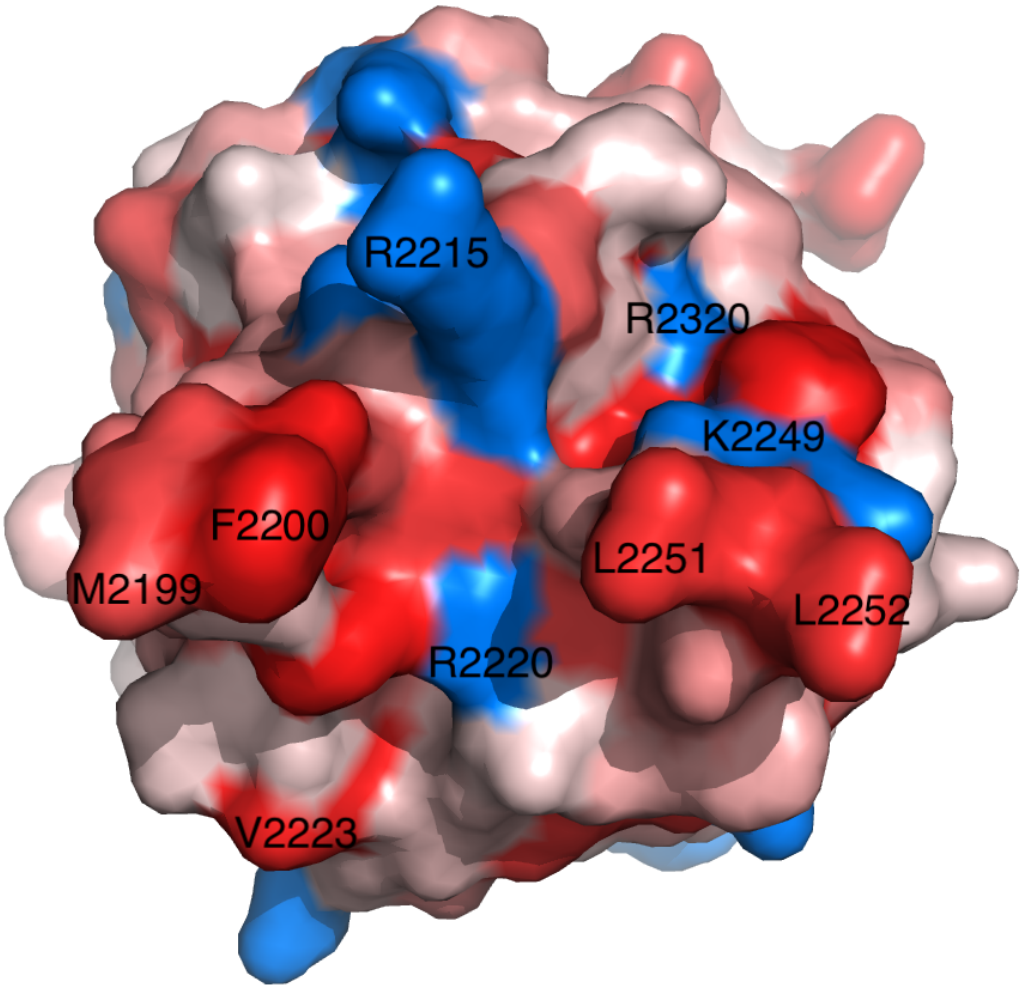


Figure 1.4: Crystal-structure of the apo FVIII C2-domain solved by Pratt *et al.* [60]; Hydrophobic residues (red) especially those at the tip of β-hairpins as well as positively charged residues (blue) are suggested to contribute to phospholipid binding.

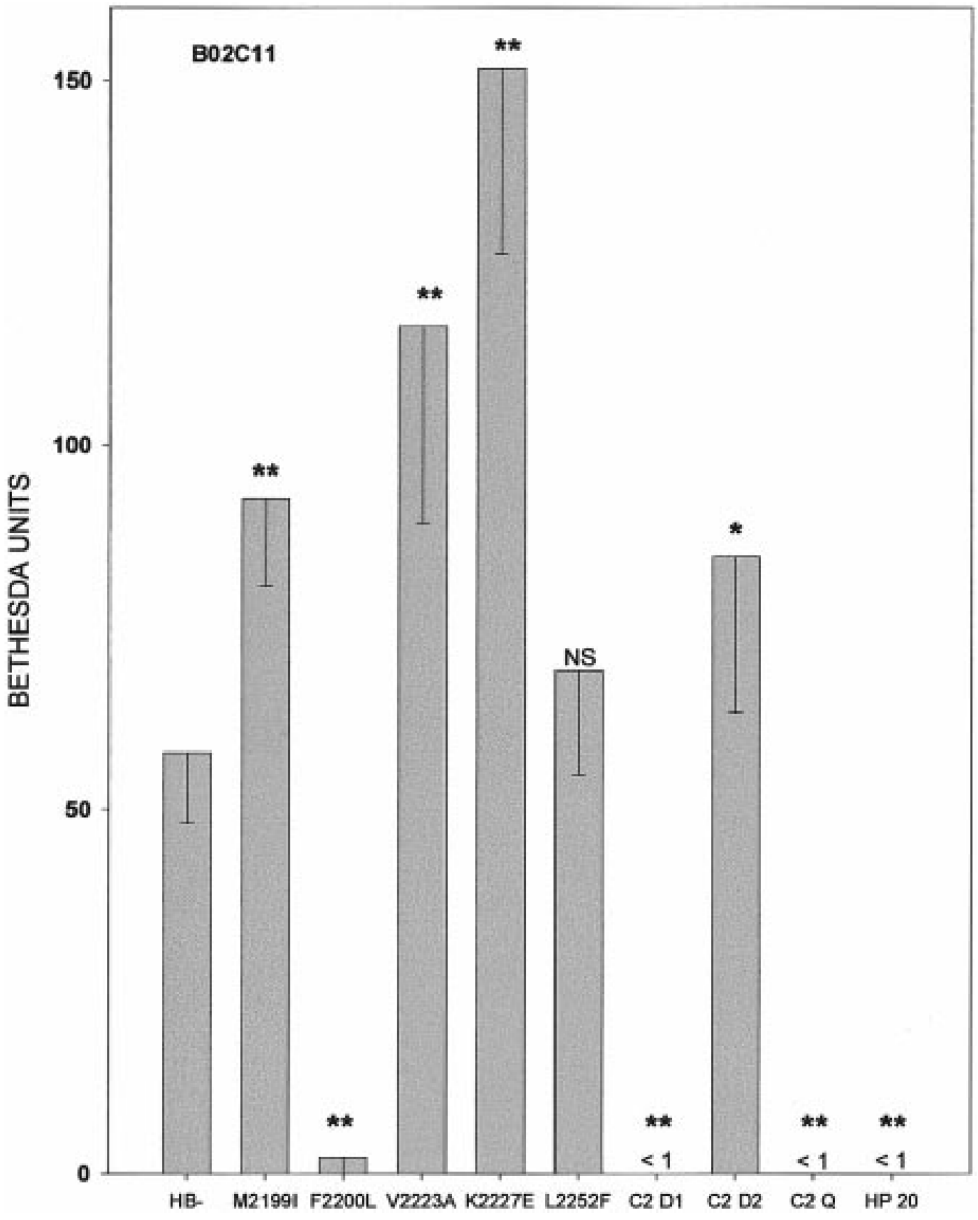


Figure 1.5: Modified Bethesda assay conducted by Barrow *et al.* [62] of the B-domain deleted WT FVIII (HB-), with point mutations (M2199I, F2200L, ..) as well as with multiple mutations introduced (C2 D1, C2 D2 ...). Point mutation M2199I has a higher Bethesda unit than HB-, meaning that lower concentrations of antibody BO2C11 are needed for FVIII to lose 50% of its coagulant function, which in turn indicates that this substitution might somewhat increase binding affinity. On the other hand, F2200L has been found to be less antigenic than the WT (reproduced from [62]).

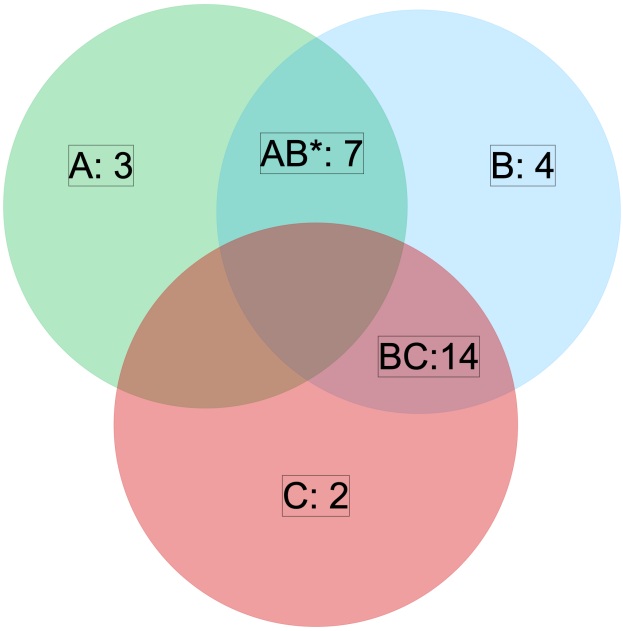


Figure 1.6: Antibody grouping by epitope regions as described by Meeks *et al.* (reproduced from [63]); *Group AB contains BO2C11, the only human antibody that was part of the ELISA experiment, which has been found to overlap with antigenic regions of both group A and B.

FVIII-C2 variant	$\Delta H_A°$(kJ/mol)	$T\Delta S_A°$ (kJ/mol)	K_D (pM)
WT-FVIII-C2	-14	54	1
F2196A	-18	38	154
T2197A	-12	52	6
N2198A	-13	45	68
M2199A	-16	44	30
F2200A	-18	37	230
R2215A	-18	49	100
S2250A	-3	57	30
L2251A	-31	35	3
L2252A	-49	14	9

Table 1.3: Thermodynamic (van't Hoff) analysis of binding site affinity carried out by Lin and co-workers (reproduced from [44]). Measurements were taken over a range of 10°C to 40°C with 5°C increments for mutations that are at or over the limits of the equipment as measured in table 1.2. It appeared that binding is mostly entropically driven (high $T\Delta S_A°$ values). Measuring errors were reported as under 1 kJ/mol, however N2198A showed the most variability in the included van't Hoff plots (refer to Figure 2 in the original paper [44])

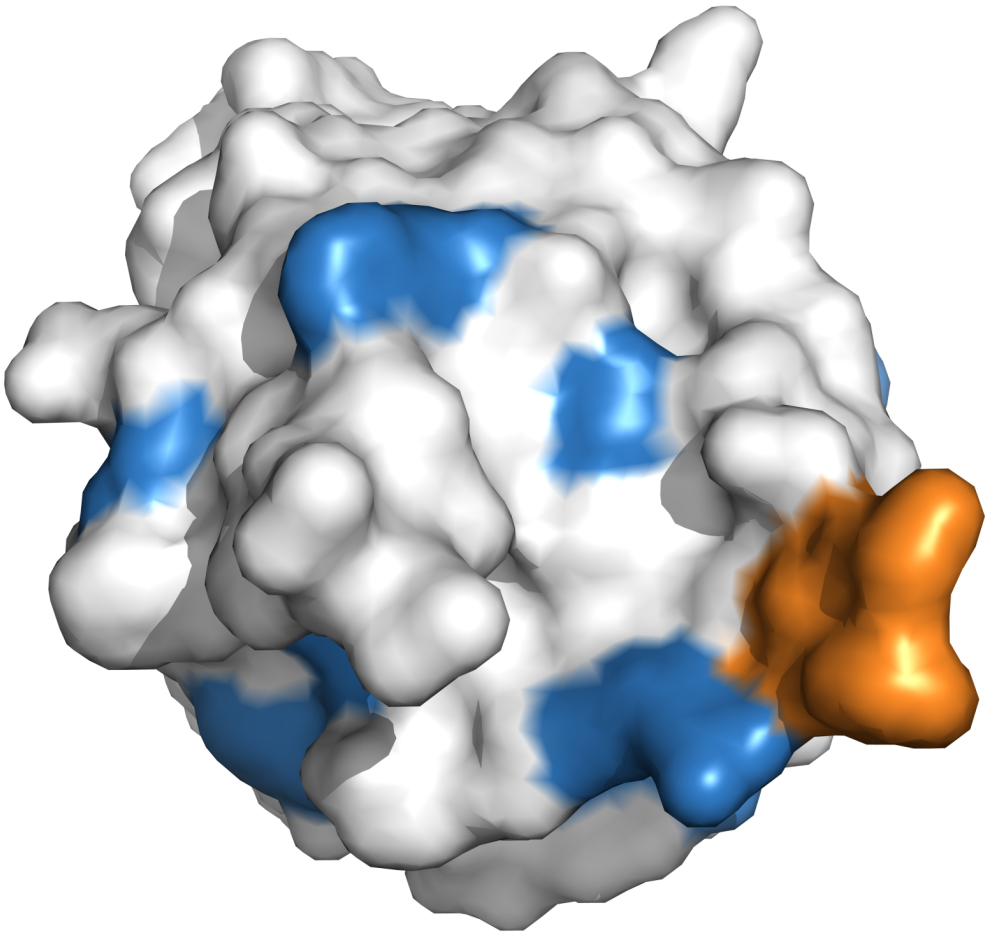


Figure 1.7: Positively charged surfaces (blue) in the vicinity of β-hairpin M2199/F2200 (orange); antibody I109 belonging to the same antigen-recognizing group as antibody BO2C11 inhibits the function of FVIII by masking VWF and phospholipid membrane binding sites on the C2-domain; the double mutant M2199I/F2200L was found to abrogate binding to antibody I109 [63]. If this mutant influences the binding to phospholipid membranes has not been investigated in this study. The structural analysis of Spiegel *et al.* [61] points out the importance of the β-hairpin M2199AF2200 since positively charged surfaces are located in its vicinity that are supposed to bind to negatively charged regions found on phospholipid membranes.

Complex	$K_D=k_d/k_a$ (pM)
WT-BDD-FVIII/VWF	290
BDD-FVIII-F2196K/VWF	200
BDD-FVIII-M2199A/VWF	250
WT-BDD-FVIII/BO2C11	21
BDD-FVIII-F2196K/BO2C11	580
BDD-FVIII-M2199A/BO2C11	92

Table 1.4: B-domain deleted FVIII affinities to BO2C11 and to VWF (reproduced from [44]). Lin *et al.* have shown that mutations F2196K and M2199A decrease antigenicity while retaining the ability to bind to VWF.

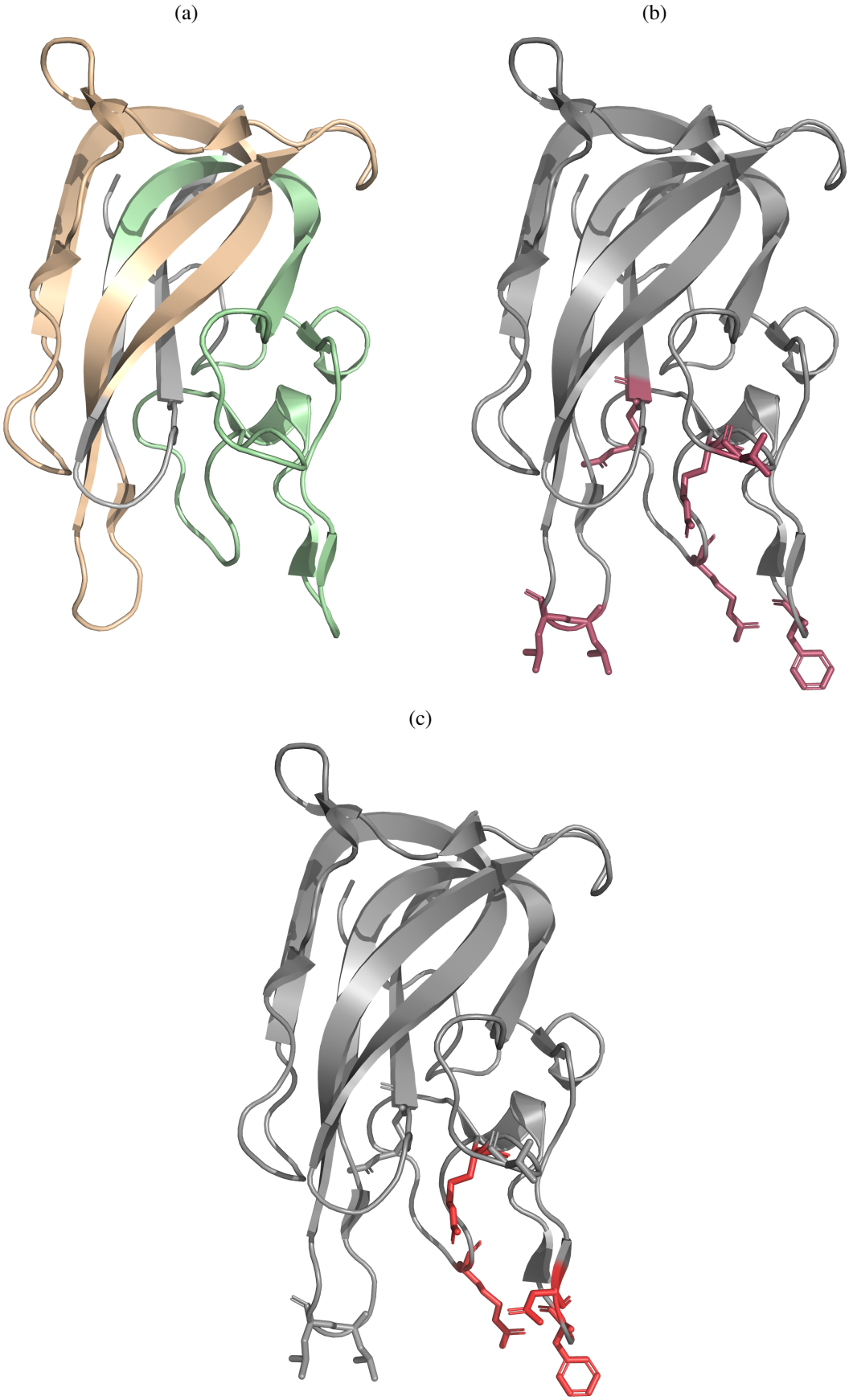


Figure 1.8: Antigenic regions on the FVIII C2-domain identified by (a) gold: Shima *et al.* [54] and Scandella *et al.* [55]; green: Jacquemin *et al.* [56] and Healey *et al.* [57] (b) plum: Pratt *et al.* [60] and (c) red: Lin *et al.* [44]

1.2 Molecular dynamics simulations to investigate protein dynamics

Molecular dynamics (MD) is a computational simulation technique dating to the 1950s, with the first simulation of a protein in 1977 [71, 72]. Its attractiveness lies in the level of spatio-temporal detail, giving a complete description of the system at atomic resolution at user-defined time-steps. The ongoing advance of computational power available from commodity hardware, especially the advent of graphical processing units (GPUs) that speed up computing by performing calculations in a highly parallelised manner, provides the computational power to simulate time spans that cover some biologically relevant processes [73]. MD is typically used with a molecular mechanics interpretation of forces between atoms, that are comprised of van der Waals, bond length and angle as well as dihedral angle and electrostatic interactions.

MD proved to be a sensible additional route of investigation in the domain of protein folding [74], regulatory processes [75], structure refinement [76], and gave insights into transient states [77].

Since most biological macromolecules exist in a state close to equilibrium with their environment, MD supplements experimental techniques that are only able to capture snapshots of averaged ensembles on the atomic level. Information about dynamic conformational changes has improved the understanding of ligand binding motifs in many cases [73].

In addition to analysing conformational system states, it is possible to analyse the coordinates and calculated forces produced by molecular dynamics simulations to get insights into free energy of binding or, more generally, into energetic differences between system states. By these means, MD is able to score/rank protein complexes, which has made it a standard tool in drug-development [73, 78, 79, 80, 81, 82, 83, 84, 85].

For an MD research project, it is necessary to have an atomic structure of the system under investigation. The Protein Databank holds more than 150,000 structures that could potentially be used for MD [86]. Given a structure, key choices

need to be made concerning the force field, water model, simulation engine and length of simulation.

1.2.1 Force Fields: The basis for calculating interatomic forces

As mentioned above, molecular mechanics distinguishes five different energy terms. A summation of these energies is generally referred to as a force field and gives the total potential energy E_{total} of an atomistic system:

$$\begin{aligned} E_{total} &= E_{bonds} + E_{angles} + E_{dihedrals} + E_{non-covalent} \\ E_{non-covalent} &= E_{vdW} + E_{electrostatics} \end{aligned} \tag{1.1}$$

where E_{vdW} is the van der Waals energy contribution.

The force field used in this work is the ff14SB which is the recommendation for protein dynamics by AMBER [87, 88] and has been shown to produce meaningful results in many cases [89, 90, 91]. This force field represents the refinement of the ff99SB force field in that its side-chain dihedral parameters were updated based on empirical findings. All AMBER force fields share the following functional form [92]:

$$\begin{aligned} E_{total} = \sum_{bonds} k_b(l-l_0)^2 + \sum_{angles} k_a(\theta-\theta_0)^2 + \sum_{dihedral}\sum_{n}\frac{V_n}{2}(1+cos(n\theta'-\xi)) \\ + \sum_{j=1}^{N-1}\sum_{i=j+1}^{N} k_{vdw}\left[(\frac{r_{0,ij}}{r_{ij}})^{12} - 2(\frac{r_{0,ij}}{r_{ij}})^{6}\right] + \sum_{j=1}^{N-1}\sum_{i=j+1}^{N}\frac{C}{\varepsilon_0}\frac{q_i q_j}{r_{ij}} \end{aligned} \tag{1.2}$$

Covalently bonded atoms are modelled using a harmonic description k_b that is using the bond length l minus its distance from equilibrium l_0 as input. The same is done for triplets of atoms in the calculation of bond angle energy. Van der Waals energy is modelled using a Lennard-Jones 12-6 potential where the distance of two atoms at the minimum of the potential $r_{0,ij}$ is divided by the actual distance of these atoms r_{ij}. Energy stemming from the 4-body dihedral angle θ' is modelled with a Fourier series and a phase shift of ξ. The electrostatic potential depends on the atomic charges q and is modulated by the Coulomb constant C and the dielectric

constant ε_0.

The calculation of the potential energy by a force field as the one mentioned that uses the coordinates of atoms as input is a computationally expensive task. Short integration time steps of 1 femtosecond are needed for the accurate numerical integration of the equation of motion for hydrogen bond vibrations [93]. Fortunately, hydrogen bonds can be constrained by using fixed ideal values which makes it possible to sample the system at larger time intervals with neglectable effects on geometry [88, 94, 95]. SHAKE [94] and LINCS [96] are two algorithms that are used frequently to decrease the sampling rate from 1 to 2 femtoseconds with SHAKE being used through this work. Another method to increase the length of time steps is 'virtual sites' which defines virtual interaction sites for hydrogens that account forces of hydrogen interactions to the closest heavy atoms [97]. Hydrogen mass repartitioning is another method to enhance the sampling of hydrogen that assigns a proportion of the mass of a heavy atom to its covalently bonded hydrogen atom. Due to the increase of hydrogen mass it is possible to increase the time step of integration to 4 fs and more [98]. Still, the real-timescale of motion of larger biomolecules such as e.g. loop motion is in the order of nano seconds and therefore a vast amount of time steps has to be calculated [99].

1.2.2 Water can be represented at varying levels of detail

When it comes to water models, a trade off has to be made. Very accurate explicit solvent models add water molecules that are simulated in atomistic resolution. These models exist with differing degrees of complexity concerning inter- and intramolecular modelling but all have the disadvantage of being computationally expensive. On the other hand, implicit solvation approximates the effect of a solvent by a solvation energy term that is added to the force field equation as outlined in detail in section 1.2.2.2. Both models are considered appropriate for specific tasks and are used extensively [100].

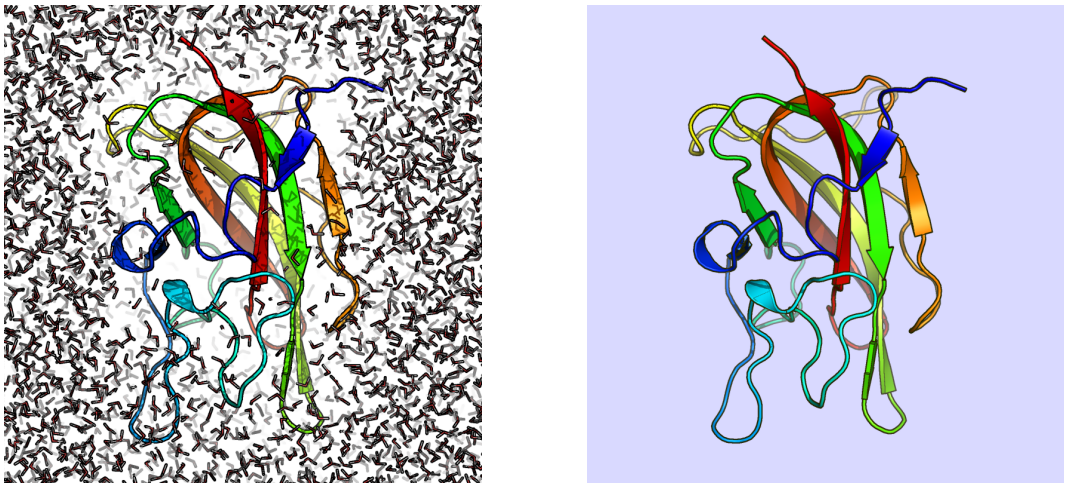


Figure 1.9: Left: Solvation with explicit (discrete) water molecules Right: Water bath represented by a continuum energy term

1.2.2.1 Explicit solvation best resembles physiological conditions

Explicit solvation gives more accurate results but might not be feasible for bigger systems. For example, the protein complex investigated in this work consists of roughly 10,000 atoms including hydrogens. After explicit solvation the system contains around 200,000 atoms. A solvation of larger structures might come with an even greater increase in atom count depending on the geometry of the solute. Because of the previously mentioned increases in computational power due to GPUs it is nonetheless possible to run expressive explicit solvent simulations in reasonable real-time.

Explicit solvation is used in this work for the equilibration productions of crystal-structures and further to produce decorrelated frames for MM/GBSA calculations (chapter 1.5). A popular choice that has been used here for explicit solvation is the TIP3P water model [101, 100]. It represents a water molecule by a rigid 3-body entity that has empirically derived distances and angles and takes into account the point-charges of hydrogen and the oxygen [102, 101]. A simplified force field equation is used to calculate the pairwise energy of water molecules, with the goal to make the calculation of the solvent more efficient:

$$E = \sum_{pairs} \left(\frac{A_{LJ}}{{r_{oo}}^{12}} - \frac{B_{LJ}}{{r_{oo}}^{6}} + C\frac{q_i q_j}{r_{ij}} \right) \tag{1.3}$$

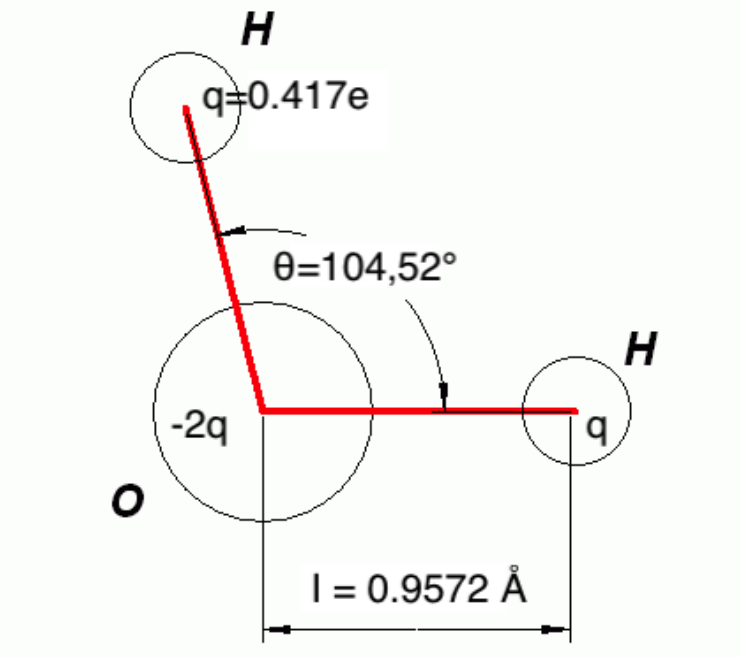


Figure 1.10: Parameters and geometry of the TIP3P water model. The three body system has a negative charge in the centre that is two times that of its hydrogen point charges. Distance, angle and charges are empirically derived quantities.

where r_{oo} is the distance of two oxygen atoms with Lennard-Jones parameters A_{LJ}, B_{LJ}, C the Coulomb constant and q_i, q_j charged sites of distance r_{ij} (cut-off 9Å) [103]. In comparison to the FF14SB formulation, here, covalent bonds are not considered at all. Non-covalent interactions are approximated by interactions of oxygen and a single charged site.

TIP4P is another rigid water model and an advancement of the TIP3P model. To better describe the electrostatic properties of water, TIP4P has a Lennard-Jones interaction site for the oxygen, charged sites at the hydrogen and introduces an additional negative charge M in the center of the molecule [101].

Another approach that uses explicit solvent are all-atom force fields which differ from a united-atom force fields like ff14SB in that they incorporate the calculation of the solvent and do more accurately model the solvent as e.g. the solvent model TIP3P. All-atom force fields such as ff03ws are especially well suited for simulations of folding and intrinsically disordered proteins [104, 105].

1.2.2.2 Representing solvent as a continuum speeds up calculation

Implicit solvation models or continuum solvent models are generally less accurate than explicit models but have been found useful for approximating solvation energy of receptor-ligand complexes or reducing computational demands in the case of resource intensive simulations like Replica Exchange MD [102, 106]. Their big advantage over explicit solvation models lies in the representation of solvent as continuous medium, which reduces the requirement of calculating energies for the numerous explicit water molecules [100]. Different approaches exists with the main branches Poisson-Boltzmann model (PB) and Generalized Born model (GB). PB is the exact and theoretically sound way to describe the solvent as a dielectric medium but is very expensive to calculate. GB is the upp to the more exact (PB) equation which it is less computationally expensive. It further showed that the approximate but accelerated calculation of the PB equation using numerical solvers produced less accurate results than the GB model [107]. Hence, it became the more popular choice for bigger molecules like proteins and DNA [108, 109, 110].
The implementation of AMBER's Molecular Mechanics Generalized Born Surface Area method (MM/GBSA) splits the calculation of solvation energy into polar/electrostatic and non-polar contributions:

$$G_{solvation} = G_{pol} + G_{np} \tag{1.4}$$

Calculation of G_{pol}, the polar solvation energy, is a non-trivial task, due to polarization effects on the boundary between solvent, having a high dielectric constant, and the solute with a low dielectric constant. The polar solvation energy is approximated by solving the GB implicit solvent model (see section 1.5).

$$\begin{aligned} \Delta G_{pol} \approx -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}(r_{ij},R_i,R_j)}\left(1-\frac{exp(-\kappa f_{GB})}{\varepsilon}\right) \\ f_{GB} = [r_{ij}^2 + R_i R_j exp(-r_{ij}^2/4R_i R_j)]^{1/2} \end{aligned} \tag{1.5}$$

Important concepts in GB model 1.5 are that every atom is represented as a sphere of radius R_i with a charge q_i. The atoms of the solute are supposed to be

filled with a material that has a dielectric constant, in this work $\varepsilon = 1$ was chosen (explanation see section 2.1.3), whereas solvent atoms have a high dielectric constant, with $\varepsilon = 80$ being the widely adopted choice for water at 300K.

Atomic radii R_i, R_j are refined in 1.6 by considering the degree of burial of an atom within the solute and hence the term 'effective Born radii'. f_{GB} smooths R_i, R_j by taking into account the distance of two atoms r_{ij}. κ is the Debye-Huckel screening parameter that introduces the interaction energy of ions in the solution to the calculation.

$$\begin{aligned} R_i^{-1} &= \overline{\rho}_i^{-1} - tanh(\alpha\Psi - \beta\Psi^2 - \gamma\Psi^3)/\rho_i \\ \overline{\rho}_i &= \rho_i - offset \\ \Psi &= I\overline{\rho}_i \end{aligned} \tag{1.6}$$

Subtracting an $offset$ from the dielectric radius of an atom ρ_i comes from the notion that its dielectric potential alters the $effective$ radius of a water molecule. The default value of AMBER is 0.09Å. I calculates the degree of burial of an atom in the solute, thereby excluding atoms that do not interact with the solvent. This can however introduce errors in cases where structures have cavities. The model assumes solvent accessibility even for cavities too small for a water molecule to fit or that are enclosed. The formula for I and the remaining parameters for the model that have been optimized by Onufriev *et al.* can be found in their works [108, 111]. The exact values for their 'model I' parameters are $\alpha = 0.8$, $\beta = 0$, $\gamma = 2.909125$.
The non-polar contribution of solvation is a combination of the van der Waals attraction between solute and solvent and the repelling force of altering the solvent's structure. This is approximated in the GB approach of AMBER to have a linear relationship with surface accessible surface area (SASA) of a structure:

$$G_{np} = \gamma \cdot SASA + const. \tag{1.7}$$

This equation is called the 'cavity' term since it calculates the energy needed to create a cavity in the solvent to fit the solute. It combines the calculated SASA, as

in [112], of the solute and the surface tension γ of the solvent. For GB 'model I' of AMBER the exact values are $\gamma = 0.005$ and $const. = 0$.
For the hydrophobic entropy S that should by the definition of free energy be part of the equation 1.4 there is no method to speak of to approximate its contribution and therefore it is not part of the AMBER formulation. The only relation to hydrophobic entropy is through the empirically derived constant $const.$ and surface tension term γ in 1.7 in the formulation of non-polar solvation free energy. Entropy calculation is clearly a weakness that comes with the method and is stated to "contribute to the largest fluctuations in the overall free energy" [113] which in turn can lead to unpredictable performance of the method, e.g. when calculated binding free energies are compared to experimental results [114].

1.2.3 Thermostats and barostats provide a first impression of thermodynamic stability

Before the investigation of a phenomenon by the analysis of a simulation, it is common practice to grant crystal-structures a generous so called equilibration time until a relatively stable thermodynamic state has been reached and the structure has settled under the novel condition, which include alterations to the structure and/or solvation and changes in temperature and pressure. An initial idea about such a stable state in a NPT simulation (constant number of atoms, constant pressure, constant temperature) as the one carried out in this work can be given by the curves of the thermostat and barostat. After the heating of the system the curve of the thermostat, which gives information about fluctuations in temperature, and the size of the box which is regulated by the barostat, should converge. Another measurement that indicates whether the system has reached a thermodynamically stable state is by investigating the potential, kinetic and total energies. However, these quantities provide only global information about the dynamics of the structure under investigation and usually a more fine grained picture is aspired, which will be discussed in section 2.1.2.

1.2.4 Simulation Engines differ in usability and accessibility

A variety of simulation softwares exists that are optimized to efficiently calculate atomistic forces mentioned above. AMBER [88], GROMACS[115], NAMD[116], CHARMM[117] and DESMOND[118] are examples of widely used simulation engines. Criteria that determine the choice of an engine include the ease of use, the speed of calculation as well as the costs for a licence. GROMACS is open-source software that is very well curated and has an active user community. The simulation engine AMBER is proprietary with comparable software maintenance and online support as GROMACS. Depending on the GPU used for simulations, AMBER outperforms GROMACS whereas GROMACS might have advantages when run on CPUs [119, 120]. The Shepherd Group at Birkbeck maintains powerful GPU servers running AMBER and has made good experiences with this setup investigating antibody complexes [89] which is why all simulations have been carried out with the AMBER simulation engine in this work.

1.3 Accelerated Molecular Dynamics enhances sampling

Nowadays, Classical Molecular Dynamics (CMD) that describes the dynamics of all atoms as a function of time can be used to investigate protein dynamics in the order of nanoseconds to milliseconds, depending on system size, computational resources and the overall time frame of the project [121]. Global conformational changes however can be in the order of microseconds and quickly get computationally intractable for larger systems [122, 123, 124, 125]. Accelerated molecular dynamics (AMD) is an enhanced sampling technique that is able to access meta-stable states that otherwise would take tremendous computing time with CMD [123]. It further has an advantage over Metadynamics, a related enhanced sampling technique, that no prior knowledge about the system at hand is needed to apply. Still, parameters of the method, namely boost potentials as well as thresholds must be chosen appropriately for expressive simulations [124]. A probing of a range of parameters is often advisable. Little a priori knowledge is also necessary for replica exchange molecular dynamics simulations but a proper setup is typically much harder to achieve and very powerful computing resources are needed to run such simulations.

1.3.1 A modified potential to overcome energy barriers

CMD simulations might be stuck in a local minima with energetic barriers too high to overcome in a reasonable time frame. The quasi non-ergodic nature of large biological molecules means there is a possibility that significant conformational changes and other effects may be missed [126]. A path in the potential energy landscape from one meta-stable state to another is typically a sequence of rarely-visited conformations [127]. The AMD approach implemented as part of AMBER adds a bias potential to the potential energy landscape (the potential energy of the system calculated by the chosen force field) so that simulations more quickly get out of low energy basins and are able to sample sequences of rare conformations that lead from one meta-stable conformation to another [128]. Low energy wells get filled up by a boost potential $\Delta V(r)$ if the potential $V(r)$ falls below a predefined

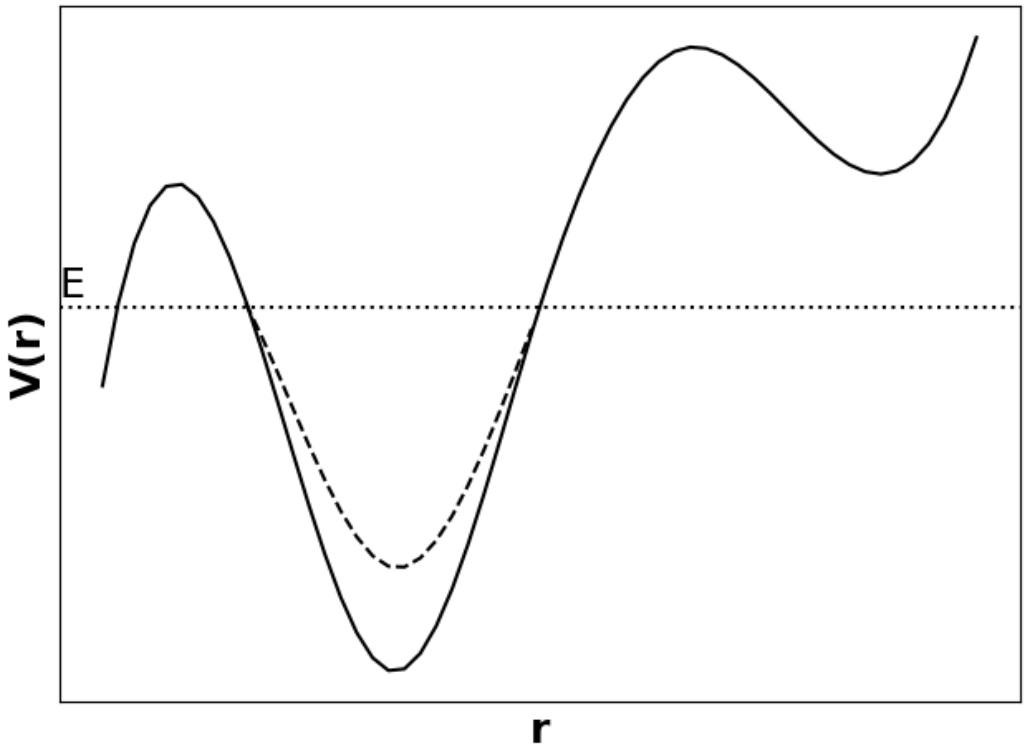


Figure 1.11: Exemplification of the normal potential as calculated by the force field (solid line) and boosted potential (dashed line). r is the system state and $V(r)$ the potential energy of that system state. AMD fills low energy wells in the potential energy surface by adding a boost potential.

threshold as defined by

$$V^*(r) = \begin{cases} V(r), & V(r) \geq E, \\ V(r) + \Delta V(r), & V(r) < E, \end{cases} \tag{1.8}$$

where $V(r)$ is the normal potential of a system state r. This is also illustrated in figure 1.11. Parameters to set for an AMD simulation include the threshold for the energy E and the boost factor α which together influence the bias potential $\Delta V(r)$. For the approach of Hamelberg *et al.* [128] used here the bias potential is given by

$$\Delta V = \frac{(E - V(r))^2}{\alpha + (E - V(r))}. \tag{1.9}$$

The AMD implementation in AMBER adds a boost to torsional energy with the

option for an additional boosting of potential energy. Since this work uses the latter approach, two boost potentials and two thresholds were calculated. In the course of this research a script was developed that uses an initial 2 ns CMD simulation to estimate appropriate boost potentials and thresholds as recommended by the relevant AMBER tutorial [129]. The calculation of the threshold for potential energy E_{thres_P} and dihedral energy threshold $E_{thres_{dihed}}$ as well as respective boost potentials α_P and α_{dihed} use the number of atoms in the system (n_{all} = 221528 for the solvated structure), the number of residues of the structure under investigation (n_{res} = 156 for the wild-type C2-domain), the average total potential energy (E_{pot}=-688815.6798) and average dihedral energy (E_{dihed}=7284.1652) which have been calculated from a short 1 ns simulation and is given in the following:

$E_{thres_P} = E_{pot} + 0.16\frac{kcal}{mol} * n_{all} = -688815.6798\frac{kcal}{mol} + 0.16\frac{kcal}{mol} * 221528 = -653371.1998\frac{kcal}{mol}$

$\alpha_P = 0.16\frac{kcal}{mol} * n_{all} = 0.16\frac{kcal}{mol} * 221528 = 35444.48$

$E_{thres_{dihed}} = E_{dihed} + (4\frac{kcal}{mol} * n_{res}) = 7284.1652\frac{kcal}{mol} + (4\frac{kcal}{mol} * 156) = 7908.1652\frac{kcal}{mol}$

$\alpha_{dihed} = \frac{1}{5} * (4\frac{kcal}{mol} * n_{res}) = 4\frac{kcal}{mol} * 156 = 624$

Sometimes it is advisable to increase the boosting of the system if progress is deemed slow. In such a case the Amber manual [113] suggests to add multiples of the boosting potential α_{dihed} to $E_{thres_{dihed}}$.

1.4 Umbrella sampling improves the understanding of transitions

Often, a protein or other biomolecular structure prepossess more than one stable conformation. These meta-stable states are found in basins of the potential energy landscape and typically have a high energetic barrier between them. With both CMD and AMD, energetic basins are sampled very well whereas it might take very long to get out of such a basin or, in other words, sample the transitional states that lead over a high energetic barrier (figure 1.12). It is impractical to develop a good understanding of the energy landscape between meta-stable states with the mentioned approaches. If the transition between two meta-stable states can be described by one or, perhaps two, degrees of freedom such as torsion angles, distances and/or RMSD values, Umbrella sampling (US) is a technique that can be used to sample transitional states along a path defined by these so called reaction coordinates. Using this approach, experimenters can not only determine the energetic landscape between two states but also the absolute difference in free energy of the conformations determined by the reaction coordinate and thereby the reaction coordinate's contribution to the free energy of the whole system. US was initially developed in the context of Monte-Carlo simulations by Torrie and Valleau but has proved useful for MD simulations as well [130, 131, 132, 133, 134].

1.4.1 Biased potentials lead the reaction coordinate along a predefined path

US is a biased molecular dynamics method that transitions a structure under investigation from one state into another by applying forces along one or multiple predefined reaction coordinates ξ. This transition is done gradually in a windowed fashion. The shape of windows along the path of transition are determined by the bias ω_i that ensures that the sampled states do not deviate to far from a reference state ξ_i^{ref} which is defined by the reaction coordinate. The bias is typically harmonic, pulling the reaction coordinate back to the reference value with increasing force as the two diverge (figure 1.13). With K as the strength of the bias, the bias

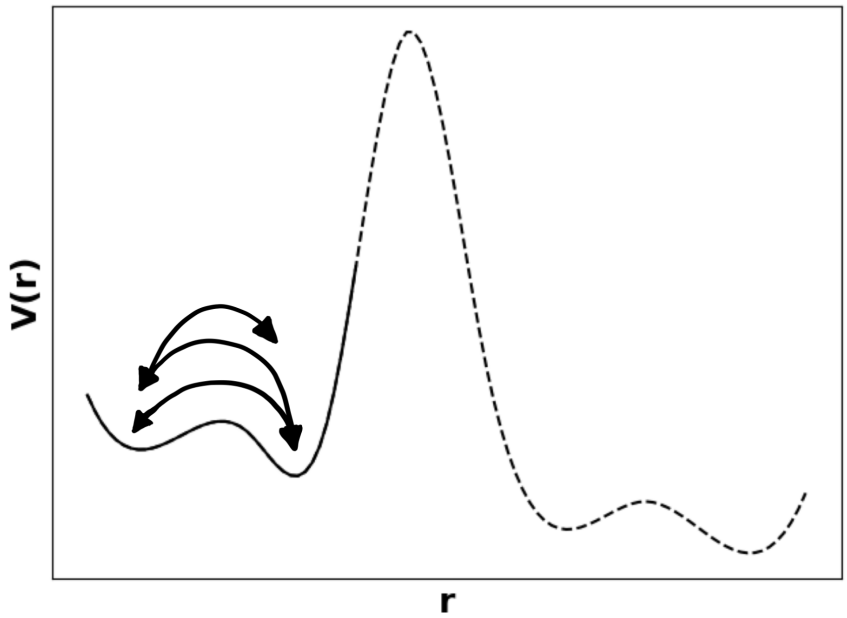


Figure 1.12: Boosting of the potential in Accelerated Molecular Dynamics: r is the system state and $V(r)$ the potential energy of that system state. Because of an energetic barrier the low energy basins on the right do not get sampled in reasonable time nor do the transitional states leading over the energy barrier.

for window i is calculated as:

$$\omega_i(\xi) = K/2(\xi - \xi_i^{ref})^2 \tag{1.10}$$

After the reaction coordinate made the structure transition from one state to another the probability distributions of the intermediate states get combined and unbiased by the Weighted Histogram Analysis Method (WHAM) [134] to reconstruct a cumulative probability distribution along the transition. This distribution can further be transformed into the free energy along the chosen reaction coordinate which is called the potential of mean force. An interesting measurement is the difference in free energy from an initial state to an end state. Because the representation of such initial and end states merely by conformations belonging to the singular state-defining reaction coordinate value would not comprise naturally occurring thermodynamics it is good practice to incorporate the probabilities p in the vicinity of the state up to a sensible amount n. The free energy of a state i by a set of probabilities

can be calculated as follows:

$$\begin{aligned} \Delta G &= G_i - G_j \\ G_i &= -k_b T ln(P_i^*) \\ P_i^* &= \frac{\sum_{l=i-n} p_l}{\sum_{k=i+n} p_k} \end{aligned} \tag{1.11}$$

Where k_b is the Boltzmann constant, T the temperature and P_i^* is the sum of probabilities p including those in the vicinity of state i (figure 1.14).

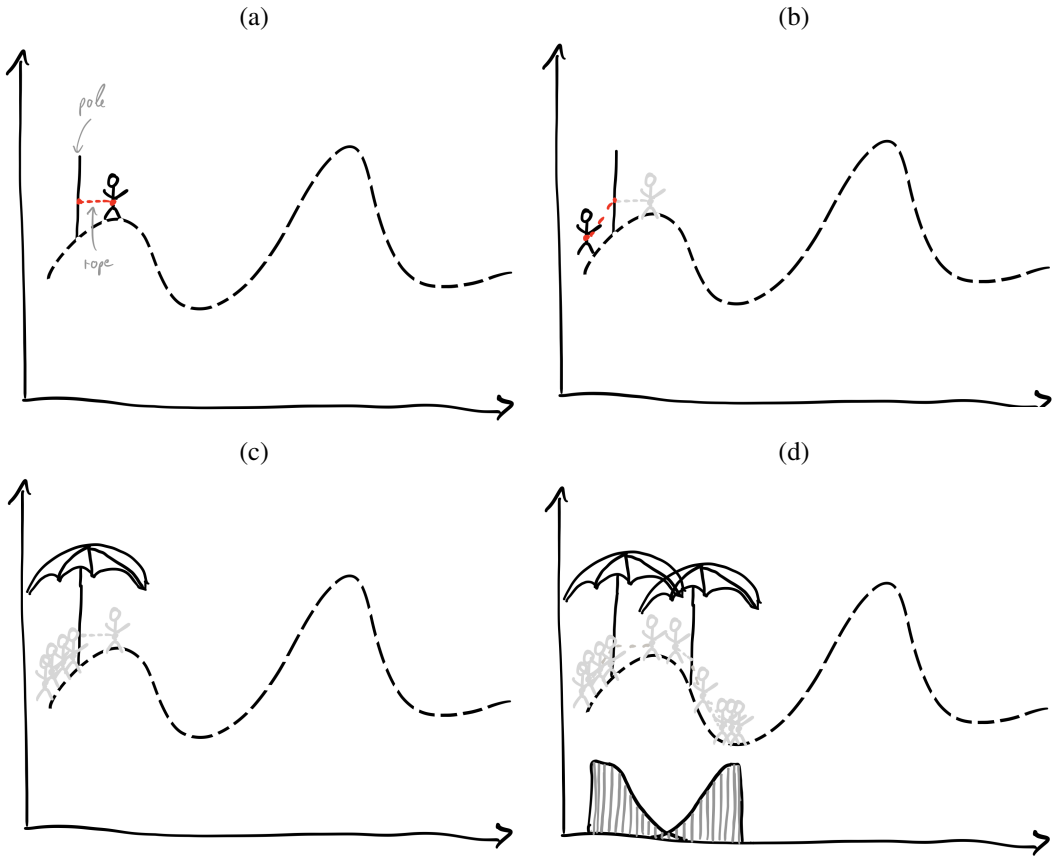


Figure 1.13: To reproduce an unknown potential of mean force (dashed curve) of a reaction coordinate the method of Umbrella Sampling could be imagined in what follows (the terms are taken from equation 1.10): (a) a pole (ξ_i^{ref}) is put in the ground and a 'stroller' (ξ) is restrained to it via a rubber band (ω_i) (b) as the stroller strolls about the rubber band pulls him or her back to not diverge to far from the pole. We further count how often positions on the x-axis have been visited by the stroller (c) After some time, all positions that are in range of the rubber band are visited or the stroller has visited all positions under an imaginary umbrella. Because the stroller does not like to go uphill too much most of the time is spent in lower regions (d) We repeat the above procedure with the pole positioned at specified gaps, with the requirement that umbrellas overlap. Now, by looking at our notes of positions (the histogram and probability curve drawn above the x-axis) we can comprehend how the landscape must have looked. Higher regions will be populated much less than lower regions. Umbrellas should have a considerable overlap so that even rarely-visited positions are sampled. Gaps between umbrellas lead to an inaccurate reconstruction of the potential of mean force. Besides the distance between poles the restraining force, determining the width of the umbrellas, could be chosen appropriately to sample these regions sufficiently.

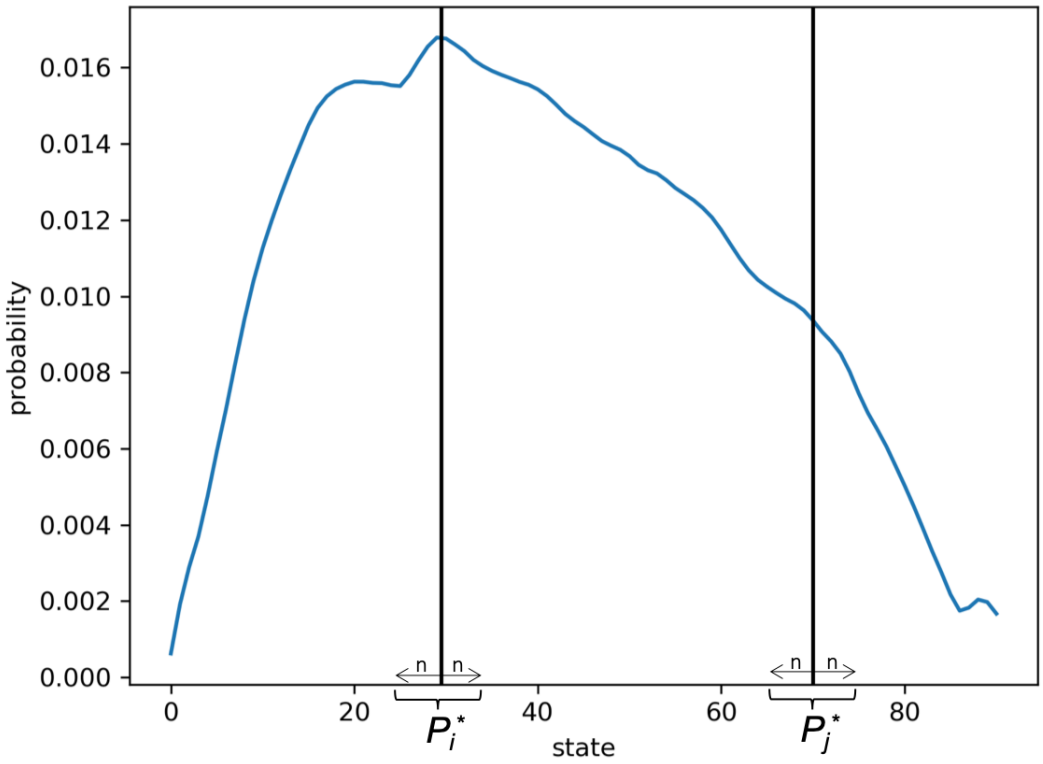


Figure 1.14: Free energy calculation from probability. To calculate the difference in free energy of states i and j it is good practice to include the probability of a number n of neighbouring states. These probabilities P_i^* and P_j^* can then be used in the calculation of free energy as in equation 1.11.

1.5 MM/GBSA: The compromise between accurate and rapid free energy calculations

Molecular Mechanics Generalized Born Surface Area (MM/GBSA) is a computational method that is used to calculate the free energy from molecular systems. It is popular in computational biology for estimating the strength of small ligand binding [135] but has been successfully applied to characterize larger complexes as well with latest applications in the course of SARS-Cov-2 related research [136, 137]. In benchmarks against more rigorous methods like free energy perturbation (FEP) and thermodynamic integration (TI), MM/GBSA has been shown to perform less accurately, but still comparatively well. An at least 8-fold reduction in computation time and the possibility to decompose binding free energy down to the residue level justifies its use - even more so with larger complexes that need a lot of sampling with more rigorous methods [138, 139, 140]. Molecular docking approaches have been found useful to quickly find binding poses but binding affinity scoring is not reliable [141] though new docking approaches combined with a mix of models has shown promising results in the recent GC4 challenge held in January 2020 [142]. Free energy calculations using the MM/GBSA approach is the middle way between empirical scoring and very computationally heavy but theoretically sound methods [114].

1.5.1 Change in free energy as an estimate of binding affinity

Free energy describes the state of a thermodynamic system as the amount of reversible work at a constant temperature and constant pressure and is given in equation 1.12:

$$G = H - TS \tag{1.12}$$

where H is the enthalpic and TS the temperature dependent entropic contribution. In chemistry it is common practice to compare the free energy of two system states to draw conclusions if a chemical reaction will happen spontaneously as is the case when the free energy of the initial state is higher than the free energy of the final state. This difference, the ΔG value, is calculated as the difference of enthalpy

minus the difference in entropy times temperature:

$$\Delta G = \Delta H - T\Delta S \tag{1.13}$$

One could also calculate change in free energy by subtracting the free energy of one state against the other.

$$\Delta G = G_{state1} - G_{state2} \tag{1.14}$$

In this work, the system comprises the molecular complex and the explicit solvent molecules. Varying conformations of the solute and rearrangement of water are effects that are naturally occurring in the thermodynamic equilibrium [143]. To account for this dynamic behaviour, free energy is calculated over a range of system states. Since a quantification of the binding free energy between the parts of the complex is desired, one could naively subtract the free energy of the complex from that of the separated binding partners as illustrated in figure 1.15. However, the free energy of systems containing explicit water is largely determined by the contribution of solvent-solvent interactions that fluctuate by an order of magnitude larger than the binding free energy. Lengthy simulations of both system configurations (in complex and with separated partners) would be needed to get convergence in the presence of solvent-solvent interactions. Neglecting the solvent completely is far from optimal, since hydrophobic residues that are shielded from water in the binding site typically contribute a greater portion of binding free energy. The MM/GBSA method replaces the explicit solvent by an implicit solvent model, removing free energy calculations of solvent-solvent interactions. Polar solvation free energy is then approximated with the Generalized Born approach whereas interactions between binding partners are determined by molecular mechanics (MM) 1.15.

$$\begin{aligned} \Delta G_{bind,solv} &= G_{complex} - G_{separated} \\ G &= G_{MM} + G_{solvation} - TS \\ G_{MM} &= E_{bnd} + E_{el} + E_{vdW} \\ G_{solvation} &= G_{pol} + G_{np} \end{aligned} \tag{1.15}$$

G_{MM} is calculated for the solute in vacuum consisting of E_{bnd} which includes energies from atoms linked by a covalent bond and E_{el},E_{vdW} that capture long range interactions. These energies are described in more detail in chapter 1.2.1. Interactions with the solvent are split into G_{pol} which is estimated using the Generalized Born approach and a non-polar contribution G_{np} approximated by the solvent accessible surface area. Both are outlined in more detail in subsection 1.2.2.2.

The entropy S in equation 1.15 can be split into configurational entropy and hydrophobic entropy. The methods to calculate configurational entropy, quasi-harmonic and normal mode approximations that are part of the AMBER package, were found to introduce additional statistical errors, rendering free energy calculations less expressive [113].

In detail, entropy calculations employing quasi-harmonic approximation have been shown to have severe convergence problems and conceptually expect only one energy minima which is a an approximation too crude for flexible structures such as proteins [144, 145, 114]. Normal mode approximation (NMA) is a very computationally expensive technique which is why typically a truncated NMA approach is used that restricts calculations to an area of interest [146]. It has been shown to improve correlation to experimental results in some cases but also worsening free energy calculations in other cases [147]. Because of the tremendous computation times of even truncated NMA it cannot be easily ascertained if energy predictions benefit from the inclusion of a configurational entropy term.

An assumption often made is therefore that the configurational entropy is not significantly different between holo (bound) and apo (unbound/solitary) state nor is influenced by the amino acid change that is associated with point substitutions. The omission of configurational entropy calculations has become standard practice in free energy calculations [113].

However, research on configurational entropy using the computational method 'Mining Minima algorithm' [148] that is able to split free energy into enthalpic and entropic contributions shed light on the importance of configurational entropy and how it is influenced by dynamics, namely rotational and translational degrees of

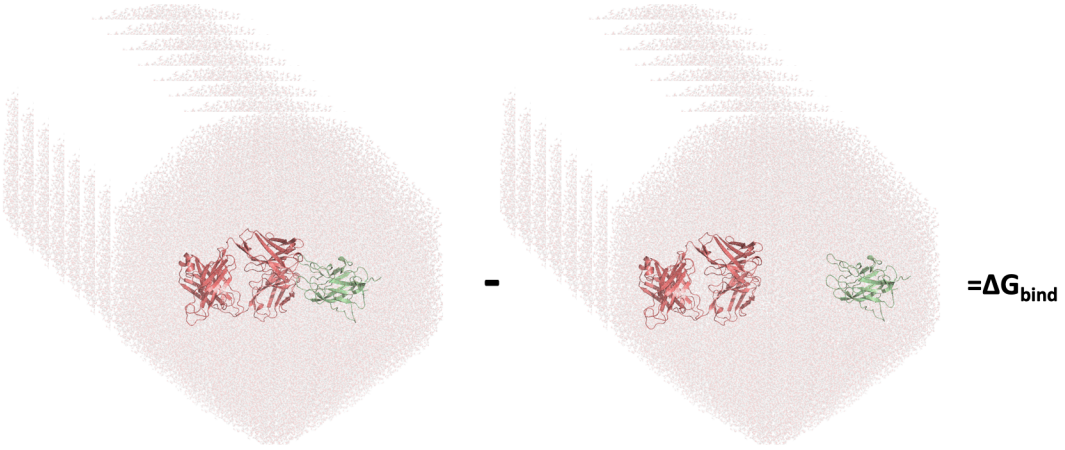


Figure 1.15: Naive calculation of binding free energy by subtracting free energies of a set of binding partner trajectories in complex and separated into ligand and receptor including solvent-solvent interactions

freedom and overall rigidity of a ligand [149]. The results of the study of Chang et al. suggest that binding affinity is as much influenced by entropy as by electrostatic interactions and hydrophobicity. Their conclusion is therefore that the ranking capabilities of computational techniques omitting configurational entropy will by design have a limited accuracy.

As discussed in subsection 1.2.2.2 entropy of water molecules is only part of empirically determined parameters and thereby cannot be evaluated directly. Overall, the lack of proper entropy estimations introduces an offset between predicted ΔG and experimentally determined free energies. Since, in many experiments and also in this work the goodness of fit between experimental and predicted binding free energies relies on $\Delta\Delta G$ values (differences in free energy relative to the wild-type), direct comparison of predicted ΔG values to experimental ΔG values is not required.

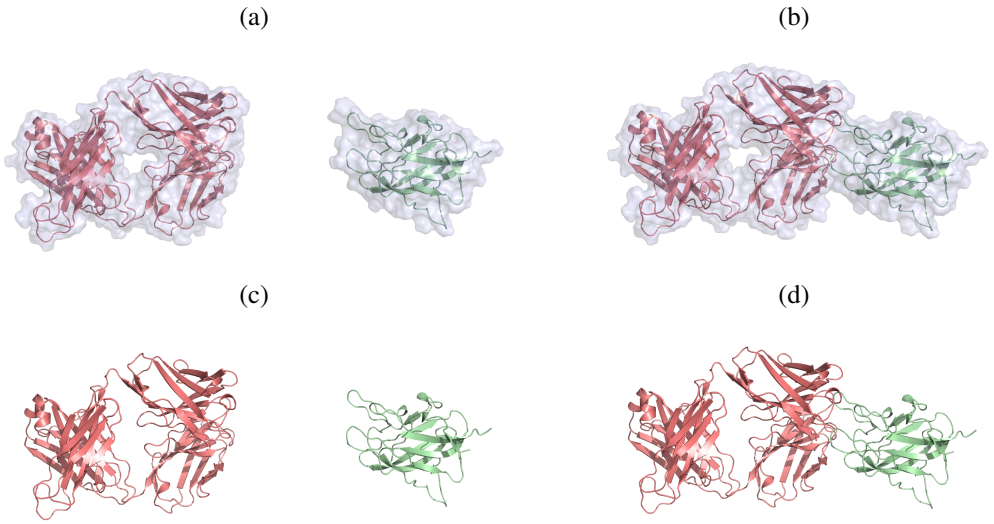


Figure 1.16: Calculation of $G_{bind,solv}$: (a) implicit solvation of separated binding partners, (b) implicit solvation of complex, (c) separated binding partners in vacuum, (d) complex in vacuum. Polar and non-polar solvation free energy ($\Delta G_{pol}, \Delta G_{np}$) are approximated using the GB approach (figures (a),(b)). MM terms are calculated in vacuum (figures (c),(d)). The cumulative energy of systems (a), (c) is then subtracted from (b),(d) to get the energy upon binding.

1.6 Existing structural analysis techniques reduce dimensionality insufficiently

The sheer amount of spatio-temporal data produced by MD simulations poses a challenge to researchers trying to pin down the most interesting phenomena of the system under investigation. This problem gets even harder when effects across simulations that use slightly altered structures (e.g. point mutations) should be considered. Often, structural changes can be described as RMSD/RMSF values. Such an approach is legitimate in situation where the researcher knows in advance what to focus on or for binding sites that contain only a few residues. Probing a wider range of atomic positions across different simulations exponentially grows the possibilities for RMSD/RMSF calculations resulting in an insufficient reduction of dimensionality to spot meaningful phenomena.

Another method popular in the field is principal component analysis. It represents the motion of an atomic system by principal components that range from large scale conformational changes to local motion and thermodynamic noise. Since motion differs from simulation to simulation the calculated components are differing as well and cannot be compared computationally. An investigation of components by eye is cumbersome and error-prone across a set of simulations.

1.6.1 A decomposition of MM/GBSA energies helps to elucidate binding patterns

One advantage of the MM/GBSA method is that binding free energy can be decomposed to a pairwise-contribution of residues in the epitope. Further, by comparing decomposed binding free energies of mutants to those of the wild-type, weaker as well as stronger bonds can be investigated across simulated frames and visualized as in figure 1.17 and figure 1.18. It is thereby possible to develop an intuition about underlying binding mechanisms and differences introduced by mutations, which has proven insightful in a study of Shepherd and co-workers involving stalk binding antibodies in the context of influenza virus [89]. By decomposing binding free energies they were able to determine the importance of residues in the epitope of three

antibodies and illustrate the difference in interaction energy for selected mutants with the aim to inspire the development of stalk binding antibodies.

In the context of the apoptosis inhibitor survivin that is overexpressed in solid tumors, Sarvagalla and co-workers employed a pairwise decomposition of frames of a simulation of the wild-type using MM/PBSA to identify residues that are involved in biologically relevant interactions which they call 'hot spots' . They subsequently substituted hot spot residues to alanine, ran MD simulations, and calculated binding free energies. By that, they were able to define a pharmacophore model that further informed a virtual screening that identified the HIV protease inhibitor indinavir as a potential survivin inhibitor [150].

It is however challenging to undertake an analysis of a set of simulations including amino-acid substitutions using a pairwise decomposition since random fluctuations can often not be distinguished from changes introduced by a substitution. The usefulness of interaction plots is limited to the investigation of a few simulations, residues or phenomena.

Something to consider when using a pairwise decomposition of free energy is that for a rigorous decomposition of total free energy into partial free energy contributions, interaction energies must be considered in isolation which would make it necessary to separate the molecule into independent parts. However, such a separation would result in a new system where populated states or phase space in general differs from the unseparated/original system since interactions are typically nonadditive. Free energy calculated for the new system would not be precise. Differences in free energy calculated by subtracting the free energy of the new system from the free energy of the original system therefore do not reflect the free energy contribution of an interaction. Calculated differences in free energy using such an approach would depend on the correlation of the interaction of interest with all other interactions. A disentanglement of interactions into independent parts remains a challenging task that needs considerable simulations lengths [151]. Non-the-less an energy decomposition of enthalpic contributions can be a helpful addition to studies aiming to predict the impact of mutations or alterations of interactions. Yet, one has

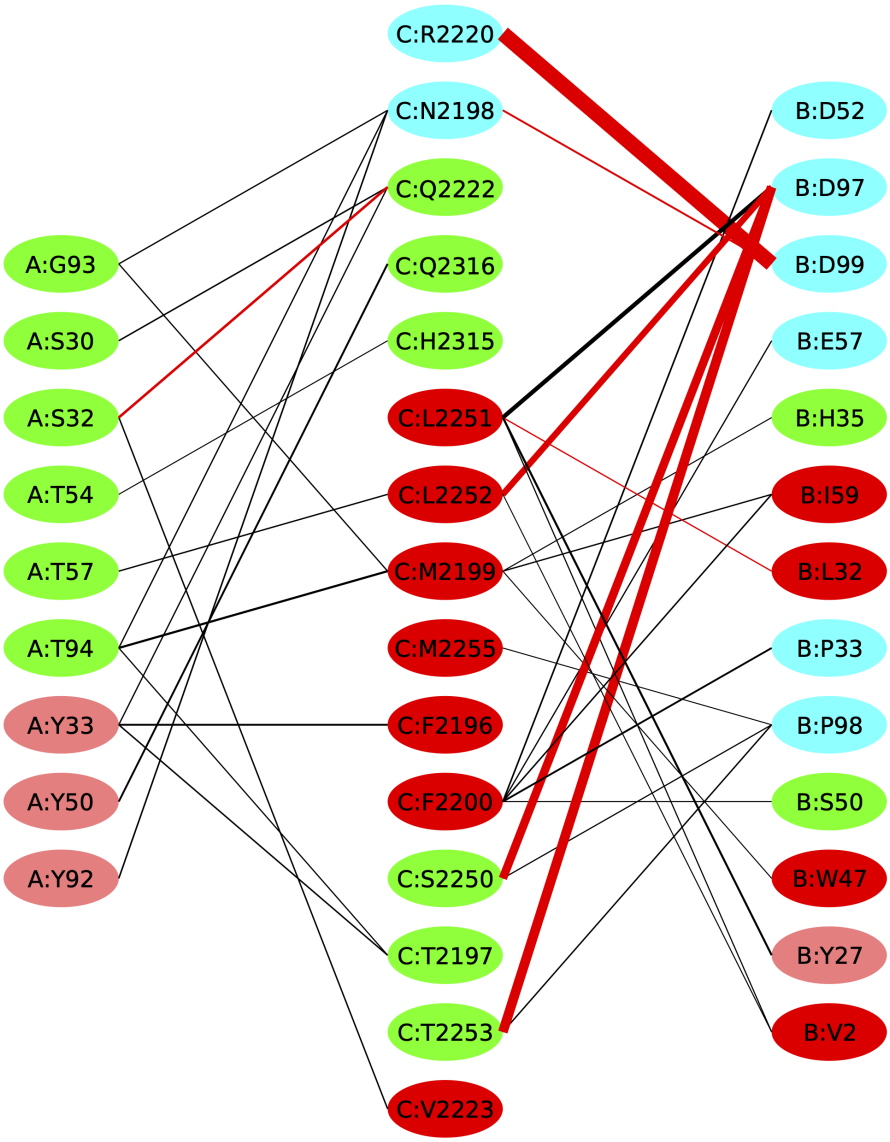


Figure 1.17: Visualizing interaction energies of residues: Identifiers given by chain letter (e.g. 'A','B','C') followed by the amino acid one letter code and residue number after the colon; hydrogen bonds (the hydrogen bond has to be present in at least 150 out of the last 200 frames of the equilibration) coloured in red, remaining interactions in black; The thickness of lines is determined by the represented energy (the thicker the line the more energy); residue colouring based on the hydrophobicity at pH 7: pink and red reflect slight and high hydrophobicity respectively, green neutral, cyan hydrophilic; energy values below 1 kcal/mol are not displayed.

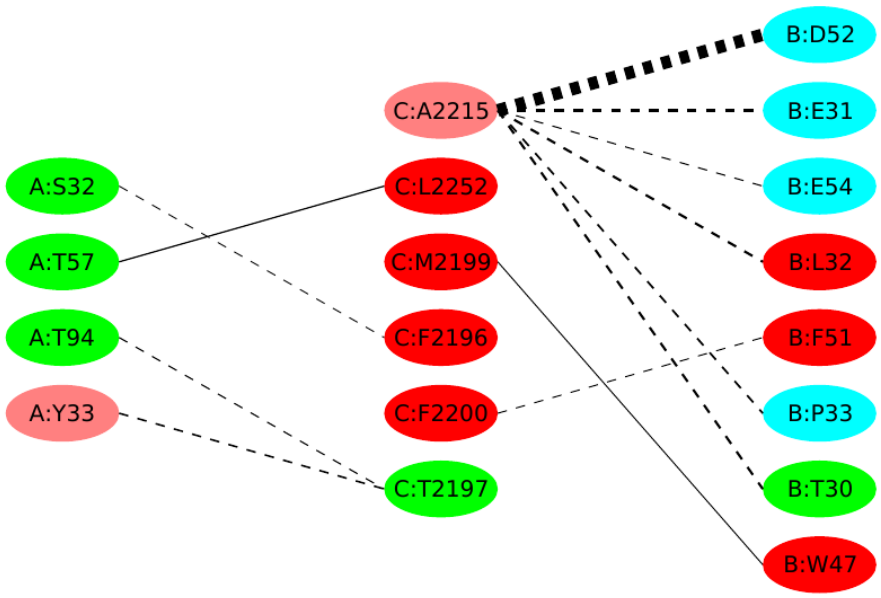


Figure 1.18: Relative interaction energies to wild-type: Dotted lines indicate a loss in binding energy, solid lines a gain; colouring, line thickness and threshold as in figure 1.17

to be cautious to not over interpret such enthalpic decompositions since the actual effect of a structural alteration on free energy is ultimately a mix of thermodynamic properties where the role of enthalpic contributions is hard to come by [152].

Chapter 2

In-depth analysis of the BO2C11 FVIII C2-domain binding site

So far, the human antibody BO2C11 is the only human anti-Factor VIII antibody studied in detail. Pratt and co-workers used SPR to measure the binding kinetics of 43 non-silent point mutations introduced to the epitope of the BO2C11 FVIII C2-domain complex (from now on referred to as the holo FVIII C2-domain). In this work, 16 of the 42 amino-acid replacements to alanine and one replacement to methionine have been characterized computationally. This subset comprised all residues that were described previously by Pratt and co-workers to mediate binding and residues that stabilize binding as well as such that did not show a significant impact in experiments. The latter two sets of amino-acid replacements were chosen so that the full range of experimental binding affinity measurements was represented in simulations. Since the majority of experimental affinity measurements differ less than 1 kcal/mol from wild-type it is sensible to restrict calculations to a representative subset of replacements since such small changes in affinity would not be picked up by the computational approach chosen here. Different equilibration times were evaluated and the number of frames for binding free energy calculations was chosen to reduce the statistical error to under 1 kcal/mol.

Even though entropy has been shown to be a major contributor to binding affinity in experiments (referring to high $T\Delta S_A$-values in table 1.3), the computational approach adopted here only considers entropic contributions stemming from the

calculation of solvation free energy by the Generalized Born model (as discussed in detail in section 1.5), where entropy is represented in the empirically derived constants for the estimation of non-polar as well as polar Gibbs free energy. The importance of configurational entropy, which is the entropy neglected in this work, has been investigated computationally by Sun et al. [147]. In their study, the calculation of entropy using the method 'interaction entropy' has been shown to improve the correlation to experimentally determined binding free energies with the force field FF14SB with a dielectric constant of $\varepsilon = 1$, which were also the choices for MM/GBSA calculations in this work. However, since relative binding free energy was estimated comparatively well with a Pearson correlation of $r_p = 0.62$ in the initial MD simulation carried out in the course of this work, I refrained from entropy calculations that are labour-intensive to implement and/or take tremendous time to converge.

2.1 MD and MM/GBSA protocols

2.1.1 Preparation of crystal-structures for simulation

The holo FVIII C2-domain crystal-structure (where the BO2C11 fab fragment is bound to the recombinant FVIII C2-domain; PDB 1iqd) [61] and the apo FVIII C2-domain crystal-structure [60] (PDB 1d7p) were downloaded from the Protein Data Bank [86]. The 2 Å crystallographic structure of the holo FVIII C2-domain contains three segments of three, six and nine residues that are unresolved in the structure of the antibody fab fragment at locations remote from the binding interface. MODELLER v9.17 [153] was used to model these missing loops and to introduce amino acid substitutions.

The apo crystal-structure of the FVIII C2-domain has a resolution of 1.5 Å with no missing segments but with crystal contacts as outlined in figure 2.1. Crystal contacts are artefacts that occur due to the process of crystallization and potentially influence the structure of individual crystals [154].

Structures were protonated with MolProbity 3.3.160602 [155]. Topologies for simulations were generated by tleap from AmberTools 18 [88] with the FF14SB

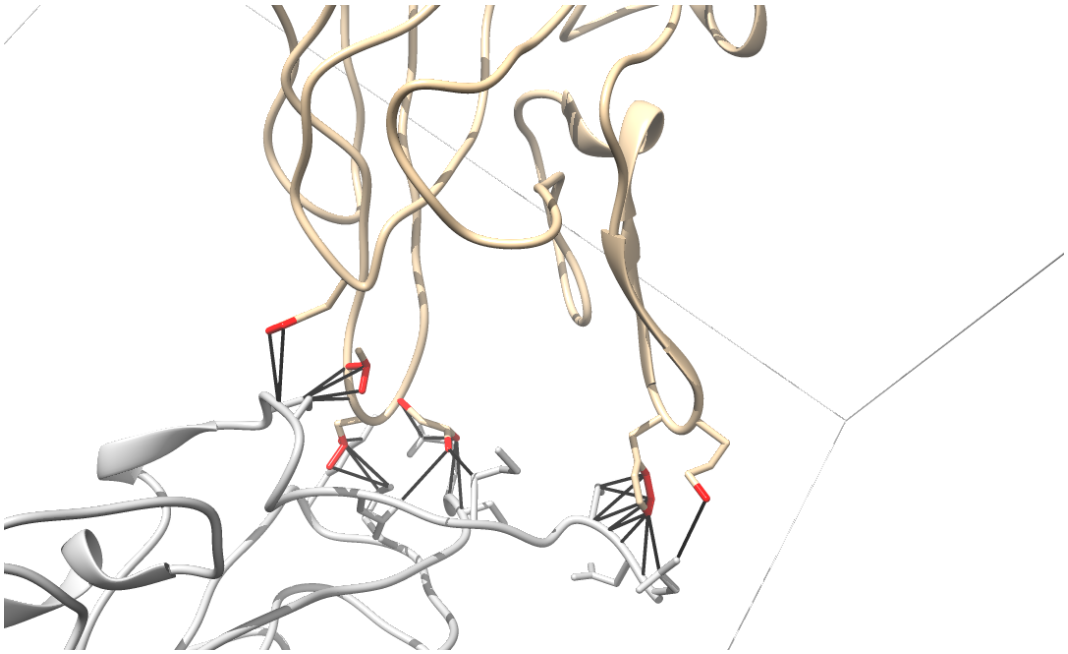


Figure 2.1: Crystal contacts of the apo C2-domain structure (PDB 1d7p); epitope region of the C2-domain (light-brown) including two β-hairpins that contain numerous contacts (black lines) to a neighbouring unit (grey).

protein force field and mbondi2 radii set. Structures were solvated in a TIP3P octahedral water box with charge neutralized by Cl- ions. The distance between the box boundary and a given structure was set to a minimum of 20 Å. Short range van der Waals and electrostatic interactions were restricted to a distance of 8 Å to reduce computational time. For van der Waals energies, that are modelled using a Lennard-Jones potential, attractive forces are nearly zero after the cut-off distance chosen here. Long-range electrostatic interactions above this threshold were calculated employing particle mesh Ewald summation (PME) [156].

2.1.2 An equilibration time of 40 ns was ascertained employing statistical methods

Residence times for protein complexes in this work are in the order of hours (see table 1 in [44]). Even with supercomputers, simulation times of hours are insurmountable in reasonable real-time and it is necessary to find simulation times that are computed in the time frame of the project and give the experimenter enough confidence that an equilibrated state has been reached. In practice, the real time

duration of MD simulations typically ranges from a few hours to several days or month [157].

Another consideration concerning equilibration is that structures used in this work have been produced by X-ray crystallography [61, 60]. This method makes use of the scattering of X-rays that give information about electron densities and thereby atom positions [141]. By recording the diffraction pattern of the crystal at different orientations, it is then possible to define an electron density map and further draw conclusions about the atom types and relative positions. Hence atomic positions are averaged and do not necessarily reflect positional fluctuations that occur at thermodynamic equilibrium [158]. Further, since a crystal comprises many molecules of the same type, one has to consider that these contact each other and form so-called crystal contacts (figure 2.1). These can alter the electron density and therefore influence the derived conformation of the structure.

A preliminary equilibration stage has been proven to produce more accurate binding free energy values in a benchmark comprising 43 complexes compared to solely minimising energy [140].

For these reasons, it is common practice to equilibrate a structure, i.e. bring it into thermodynamic equilibrium, before running succeeding simulations upon which conclusions are based [113]. It should be noted though, that equilibrium can only be achieved for a limited amount of degrees of freedom and not for the whole structure the size of a protein domain [145]. The AMBER manual recommends further a slow increase of temperature and a restraining of force field energies before the actual equilibration. A conservative and incremental increase of temperature should prevent high velocities and abnormally high energies that could 'blow up' the structure, which is often the case if the temperature is increased from 0 K to 300 K in a single step or, in general, if temperature increments are too big [113]. Preparatory steps in this work follow the protocol implemented by Shepherd and co-workers [89]:

1. Initially there is a minimization of the system energy by 40 steps of steepest descent prior to 40 steps of conjugate gradients.

2. This is followed by a 100 ps relaxation production where backbone atoms are restrained by a force equivalent to 4.0 kcal/mol and an increase in temperature from 0 K to 50 K. Pressure is not restrained during this first relaxation.

3. The second relaxation of 2 ns raises the temperature further to match the one used in all succeeding simulations. Backbone atoms are restrained by a force equivalent to 1.0 kcal/mol.

4. To conclude the adaption of the structure to the force field another 1 ns simulation is run with no restraints on the backbone.

To date, the most widely adopted method to quantify sufficient equilibration time (besides the analysis of global parameters outlined in section 1.2.3) is by the means of a Root Mean Square Deviation (RMSD) value. Since this quantity is only a single number, counteracting motion might not be spotted, e.g. parts of the structure might stabilize while others destabilize which would cause the RMSD value to stagnate. This might convey the impression that the structure is equilibrated [159]. The equilibration time in this work has been chosen based on the similarity of probability distributions of ϕ, ψ backbone angles of residues in the binding site in a prolonged simulation of the holo FVIII C2-domain structure. The similarity between two probability distributions P, Q was quantified using the Jensen-Shannon divergence distance metric which is given in the following equation:

$$\begin{aligned} JSD(P||Q) &= \sqrt{\frac{D(P||M) + D(Q||M)}{2}} \\ D(P||Q) &= \sum_{x \in \chi} P(x) log\Big(\frac{P(x)}{Q(x)}\Big) \end{aligned} \tag{2.1}$$

where M is the point-wise mean ($M = \frac{P+Q}{2}$) and D is the Kullback-Leibler divergence. The Jensen-Shannon divergence of ϕ, ψ backbone angle probability distributions after different equilibration times (1, 5, 10, 20, 40 ns where each nanosecond contains 100 frames) has been calculated against the probability distribution of a 1 μs run in this way (figure 2.2). If the two probabilities of backbone angles are

fairly similar, it means that the structure does not explore new backbone conformations after the simulation time of the shorter simulation up to the simulation time of the longer one and the divergence approaches zero. This means that side chain and backbone angles in the shorter simulation populate virtually the same values as in the longer simulation, which indicates that during the time span from after the shorter simulation up to the end of the longer simulation, no vital new angles are explored. By this logic the shorter simulation time is sufficient to reproduce the longer simulation time, which proposed that the binding site is just as well equilibrated after 40 ns as after 1 μs. However, the approach described here might miss some bigger conformational changes. The orientation of a prolonged structure, such as a β-hairpin might change disproportionately high in relation to backbone angles at the base of the structure, which might act as a lever, and would not be represented in the calculated Jensen-Shannon divergence of angles. It might be useful to additionally investigate such effects using RMSD values. Another improvement of the method could be to calculate the Jensen-Shannon divergence in a window like fashion. If a simulation is well equilibrated, a comparison of e.g. two windows, containing the first and second half of the frames to a prolonged simulation run should come to the same conclusion. By that, a better understanding of trends and equilibration times in general could be developed.

2.1.3 Fluctuations of MM/GBSA values dropped below 1 kcal/mol using 150 decorrelated frames

To remove bias through autocorrelation between frames, a set of decorrelated frames for energy calculations was extracted from 30 simulations of 1 ns each. These short simulations were independent from each other, meaning each used the position coordinates which were the ones from the last frame of the equilibration but assigned new randomized initial velocities to the system as recommended by Genheden et al. [160]. Further, only 5 frames in 100 ps intervals from the second half of these runs were used in the MM/GBSA binding free energy calculations in order to give the complex sufficient time to deviate far enough from the input state resulting in a total of 150 frames. The sampling frequency of 100 ps was based on

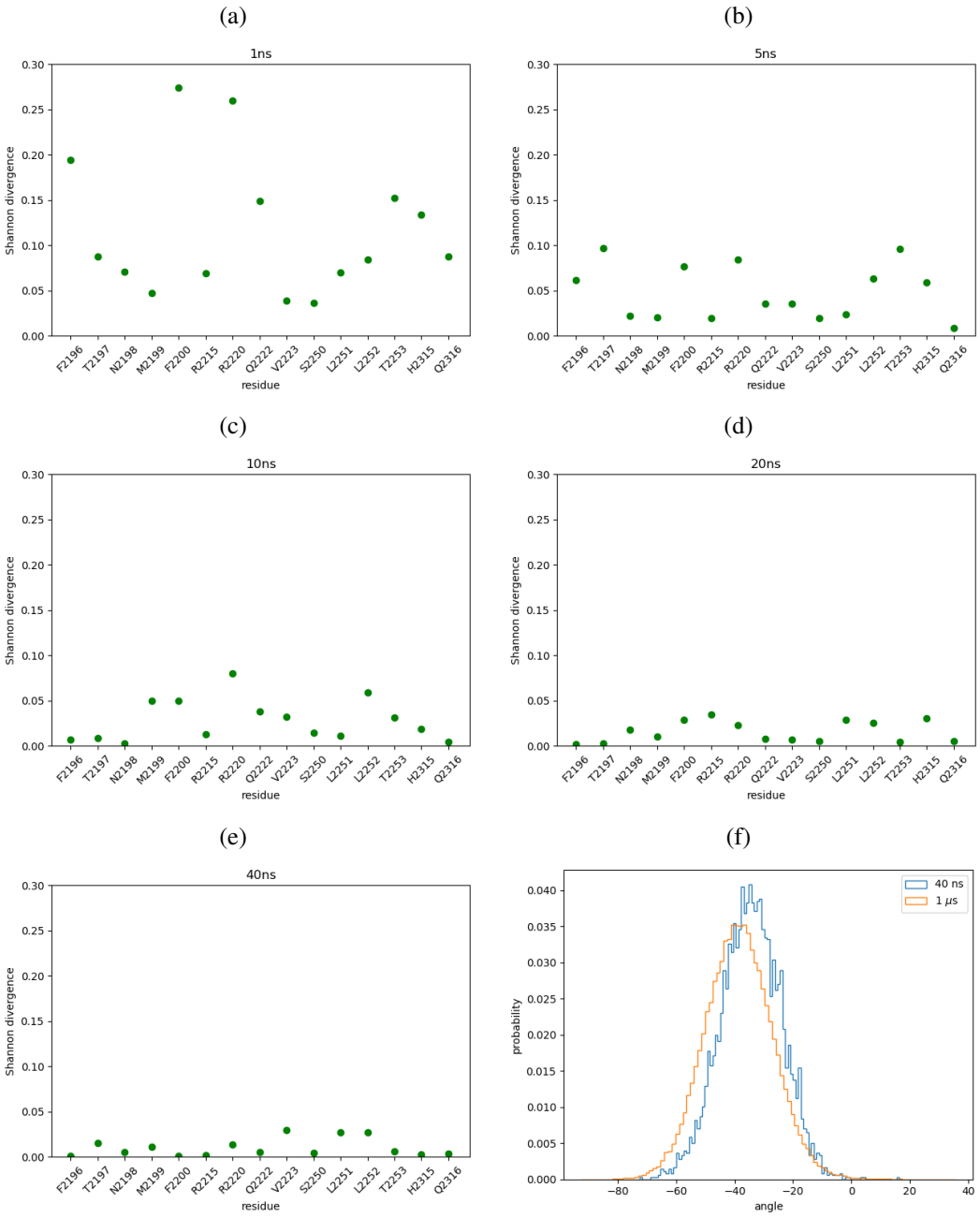


Figure 2.2: (a)-(e) Jensen-Shannon divergences of the probability distributions of binding site residue backbone angles. The reference probability distribution is a 1 µs simulation; (e) It showed that after 40 ns probability distributions are fairly similar; (f) Leucine 2252 is the least converged residue after 40 ns, yet distributions do not differ significantly.

the time taken for the time-dependent autocorrelation function of calculated ΔG values to reach 0. This protocol has been shown to produce good results as in Shepherd and co-workers [89].

The impact of the dielectric constant ε on MM/GBSA binding free energy calculations has been extensively studied in the papers of Hou et al. [140, 161]. In their work dating back to 2011, they propose a method to deduce most promising dielectric constants, in terms of high correlation coefficients to experimental results, for protein-ligand complexes. This method is based on the solvent accessible surface area of polar atoms in the binding site and suggests a solute dielectric constant of $\varepsilon = 2$ for the complex structure used in this work. However, in their more recent study, a dielectric constant of the solute of $\varepsilon = 1$ did successfully identify the binding pose that was closest to the crystal-structure out of a set of generated decoy poses. An increase of the dielectric constant worsened this ranking of binding poses. Therefore, $\varepsilon = 1$ was used throughout this work.

SHAKE [94] was used in all steps except for the minimization. Atom coordinates escaping the simulation box were set to reappear on the opposite side of the box ($iwrap = 1$). The remaining parameters were set to the AMBER 18 defaults. A Monte Carlo barostat was used throughout all simulations to keep pressure constant. The Monte Carlo barostat was used over a Berendsen barostat because of its recommendation by AMBER [88]. The volume was kept constant in all but the first relaxation run. A Langevin thermostat was used throughout the simulation with a collision frequency of 1 $ps^{-}1$. This value was set following a recommendation by the developers of AMBER (Re: [AMBER] NVT Vs NPT from Ross Walker on 2011-03-17 (AMBER Archive Mar 2011), n.d.). Simulations were carried out with the GPU-based simulation engine *pmemd.cuda* [162] except for the minimization step which used the CPU-based sander [113] as recommended because of the double precision floating point numbers of the CPU-based implementation. On a commodity hardware server, using Nvidia GTX 1080 GPUs, the simulation of one structure lasted about 4 days when run on a single GPU and resulted in a total of 73.1 ns of simulation time consisting of 3.1ns relaxation, 40 ns equilibration and

30 decorrelated productions of 1 ns each. With 17 substitutions plus the wild-type structure, the total runtime for a set of simulations was therefore around 72 days, which was further reduced by the use of multiple GPUs in parallel.

2.1.4 Investigating reproducibility by using different setups and repetition

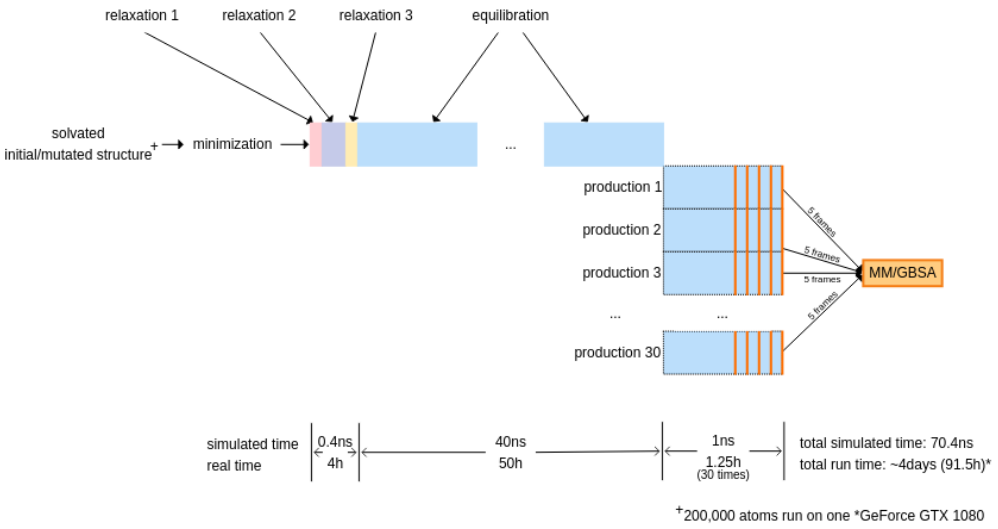
Using the setup outlined in the last section, 5 sets of simulations were run differing in temperature, initial structure and/or equilibration time.

- set CS25: conformation of the crystal-structure simulated at 25°C
- set CS37: conformation of the crystal-structure simulated at 37°C
- set CS25': conformation of the crystal-structure pre-equilibrated for 40 ns in a simulation at 25°C before amino acid substitutions were introduced
- set WT25': conformation of the wild-type pre-equilibrated for 40 ns in a simulation at 25°C before amino acid substitutions were introduced
- set WT37: conformation of the wild-type simulated at 37°C
- set WT25*: conformation of the wild-type with parameters $mbondi = 3$ and $igb = 8$ instead of $mbondi = 2$ and $igb = 2$ simulated at 25°C

where 'wild-type' refers to a reversion of the mutation S2296C originally introduced by crystallographers, to reconstruct the original FVIII C2-domain wild-type sequence.

Set CS25' and WT25' use the last frame of a 40 ns equilibration run at 25°Cas illustrated in figure 2.3 and therefore simulated a time span of 40 ns + 73.1 ns = 113.1 ns in total. Each set contained the 17 substitutions inspired by Pratt and co-workers and further M2199I and F2200L from the functional study carried out by Barrow and co-workers. Selected substitutions can be split into two groups. Group one containing R2220A/Q, R2215A, N2198A, M2199A, F2186A, S2250A, Q2316A and F2200A/L are substitutions that decrease binding affinity strongly (coloured blue in figure 2.4). Group two containing L2251A, L2252A, T2253A,

(a)



(b)

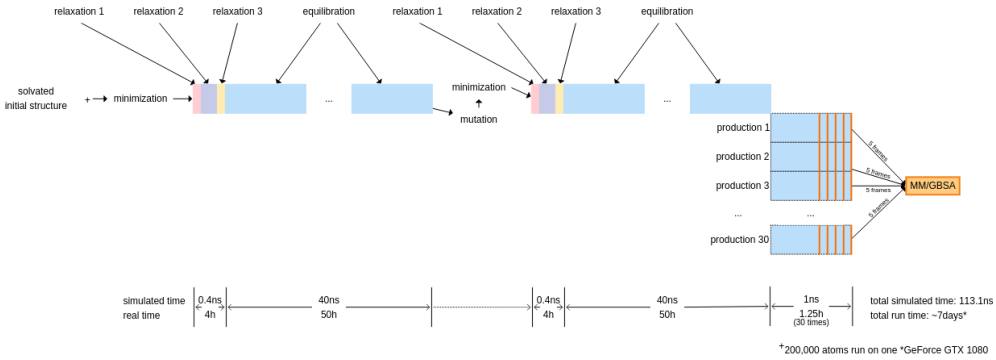


Figure 2.3: (a) Protocol for sets CS25, CS37, WT37 and WT25*: The conformation of the crystal-structure is used directly to introduce substitutions; (b) protocol for set CS25' and WT25': The difference to the other protocol is that an initial equilibration of 40 ns was granted so that the structure has more time to adjust itself to the new condition in explicit solvent. This provision has only been done once for the original starting structure (the conformation of the crystal-structure in set CS25' and the reverted wild-type in set WT25'). All following substitutions were introduced to the final conformation of this pre-equilibration; Orange lines in the production simulations indicate that chosen frames for MM/GBSA calculations are extracted from the second half and at intervals to allow structures to deviate from their initial conformation so that a range of conformations can be captured.

T2197A, H2315A, Q2222A, V2223M and M2199I are substitutions that have no or a slightly stabilizing effect on binding. Of special interest are the substitutions M2199I and M2199A that were shown to somewhat increase binding affinity [62] or reduce binding free energy [44] respectively.

The script MMPBSA.py was configured to carry out energy calculations based on implicit solvent simulations using the one trajectory approach with AMBERs Generalized Born model 2 (igb=2). This parameter resembles the model known as GB^{OBC1}, which is published as model I in the work of Onufriev et al. [108]. In the context of a study, which amongst others investigated the impact of MM/GBSA models on binding free energy calculations, this model has been shown to be superior in terms of correlations to experimental values over other choices [161]. According to the AMBER manual, the parameter for radii sets was set to 2 (mbondi=2) for this GB model. The unsolvated receptor and ligand topologies were created using ante-MMPBSA.py [113].

150 frames at 100 ps intervals were used to calculate binding free energy and pairwise contributions of binding site residues not more than 3.9 Å apart. The value of 3.9 Å is the default cut-off distance of the widely-cited programs Ligplot and Ligplot+, which are used to identify hydrogen bonds and hydrophobic contacts between proteins [163, 164]. In the analysis employed here, residue pairs are included which are within this cut-off in any snapshot, meaning that the list of interacting residues will be more permissive than that generated from a single structure.

The model accounts for surface tension in the non-polar contribution of solvation by 0.005 kcal/mol per 1 $Å^2$ solvent accessible surface area. To mimic conditions in the mammalian body, the salt concentration was set to 0.2 M whilst other parameters followed the AMBER 18 defaults. The MMPBSA.py calculations were carried out using the CPU-based code 'sander' from AMBER 18. Running on 15 cores of a 20-core 2.5 GHz Intel Xeon E5-based server calculations finished within minutes.

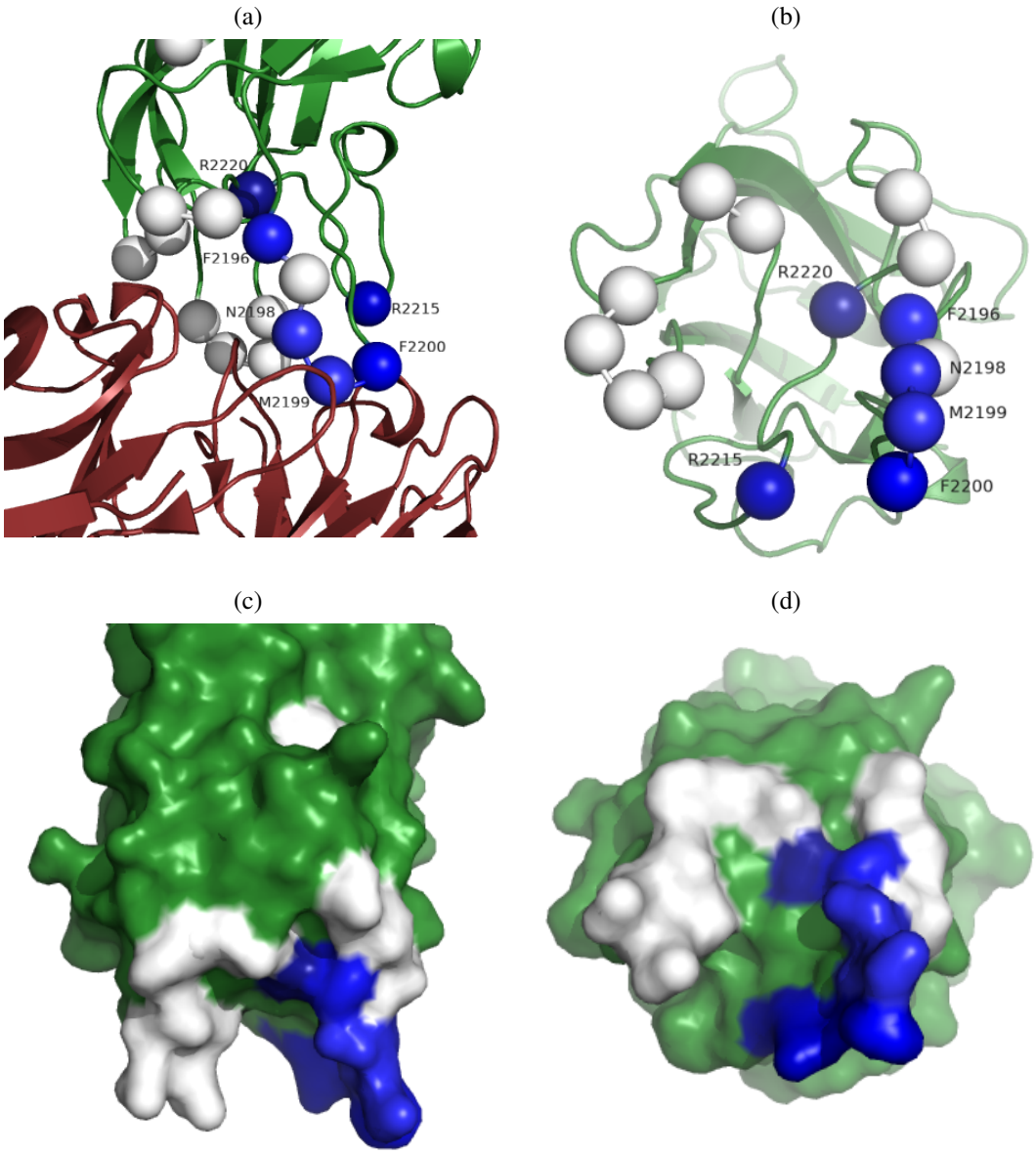


Figure 2.4: Important residues in the binding site of BO2C11 with the FVIII C2-domain; (a) spheres indicate amino-acid replacements evaluated experimentally and computationally, residues coloured blue were shown to have a significant impact on binding affinity according to experiments [44]; (b) and (c) view from the perspective of the antigen recognizing site; (c) and (d) surface plots highlighting regions of point mutations in white that have no or little effect and blue significant effect on binding affinity according to experiments.

2.2 Multiple simulation setups show good agreement with experiments

Simulations of the holo FVIII C2-domain structure were carried out using 6 different configurations: 1) the conformation of the crystal-structure simulated at 25°C (CS25); 2) the conformation of the crystal-structure simulated at 37°C (CS37); 3) the conformation of the crystal-structure where substitutions are introduced after a pre-equilibration of the original structure of 40 ns at 25°C (CS25'); 4) the conformation of crystal-structure with reversed mutation S2296C pre-equilibrated for 40 ns (further termed wild-type) at 25°C (WT25'); 5) the conformation of the wild-type at 37°C (WT37); 6) the conformation of the wild-type with parameters $mbondi = 3$ and $igb = 8$ instead of $mbondi = 2$ and $igb = 2$ simulated at 25°C (WT25*). Correlations with experimental ΔG values are shown in table 2.1.

Simulations of the crystal-structure (CS25, CS37, CS25'), containing the mutation S2296C, performed reasonably well, given the approximations of the method used, especially in light of the importance of entropy. A generous 40 ns pre-equilibration before introducing substitutions (CS25') did not have a positive effect on the correlation with experimental data. However, reverting the crystal-structure mutation S2296C to the amino-acid sequence of the FVIII wild-type and equilibrating it for 40 ns before further substitutions improved accordance with experiments in the case of WT25'. A change of the parameters for atomic radii and Generalized-Born model did significantly worsen outcomes in the set WT25*.

To gain confidence that the promising correlation of the set WT25' (figure 2.5) is persistent, this set of simulations was repeated three times with the result, that the performance of WT25' simulations were not in all cases reproducible. The third set of simulations gave somewhat poorer results, emphasising the importance of deriving mean values from multiple simulations. Correlating the mean of individual binding free energy calculations of the four sets with experimental results gave a Pearson correlation of 0.67 to experimental SPR and 0.56 to van't Hoff data and corresponding Spearman's rank correlation coefficient of 0.63 and 0.68 respectively (figures 2.7, 2.8 and 2.5). It showed that ΔG values favourably averaged out and

produced higher correlations than was expected from the individual simulations. Still, Spearman rank correlation coefficients r_s of individual simulation sets ranging from 0.46 to 0.61 were comparatively good, as were Pearson correlation coefficients ranging from 0.51 to 0.68. The best correlation that has been achieved with MM/GBSA applied to 1864 protein-ligand structures from the PDBbind database using different settings for the interior dielectric constant was r_s=0.60 and r_p=0.58 [165]. With a smaller benchmark dataset of 46 protein-protein complexes and using various MM/GBSA protocols, the best correlation achieved was r_s=0.68 and r_p=0.65 [140]. The findings presented here give reason to belief that a thorough evaluation of simulation setups is advisable before drawing conclusions about the capabilities of molecular dynamics simulations.

The overall mean error of binding free energy calculations was 5.6 kcal/mol with a higher error observed in R2220A (12 kcal/mol), WT (10 kcal/mol), N2198A (9.4 kcal/mol) and H2309A (9 kcal/mol). From figure 2.6 it can be seen that the ranking of N2198A is especially unstable. It is worth noting that N2198A had the highest standard error rate in the reported SPR measurements at 25°C, and its sensorgram and van't Hoff plot suggest that the impact of this mutation on BO2C11 binding proved challenging to characterize [44]. Whether these experimental and computational challenges are linked is unclear.

Predictions were made for two additional substitutions inspired by FVIII orthologs: M2199I (porcine) and F2200L (canine). When evaluated using a modified Bethesda assay, substitution F2200L was shown to decrease the antigenicity of FVIII with respect to BO2C11, whereas substitution M2199I somewhat increased its antigenicity [62] – in marked contrast to substitution M2199A, which induced a significant reduction in BO2C11's binding free energy [44]. Calculated binding free energies relative to WT for M2199I and F2200L were consistent with their observed (and contrasting) functional impact – M2199I was predicted to increase the strength of BO2C11 binding in a total of 6 out of 8 simulations, whereas F2200L and M2199A were predicted to decrease the strength of BO2C11 binding in all but one simulation (figure 2.9).

	SPR		van't Hoff	
configuration	Pearson	Spearman	Pearson	Spearman
CS25	0.56	0.50	0.51	0.58
CS37	0.62	0.61	0.57	0.51
CS25'	0.50	0.36	0.57	0.63
WT25'	0.65	0.61	0.57	0.78
WT37	0.63	0.68	0.57	0.64
WT25*	0.22	0.12	0.18	0.17

Table 2.1: Correlation coefficients of calculated free energies in sets of simulations with varying setups to experimental SPR and van't Hoff measurements from experiments: Each simulation set is comprised of 17 simulations of structures containing substitutions plus the unaltered structure. Non-binding mutations found in experiments are lacking an absolute measurement value and are therefore not included in the calculation of correlation coefficients. WT25' showed good ranking capability and was selected for an in-depth analysis.

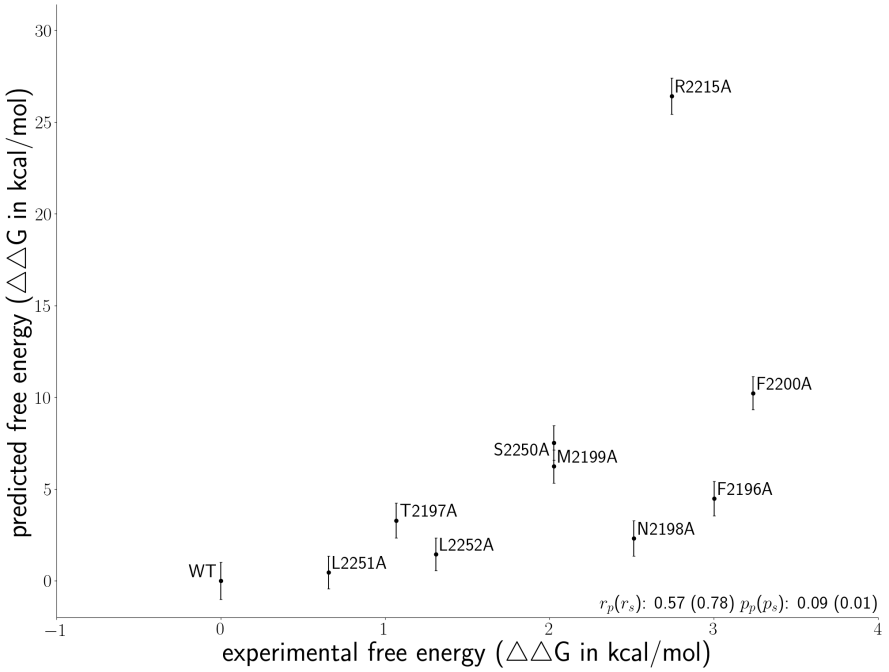


Figure 2.5: First run of the set of simulations WT25' plotted against experimental van't Hoff data ($\Delta\Delta G = \Delta G(mutant) - \Delta G(WT)$): The exclusion of the outlier R2215A greatly increases the Pearson correlation coefficient from $r_p = 0.57$ to $r_p = 0.76$ and reduces the p-value from $p_p = 0.09$ to $p_p = 0.02$. The Spearman rank correlation coefficient $r_s = 0.78$ was superior to all other simulation setups.

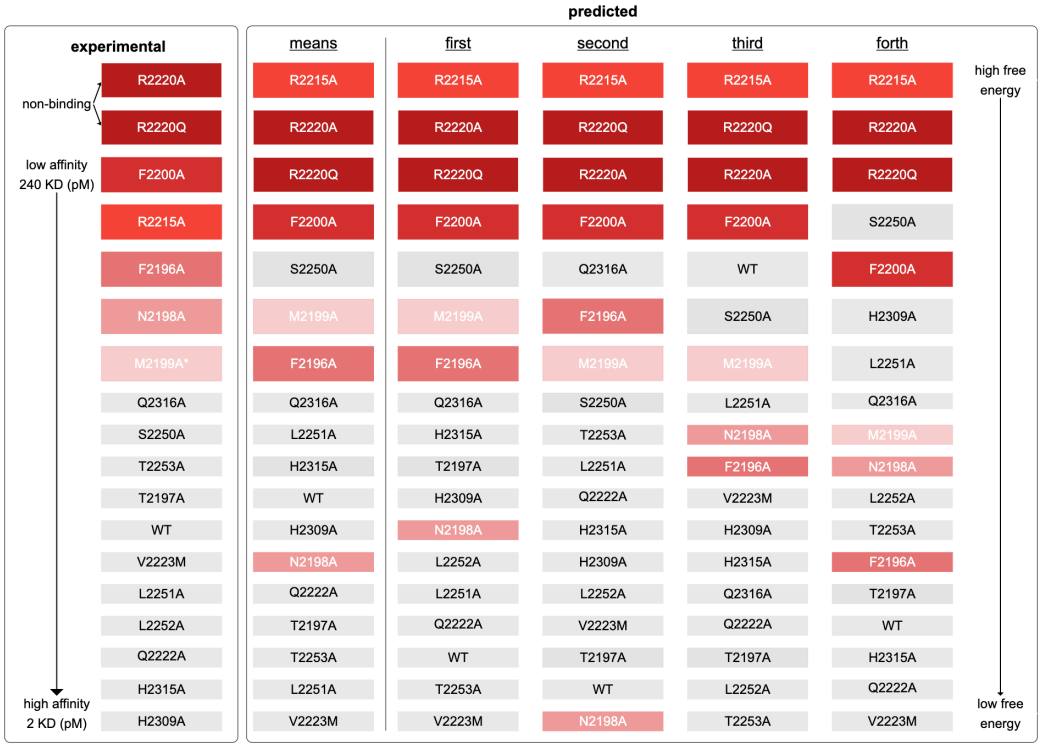


Figure 2.6: Ranking of the repeated WT25' runs: The first column depicts the ranking determined by SPR experiments. The association rate of M2199A* did exceed limits of the instrument but had been incorporated by using its van't Hoff measurement. The second column ranks calculated free energies based on the mean binding free energy values of individual simulations of the four sets whose ranking is following in the remaining columns.

Simulation set Correlation coefficient	First	Second	Third	Forth	Mean*
Pearson	0.65	0.69	0.60	0.51	0.67
Spearman	0.56	0.51	0.49	0.46	0.63

Table 2.2: Correlation coefficients of individual simulations with settings WT25' against SPR measurements: *Mean refers to the calculation of coefficients by using mean ΔG values and it is thereby different from the mean of the coefficients shown in the table; it showed that correlations coefficients of individual simulation sets were less expressive (Pearson 0.61) than when the mean of ΔG values is used (Pearson 0.67). This indicates that ΔG values which have an adverse affect on correlation favourably mean out over multiple simulations.

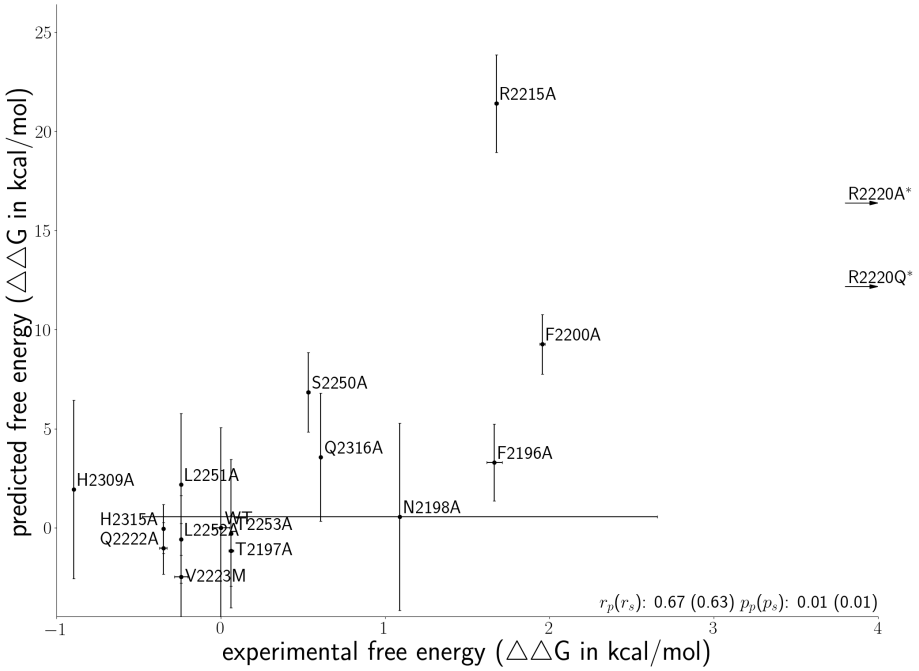


Figure 2.7: Predicted mean $\Delta\Delta G$ values of the four times repeated simulation set WT25' against experimental SPR measurements [44]: Experimental K_D values were converted to ΔG values with $\Delta\Delta G = \Delta G(mutant) - \Delta G(WT)$. R2220A* and R2220Q* have been reported as non-binders in experiments with no absolute value given. Their ΔG value would therefore be found around a very high x-coordinate which is why they are depicted as arrows. Error bars on the y axis depict the standard deviation of predicted free energies of the four times repeated set of simulations. The statistical error of 1 kcal/mol of each individual simulation is not shown here. r_p and r_s are the Pearson and Spearman correlation coefficients with p-values p_p and p_s.

2.3 A structural analysis explains the impact of substitutions

An analysis of pairwise interaction energies of the contact residues in the holo FVIII C2-domain (between the BO2C11 fab fragment and the FVIII C2-domain) in the simulation set WT25', both with and without selected substitutions, provided the means for a detailed study of the epitope.

Interaction energies of the holo FVIII C2-domain wild-type suggested that three residues in the antibody heavy chain – D52, D97 and D99 – account for

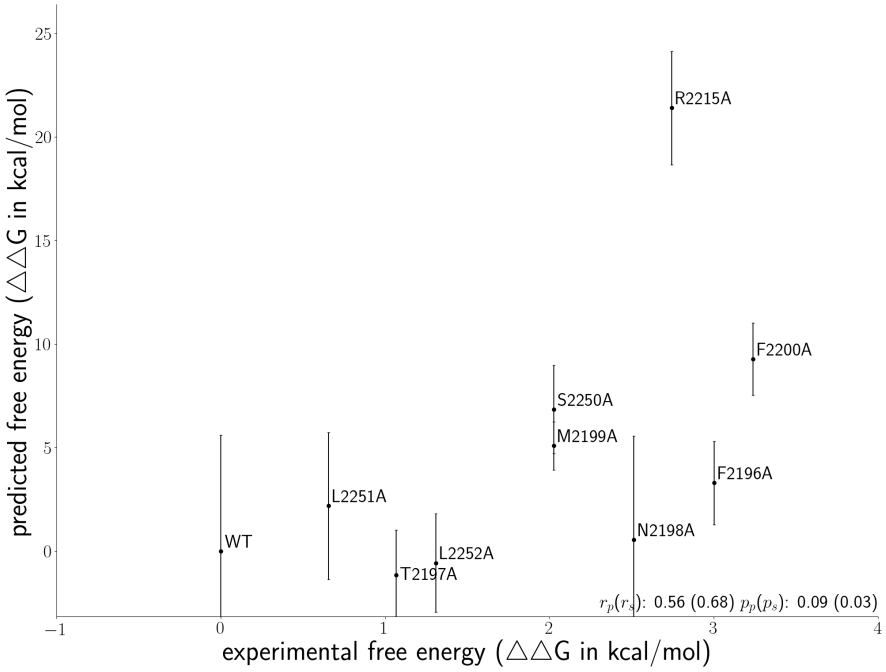
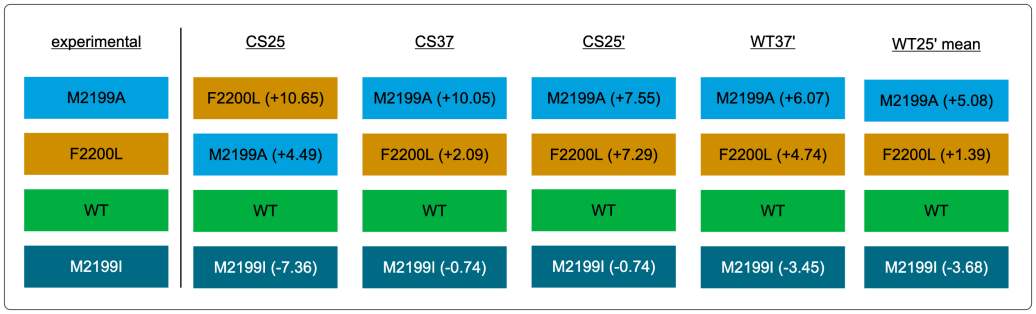


Figure 2.8: Predicted mean $\Delta\Delta G$ values of the four times repeated simulation set WT25' against van't Hoff measurements. Values were converted as in figure 2.7. r_p and r_s are the Pearson and Spearman correlation coefficients with p-values p_p and p_s.

more than half of the binding free energy (figure 2.11). On this basis, it is possible to comprehend why certain substitutions have a big impact on the binding to BO2C11. For example, substitutions R2215A, R2220A and R2220Q are predicted to break multiple hydrogen bonds between R2215 and D52, and R2220 and D99 respectively. On the other hand, it is less obvious why breaking other hydrogen bonds has considerably less impact. One such example is T2253A, which forms hydrogen bonds with key residue D97 (figure 2.11) but is located close to predicted and experimental determined binding free energy values of the wild-type (figure 2.7a). In this case, the post-substitution binding free energy decomposition suggests that the loss of bonds with D97 is partially compensated by the formation of stronger bonds at other locations, notably those associated with residues G2214, H2315, R2215 and W2203.

A comparison of substitutions M2199A (weaker binding) and M2199I (func-

(a)



(b)

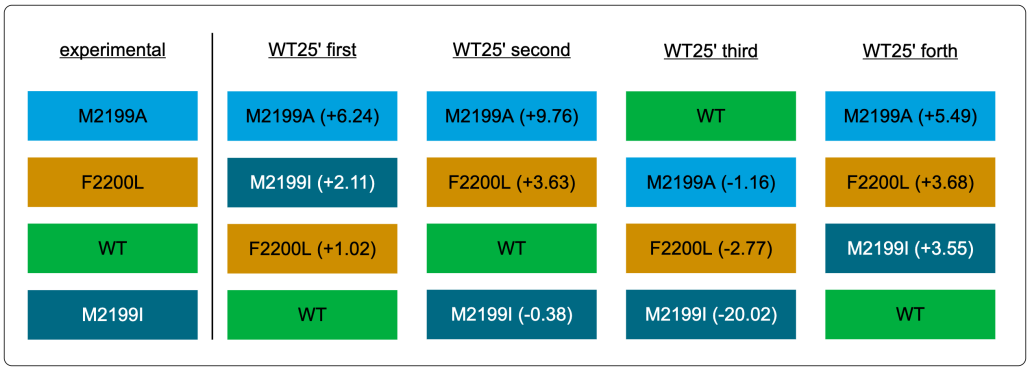


Figure 2.9: Ranking of substitutions M2199A, M2199I and F2200L with differences in binding free energy to WT shown in parentheses: The ranking of M2199A and F2200L from experimental data is not clear since they appear in different studies with methods SPR and Bethesda assay. A direct comparison is not possible but both were shown to decrease binding affinity. (a) Across different simulation setups point mutations have been ranked as in experiments; (b) An exception was the third repetition of WT25' where the WT was ranked exceptionally high. Overall, M2199I has been classified to bind stronger than WT in most cases, although some predictions were in the level of noise and moreover indicate a WT-like binding affinity.

tionally more potent, and hence potentially a stronger binder) gave insights into binding patterns that partly explained the contrary binding free energy calculations by differences in interactions of both substitutions (section 2.3.2). The predicted binding free energy did not reflect the effect of non-binding in the case of R2220A and R2220Q (section 2.3.3) and consistently exaggerated the influence of R2215A in comparison to experiments (section 2.3.1). A structural analysis suggested that this is due to shortcomings of the applied computational methods.

	T2253A	WT	T2253A - WT
F2196	-3.67	-3.18	-0.49
T2197	-4.49	-4.83	0.34
N2198	-7.33	-7.03	-0.3
M2199	-8.19	-7.58	-0.61
F2200	-7.92	-8.61	0.69
W2203	-1.08	<1	-1.08
G2214	-1.58	<1	-1.58
R2215	-25.16	-23.89	-1.27
R2220	-16.27	-15.81	-0.46
Q2222	-6.16	-6.29	0.13
V2223	-1.38	-1.32	-0.06
S2250	-10.21	-9.93	-0.28
L2251	-9.64	-10.53	0.89
L2252	-7.45	-8.08	0.63
T2253(A)	-4.6	-12.14	7.54
M2255	-1.04	-1.05	0.01
H2315	-1.32	<1	-1.32
Q2316	-3.86	-3.21	-0.65

Figure 2.10: Binding free energy decomposition of T2253A. List of binding site residues (first column) responsible for changes of up to 1 kcal/mol in binding free energy calculations of the substitution T2253A (second column) or in the wild-type structure (third column). In the forth column differences between WT and T2253A are presented. It showed that the substitution to alanine increases binding free energy by 7.54 kcal/mol at its location T2253. This however is not reflected in binding free energy calculations since the effect is compensated by stronger interactions, mainly of residues G2214, H2315, R2215 and W2203 that combined reduce binding free energy by 5.25 kcal/mol.

2.3.1 Insufficient representation of entropy and/or water-bridges may explain outlier R2215A

In all sets of simulations the effect of the substitution R2215A was overestimated. A pairwise decomposition of binding free energies of the WT suggested that bonds of R2215 contribute 25% of binding free energy, with the bond to the antibody heavy chain residue D52 accounting for 15% alone (figure 2.11). On the reduction to alanine the change in binding free energy of around 20-25 kcal/mol was apparently caused exclusively by lost bonds at the site of the substitution (figure 2.12). A possible explanation for the overestimation could be the neglected effect of water that enters the binding site that is caused by the reduction of the arginine side-chain to alanine (figure 2.13). Entropic effects as well as water-bridges could compensate for some of the loss in binding free energy upon substitution which would not be

accounted for in MM/GBSA binding free energy calculations. In the case of the substitution F2200A water is entering the binding site as well but the disregard of its effect might not be as extreme. The bonds of F2200 with the antibody are not as strong as with R2215 and a compensation of the loss in binding free energy upon substitution by water contacts might only play a minor role. Further investigations would be needed to corroborate these hypotheses.

2.3.2 A detailed structural analysis explains differences between M2199I and M2199A

Predictions for the contrasting substitutions M2199A and M2199I were consistent with experimental findings; substitution M2199A weakens the binding of BO2C11 to the FVIII C2-domain (in line with the binding assay data by Pratt and co-workers [44])), whereas substitution M2199I strengthens binding (in line with functional assays data by Barrow et al. [62])). A visualization of the difference in pairwise energies of residues in the epitope of the substitutions M2199A against M2199I highlighted that differences in binding free energy were attributable to strengthened interactions between FVIII C2-domain residue 2199 and multiple residues in both the heavy and light chain of BO2C11. To compare differences in binding patterns of single residues rather than sets, the novel visualization technique presented in figure 2.14 has been developed. By focussing on a selected residue (displayed in the middle), this technique gives the opportunity for a fine-grained analysis by reducing the threshold below 1 kcal/mol. Further, it showed that the substitution to isoleucine at position 2199 creates stronger bonds with antibody residues than the substitution to alanine which accounts for 3 kcal/mol in binding free energy (figure 2.14).

2.3.3 R2220 substitutions did not reflect abrogation of binding

Pratt and co-workers had concluded that R2220A and R2220Q muteins show “virtually no binding” to B02C11 [44]. Given the strength of BO2C11 binding, it seems highly unlikely that straightforward bonding changes induced by R2220 substitutions are sufficient to explain the abrogation of binding. This was confirmed by a pairwise decomposition of the predicted binding free energy of the wild-type where

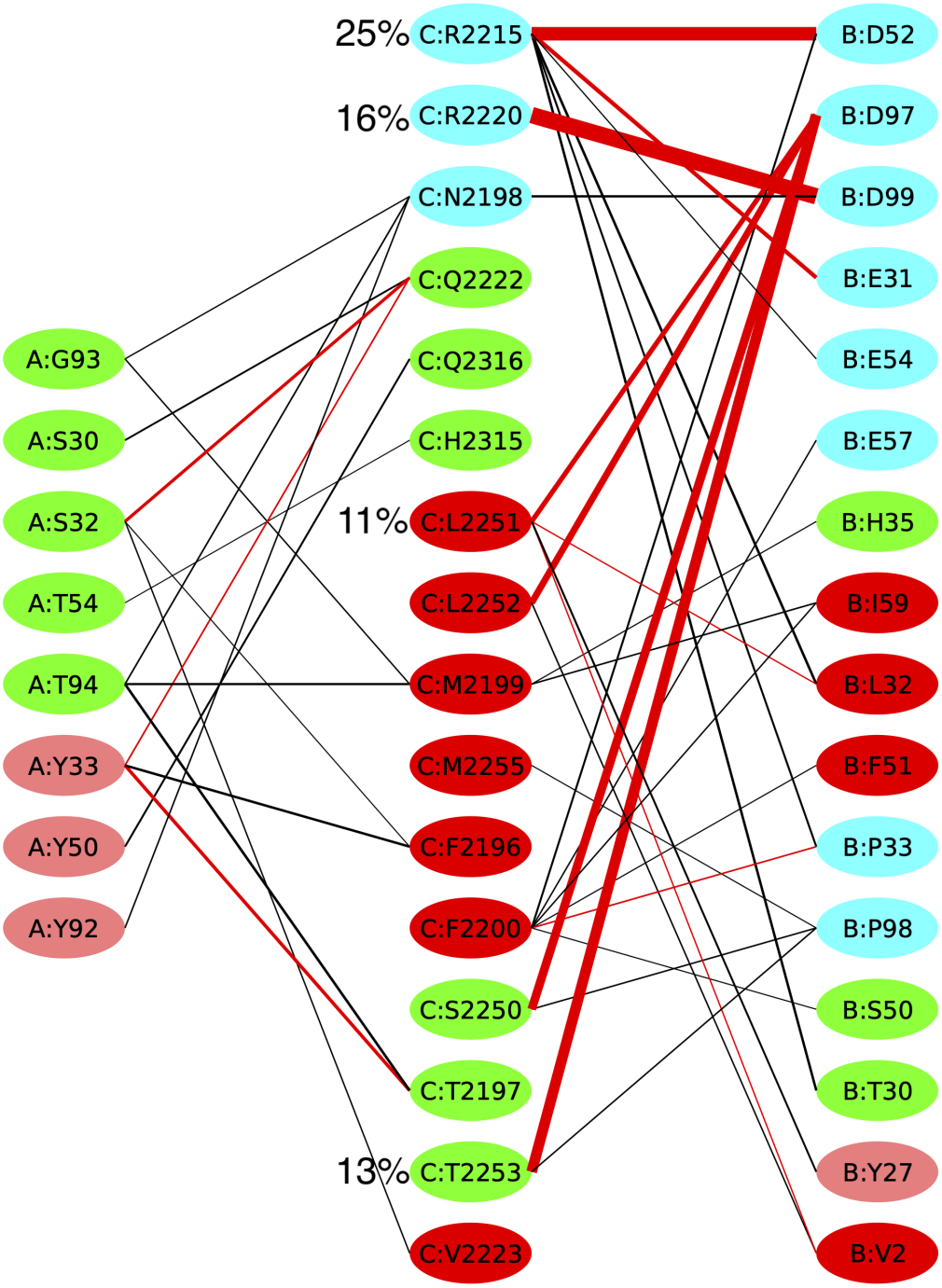


Figure 2.11: Pairwise decomposition of binding free energy of the wild-type. This suggests that R2215 has the strongest impact on binding, followed by R2220, T2253 and L2251. Interestingly, the reduction to alanine of these residue has varying effects in binding free energy calculations (figure 2.7).

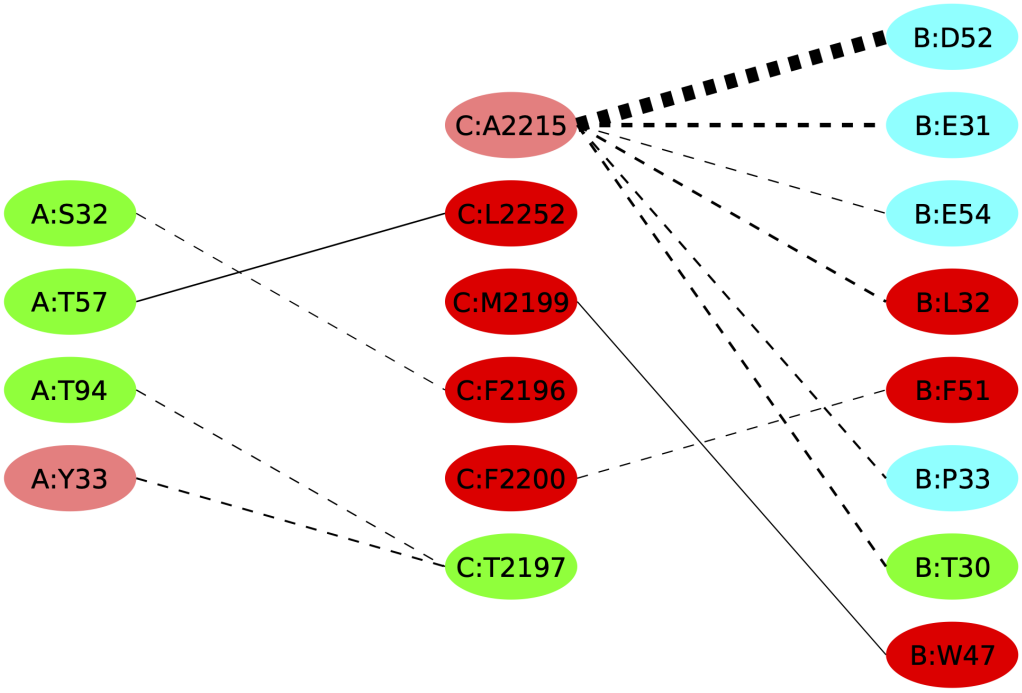


Figure 2.12: Difference of binding interactions between wild-type and mutant R2215A (denoted as A2215). The total change in binding free energy is mainly caused by the loss of bonds of 2215. No compensating bonds to the antibody form at the location of the substitution or other sites for that matter. Moreover water enters the binding site (figure 2.13).

only 16% of binding free energy reduction was attributable to the bonds of R2220 (figure 2.11). Yet, predictions of R2220 substitutions caused a larger reduction in B02C11's binding affinity than any substitution except R2215A (figure 2.7). However, figure 2.11 also shows that the salt-bridge of R2220 to D99 on the antibody heavy chain is the only means by which this residue contributes directly to binding free energy. On the other hand, R2220 has multiple contacts in the epitope, especially to the β-hairpin M2199/F2200 whose mutants were shown to have a great impact on binding (figure 2.15). Positioned between two β-hairpins, the extensive intra-domain contacts of R2220 suggest that R2220 prepossess a stabilizing function in the epitope besides its contact to D99 which has also been proposed by Nguyen et. al [70]. A reduction to alanine at this position might disturb the epitope conformation such that it does not get recognized by the antibody fab fragment. However, the substitutions R2220A and R2220Q were shown to have a relatively

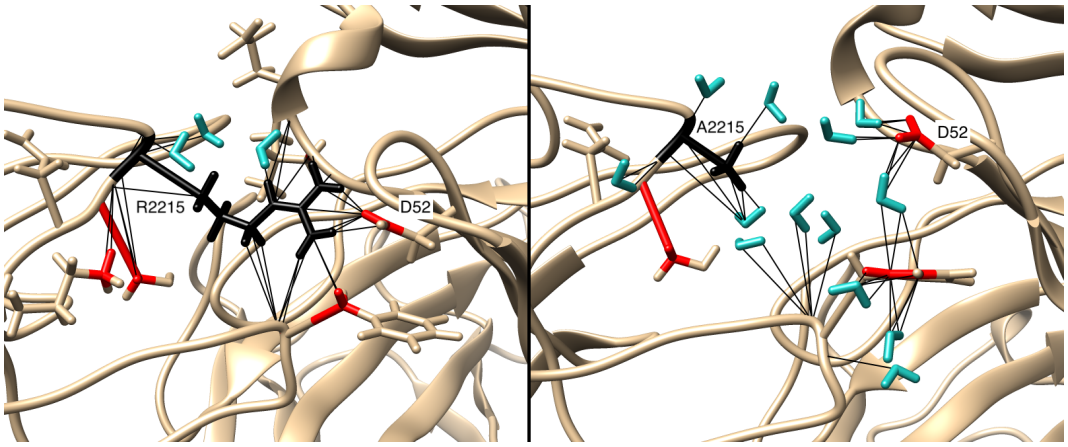


Figure 2.13: Contacts of R2215 and solvent contacts R2215A (a) bonds of R2215 and solvent contacts of R2215 contacting residues. Only few water molecules are found in the vicinity of R2215; (b) upon reduction of R2215 to an alanine water enters the binding site and binding free energy calculations indicate a very weak binding affinity. The tremendous increase of binding free energy of R2215A due to the loss of contacts with the antibody fab fragment might be counteracted by binding free energy reducing water-bridges and/or entropic contributions. Water-bridges were not considered in binding free energy calculations as well as was configurational entropy, which might explain the outlier value.

mild effect on the binding patterns of other residues in the binding site which indicates that a large-scale conformational change has not taken place (figure 2.16 and figure 2.17). To investigate this further, the Jensen-Shannon divergence used to estimate a sufficient equilibration time in section 2.1.2 can be used as a tool to visualize large scale conformational changes. For example, to comprehend the effect of a substitution the distributions of ϕ and ψ angles of all wild-type residues and of all residues in the structure including the substitution are calculated. Then the Jensen-Shannon divergence as in equation 2.1 is determined for every tuple of residues but in this case overlaid onto the structure instead of providing a histogram. By adopting this approach it is easy to spot regions that consistently differ in their backbone conformation. An investigation using this approach showed that no structural change has taken place in the epitope during a simulation of 40 ns (figure 2.18).

These findings suggest different explanations for the underestimation of R2220 substitutions in the complex structure: 1) the attractive force of the positive charge

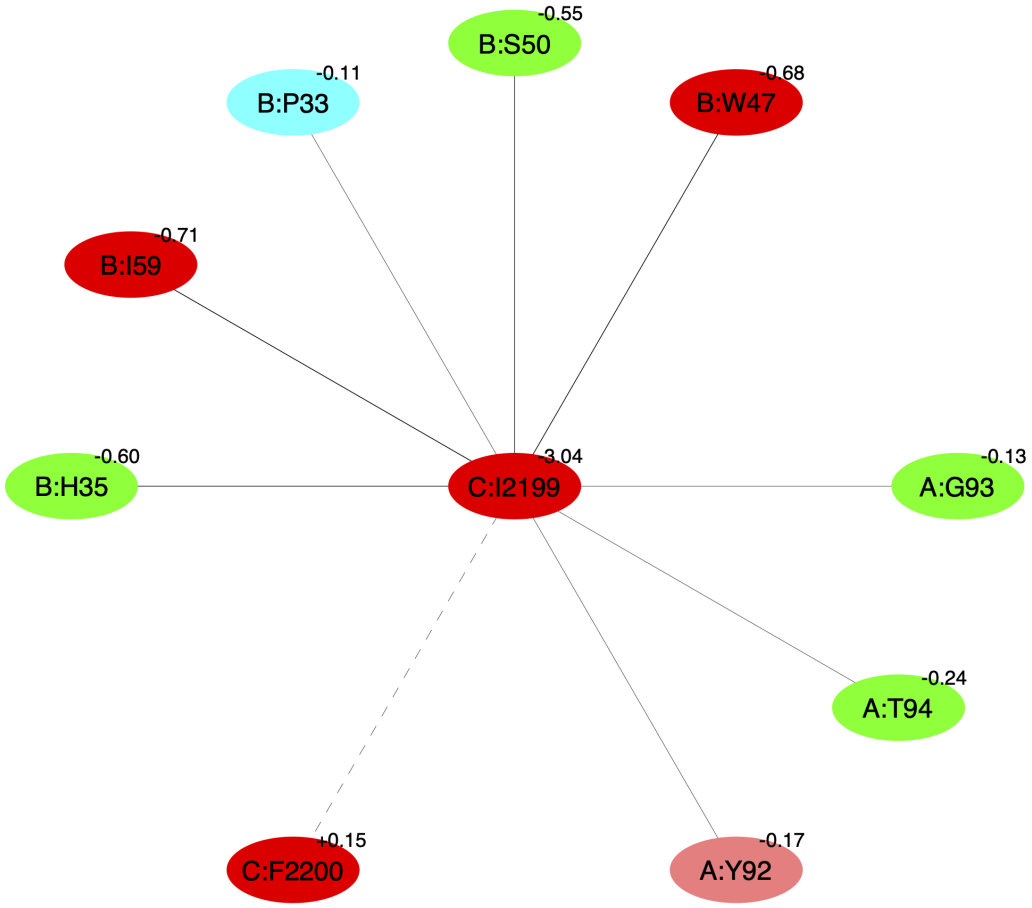


Figure 2.14: Comparison of the difference between binding patterns of M2199I (denoted as C:I2199) and M2199A; free energy values in kcal/mol are displayed alongside residues; the threshold for interactions is 0.1 kcal/mol; other elements identical to figure 1.17. The value next to 'C:I2199' in the middle indicates that the substitution to isoleucine produces around 3 kcal/mol lower free energy than the substitution to alanine. This is due to stronger bonds to the antibody heavy and light chain, e.g. the tryptophan 47 on the antibody heavy chain 'B:W47' has the value '-0.68' next to it which means that bonds of I2199 reduce the free energy by 0.68 kcal/mol more than A2199. For the light chain (letter A) contacts populate a stretch of residues from 92 to 97. Intra-domain contacts, like 'C:F2200' play a minor role.

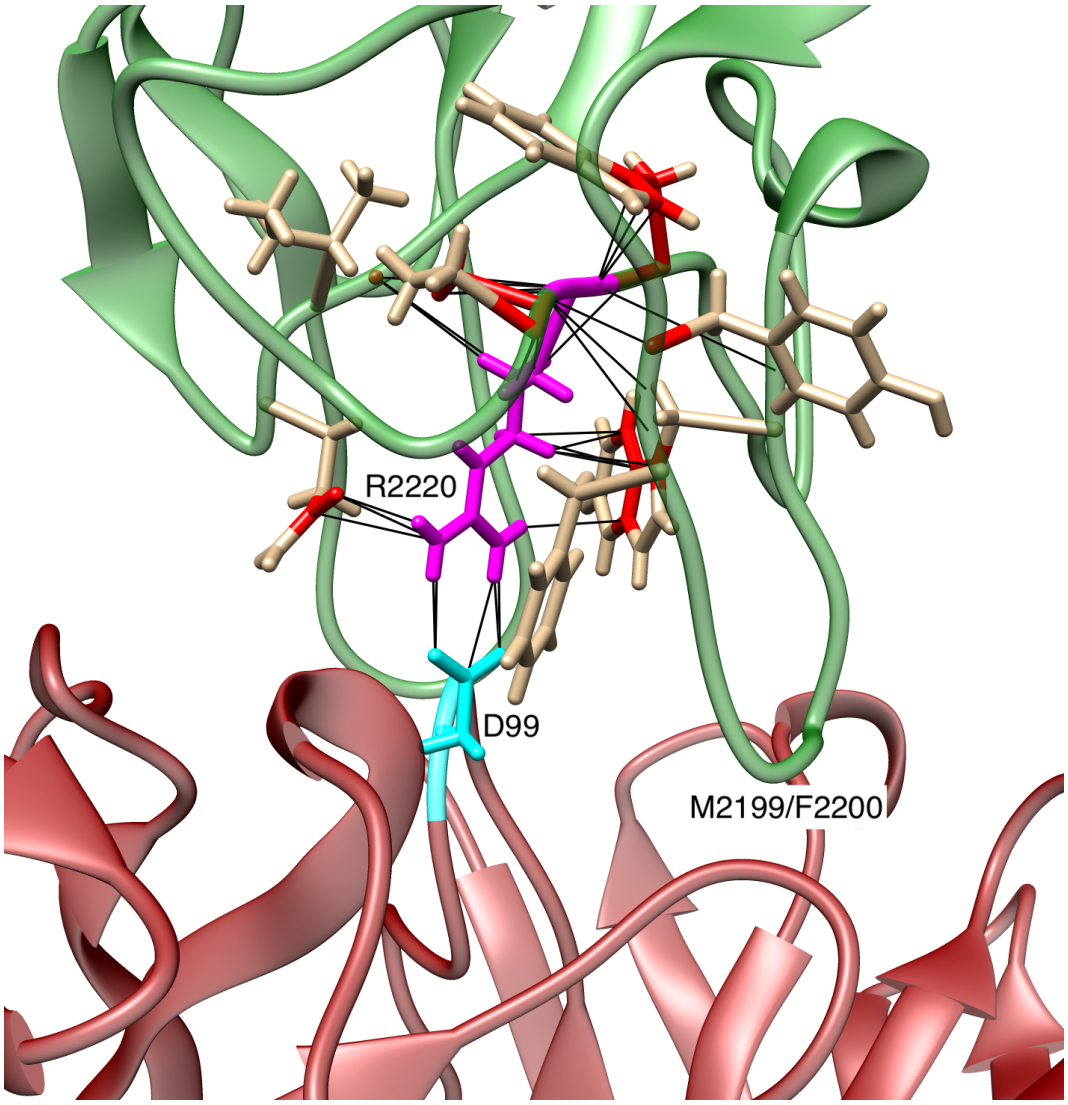


Figure 2.15: Contacts of R2220 (magenta): The only contact directly contributing to binding free energy is the salt-bridge to the antibody heavy chain residue D99.

of the R2220 arginine group is needed to initiate the binding process which is not reflected in simulations where the antibody is already bound; 2) R2220A induces a conformational change that ultimately results in dissociation from the antibody but the timescale of MD simulations is insufficient to sample this phenomenon; 3) R2220A induces a conformational change that could not take place because the antibody restricts large scale motion in the epitope once it is bound.

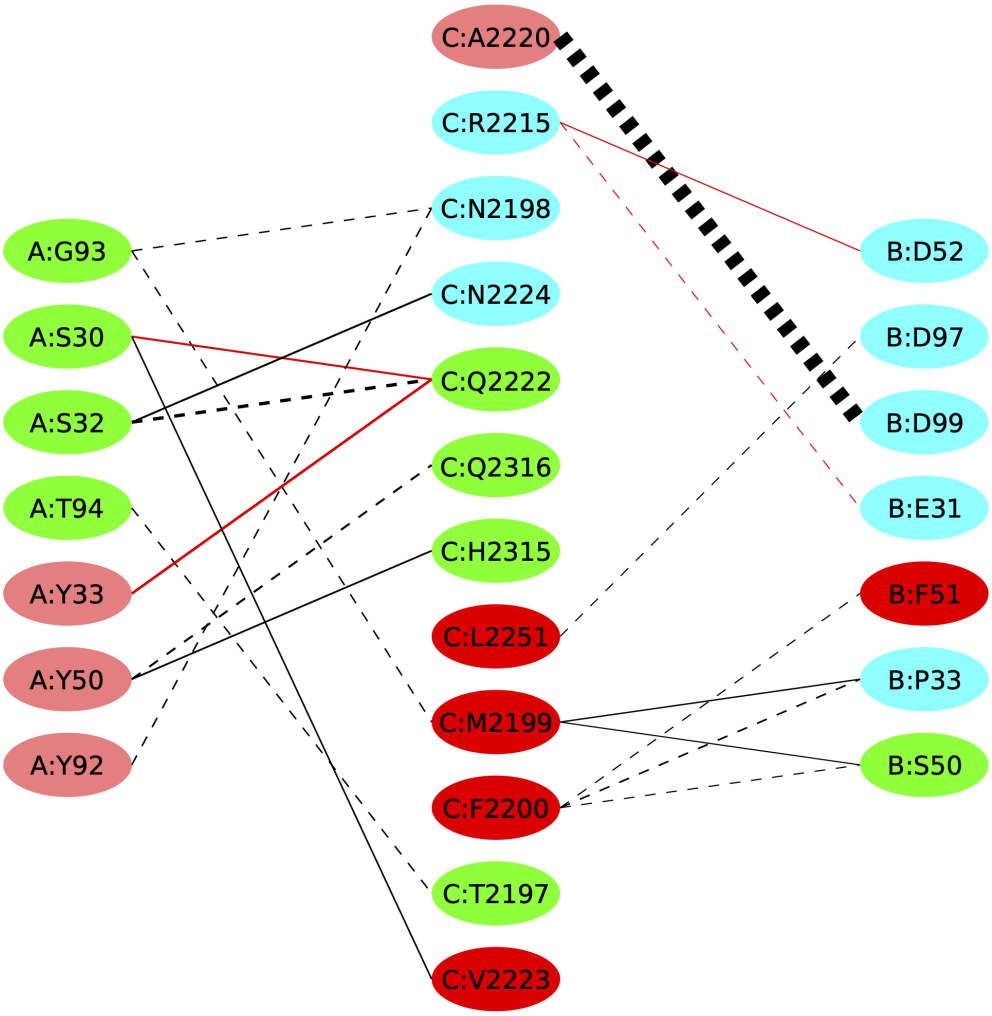


Figure 2.16: Difference of binding interactions between wild-type and substitution R2220A (denoted as A2220): The substitution of R2220 to alanine was accompanied by a change in binding free energy of 17 kcal/mol that is almost exclusively attributable to the loss of the bond to D99.

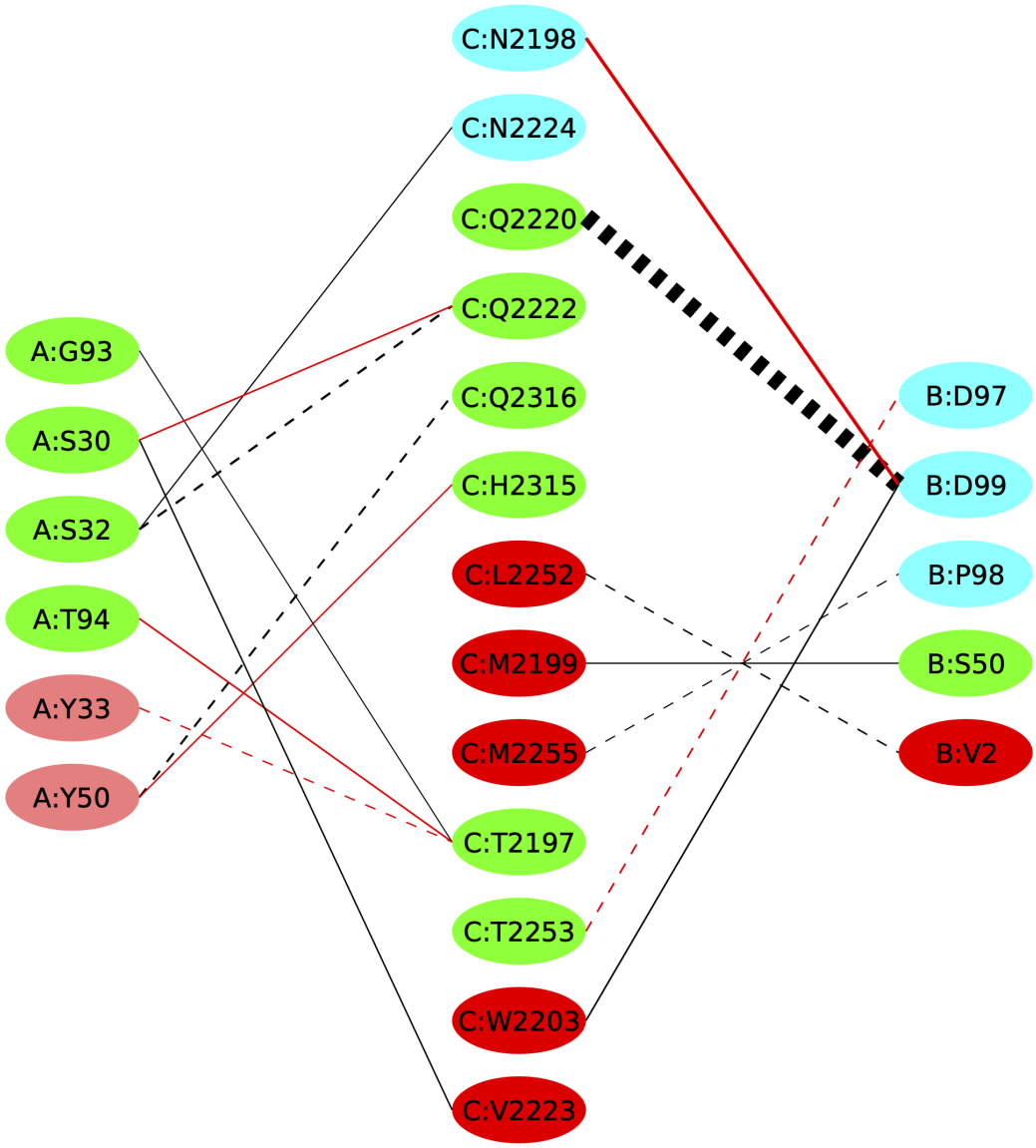


Figure 2.17: Difference of binding interactions between wild-type and substitution R2220Q (denoted as Q2220): As for substitution of R2220 to alanine the substitution to glutamine was accompanied by a change in binding free energy of 12 kcal/mol that was almost exclusively attributable to the loss of the bond to D99.

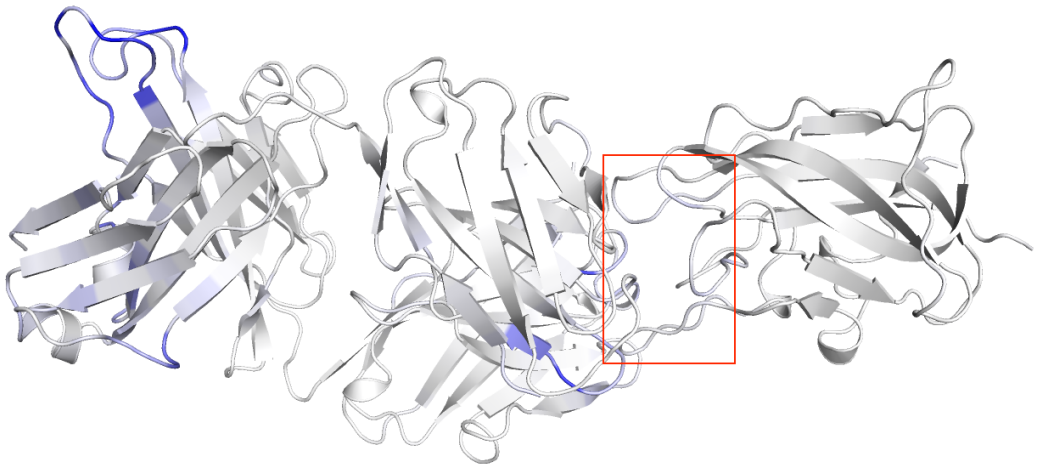


Figure 2.18: Comparison of ϕ, ψ-distributions in 40 ns simulations of R2220A and WT using the Jensen-Shannon distance measure. The red box outlines the epitope region with the antibody to the left and the FVIII C2-domain on the right. Regions in blue indicate great differences in backbone dynamics, whereas white ones indicate no or very little difference. The only region that shows a great divergence is a solvent exposed loop region of the antibody (top left) which is most probably due to insufficient sampling. The substitution of R2220 to alanine did not cause any bigger conformational changes in the epitope.

2.4 Conclusion

Multiple sets of simulations with various setups were run and evaluated consulting experimental SPR and functional binding affinity measurements [44, 62]. It showed that substitutions to alanine of residues that are supposed to mediate binding according to Pratt and co-workers [44] were successfully identified as low affinity complexes in binding free energy calculations. However, one set using a different Generalized Born model and atom radii had no significant correlation with experimental data. This introduced doubt in the method which is why a promising simulation setup was repeated three times which proved that results were reproducible.

An experimentally determined stronger binding mutation (M2199I) could be distinguished from a weaker binding one at the same location (M2199A) and a structural analysis, employing a novel visualization technique, highlighted differences in binding patterns between the two substitutions that partially explains the contrary behaviour. Further, conformations and binding patterns of an outlier in predicted free energies R2215A, and the experimentally determined non-binding mutations R2220A and R2220Q were analysed with the conclusion that the applied method of CMD with MM/GBSA binding free energy calculations might not be suitable in these cases.

An interesting avenue of research that could not be followed in the course of this PhD would have been the investigation of entropy of the holo FVIII C2-domain with methods such as interaction entropy [166]. The thermodynamic van't Hoff analysis that has been conducted by Lin et. al [44] (table 1.3) produced a range of entropy-enthalpy ratios that could potentially be used to test the usefulness and capabilities of entropy calculations.

The simulation protocol outlined in this chapter could be used to classify novel mutations and therefore provide inspiration and guidance for experimental studies. However, certain short-comings of molecular dynamics paired with MM/GBSA binding free energy calculations should be considered. The entry of water upon mutation or more generally the role of water in the binding site is poorly reproduced by MM/GBSA which might be the cause for the outlier R2215A (see section

2.3.1). Further, the one trajectory approach chosen here assumes identical conformations for the bound and separated complex. This simplification greatly reduces simulation time but flaws calculations where a bigger conformational change accompanies formation of the complex. Especially non-binding mutations such as R2220A/Q that are supposed to introduce conformational changes are difficult to classify in that manner. The omission of entropy further introduces a large offset between experimental and predicted calculated binding free energies [114]. Concerning configurational entropy, the degree of how much the formation of the complex confines the flexibility of both binding partners is typically influenced by the mutation introduced as is the entropy of water in the case where binding partners are separated.

To make the best use out of molecular dynamics simulations with MM/GBSA binding free energy calculations it is advisable to critically examine and discuss computational results before drawing conclusion or setting up further experiments [167].

Chapter 3

Accelerated Molecular Dynamics of antibody-removed and apo FVIII C2-domain

Classical Molecular Dynamics (CMD) simulations combined with MM/GBSA free energy calculations carried out in this work successfully ranked residues that contributed most to the binding affinity in experiments and gave detailed insight into the binding site between the FVIII C2-domain and antibody BO2C11. However, free energies predicted for experimentally determined non-binding substitutions R2220A and R2220Q did not suggest a drastic change in binding behaviour. This might be due to a conformational change induced by these substitutions that is inaccessible with CMD.

Accelerated molecular dynamics (AMD) is an enhanced sampling technique that reduces the time to overcome high-energetic barriers. It has been used to investigate the motion of two β-hairpin flaps in the context of HIV-1 protease by Hamelberg *et al.* [168]. They employed AMD simulations because CMD simulations did not sample the transition from holo to apo conformation upon removal of the inhibitor from the complex structure. Using AMD they were able to investigate the open and closed states and the transitional path.

A study carried out by de Oliveira *et al.* derives the open conformation and transitional states from closed to open conformation of Trypanosoma cruzi Proline

Racemase using AMD. Similar to the former mentioned work, a crystal-structure of the closed conformation, that is with bound inhibitor, was available from the Protein Data Bank. It became apparent that CMD was insufficient to sample large scale conformational changes but meaningful insights could be achieved with AMD.

To this end, AMD was used here to gain insights into molecular motions that involve overcoming comparatively high energy barriers that are unlikely to be accessible with CMD within reasonable time. Thresholds and boost potentials were calculated as has been outlined in section 1.3.1.

3.1 β-hairpin M2199/F2200 differs in the holo and apo crystal-structures

A difference between the holo and apo crystal-structure of the FVIII C2-domain epitope was described previously by Spiegel *et al.* [61] involving the β-hairpin containing residues M2199/F2200.

In this research, the twist of the hairpin mentioned by Spiegel *et al.* was quantified with respect to the angle of intersections of lines defined by the α-carbons of residues M2199 and F2200 as shown in figure 3.1. By using the definition of the twist following the definition of *reaction coordinate B* in the section 4.1, where an absolute angle of -80° was calculated for the apo FVIII C2-domain crystal-structure contrasting with a value of 4° in the case of the holo conformation, a difference of -84° was calculated. Further, results presented here are thereby comparable to the ones from Umbrella sampling simulations in the next chapter.

Since the side chain of residue R2220 whose mutants were shown to abrogate binding is in contact with this hairpin, it could be assumed that upon removal of the guanidinium group the conformation of the hairpin might be influenced (figure 3.2). Moreover, given the conformation of the aromatic rings of W2203 and F2196 it is reasonable to assume that R2220 is involved in cation-pi stacking interactions. Stacking interactions of two aromatic rings are well described by the van der Waals potential and point charge electrostatics with the force field used in this work. However, the impact of stacking interactions in biomolecules is hard to pin

down because of the interdependence of their environment [169]. A visual inspection and a calculation of angles in CMD simulations of the holo FVIII C2-domain wild-type and with the substitution R2220A showed that no large scale conformational change took place concerning the twist of the β-hairpin M2199/F2200 or in the epitope overall (figure 3.3).

A possible explanation is that epitope motion is sterically hindered by the bound antibody. To test this hypothesis, complementary CMD simulations were run where the antibody got deleted from the holo FVIII C2-domain structure (ABD). These included simulating the substitution R2220A. AMD simulations of the ABD FVIII C2-domain were conducted to investigate structural changes of the β-hairpin M2199/F2200 and also to evaluate the capabilities of the AMD simulations. The latter follows the assumption that the conformation of the β-hairpin M2199/F2200 of the holo FVIII C2-domain should revert to the conformation of the apo FVIII C2-domain crystal-structure upon removal of the antibody since the latter conformation represents a low energy basin that should be sampled using AMD. Further, the apo FVIII C2-domain structure (PDB 1d7p) was solvated and simulated using CMD and AMD. Since the ABD and apo FVIII C2-domain share the same topology and only differ slightly in their conformation (RMSD 0.4 Å) except for the β-hairpin M2199/F2200 (RMSD 1.4 Å), the analyses of their simulations should come to similar conclusions. However, results are only comparable if there is a credible proof that simulations reached an equilibrated state.

To investigate the capabilities of CMD, a 200 ns simulation was run for ABD FVIII C2-domain and the apo FVIII C2-domain. It showed that the apo FVIII C2-domain did deviate to an angle of around -50° whereas the ABD FVIII C2-domain populates states around 30° to 50° besides its initial value of 4° (figure 3.4 (a) and (b)). This plot uses a 1D Gaussian filter that aggregates values to a smooth line. General trends could thereby be spotted in the otherwise very wide spread and rapidly changing values. These plots might convey the impression that hairpin motion is rather restricted which is why the additional visualization in figure 3.5 should give a sense of angle distributions over time. The observation that both simulations

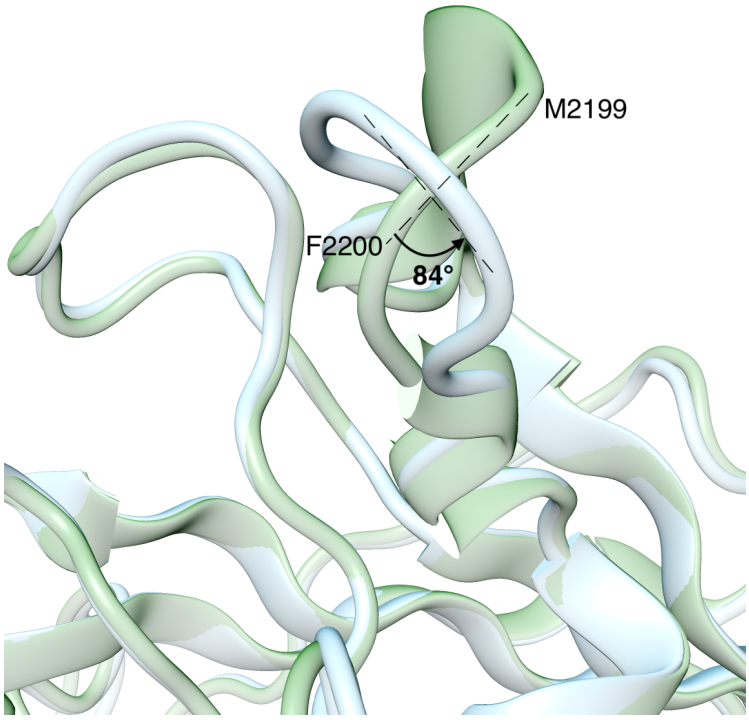


Figure 3.1: Quantification of the twist of the β-hairpin M2199/F2200: The apo FVIII C2-domain crystal-structure in green overlaid onto holo conformation of the FVIII C2-domain crystal-structure in blue. The lines defined by α-carbons M2199/F2200 intersect with an angle of approximately 84°. To increase comparability to Umbrella sampling simulations in the next chapter, reference points and calculations of this angle follows the definition of *reaction coordinate B* in section 4.1, where values of 4° and -80° were determined for the holo and apo FVIII C2-domain crystal-structure respectively.

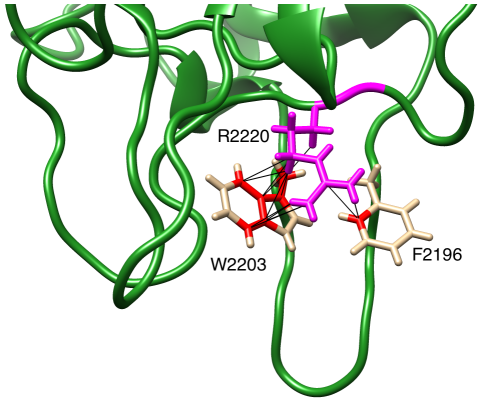


Figure 3.2: Contacts of the R2220 side-chain with residues of the β-hairpin M2199/F2200: The aromatic rings of residues W2203 and F2196 might engage in cation-pi stacking interactions involving the guanidinium group of R2220.

did not converge to the same value might be due to high energy barriers in the case of the ABD FVIII C2-domain that are not overcome in a sensible time frame with CMD.

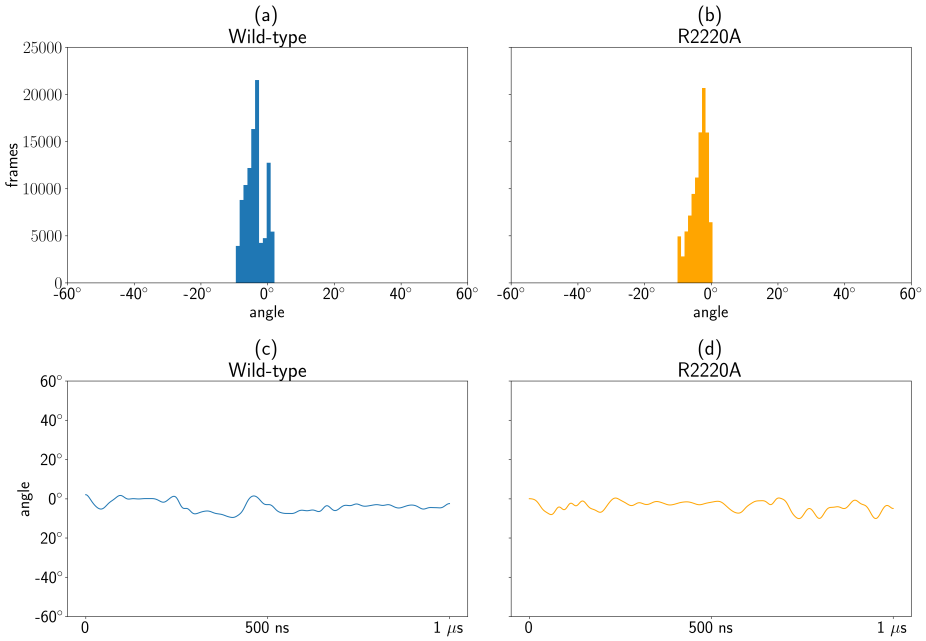


Figure 3.3: Twist of β-hairpin M2199/F2200 in CMD simulations of the holo structure: (a) and (b): The substitution R2220A did not cause a large scale conformational change of the β-hairpin compared to the wild-type; (c) and (d): No trend could be observed which might indicate that motion of the β-hairpin M2199/F2200 is hindered by the bound antibody.

Figure 3.4: Overview of trends in angles with different simulation setups: (a) and (b) CMD simulations proved to be insufficient to investigate hairpin motion since in both cases values did not converge; (c) and (d) AMD simulations of the wild-types approach a value around -45° degrees and did not resemble the value of -80° measured in the crystal-structure of the apo FVIII C2-domain; (e) and (f) Simulations with substitution R2220A did not populate a state around 4°. The absence of this state might provide an explanation for the effect of non-binding but further investigations would be needed to corroborate this hypothesis.

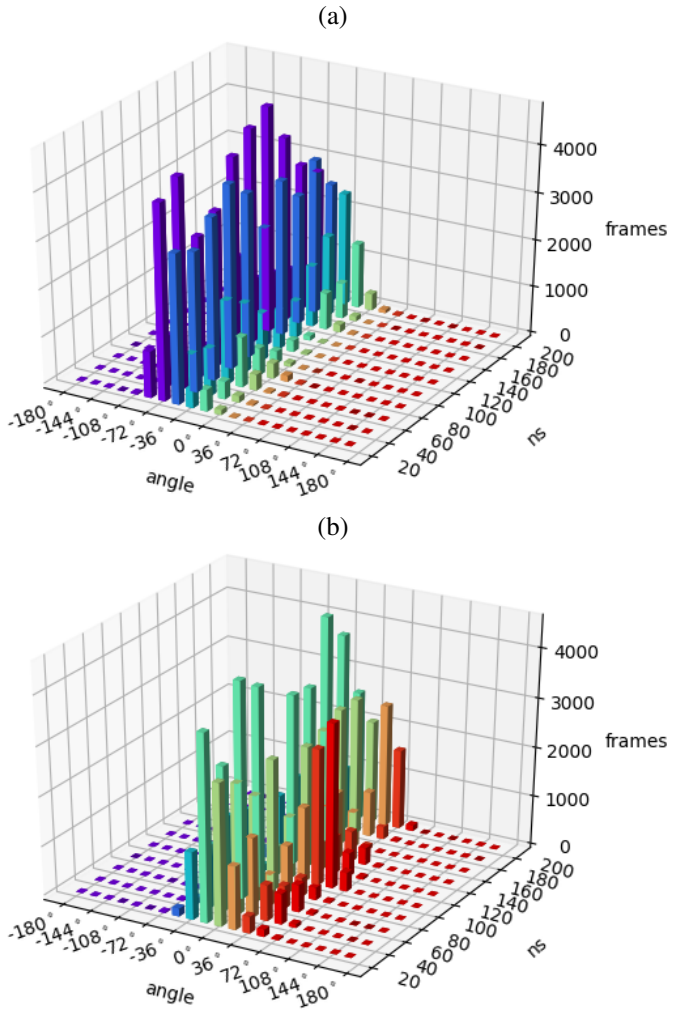


Figure 3.5: 3D histograms of CMD simulations of the (a) apo FVIII C2-domain and (b) ABD FVIII C2-domain: Even though both structures share the same topology, the twist of the β-hairpin M2199/F2200 has been relatively stable in both simulations and more importantly the ABD FVIII C2-domain did not approach the conformation of the β-hairpin M2199/F2200 of the apo FVIII C2-domain.

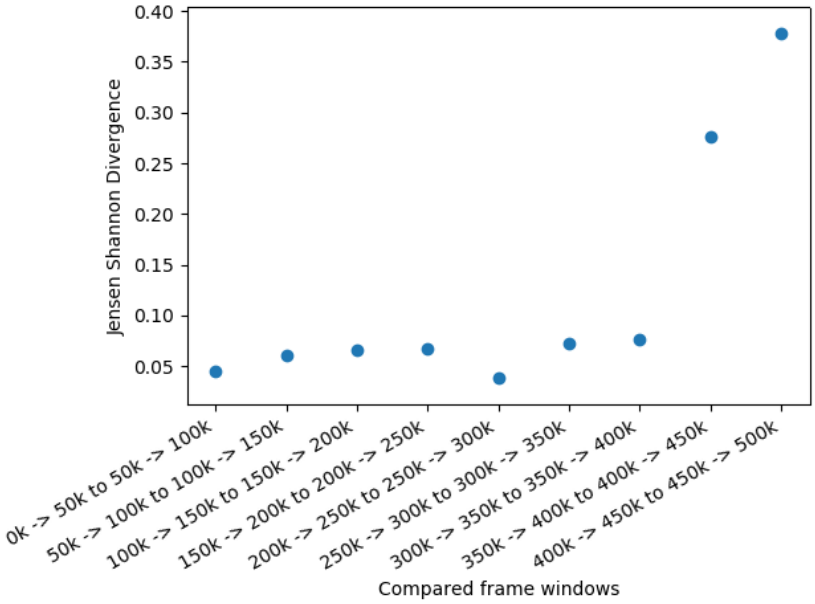
3.2 No equilibration but hairpin conformations deviate from crystal-structure

To overcome energy barriers that are not sampled using CMD, AMD simulations were performed which are adding a boost potential to enhance the sampling of the energy landscape. To check whether a biased simulation like AMD reached an equilibrated state it is sensible to get an idea whether low energy wells have been filled up by bias potentials. The energy landscape would not change its shape after equilibration since low energy wells have reached the threshold and additional boost potentials would not be applied.

However, an issue with AMD simulations is that for larger structures, like the one discussed here, it is no longer possible to recover the equilibrium due to large energetic noise [170]. Other methods, like the Jensen-Shannon divergence as employed in section 2.1.2, showed that the β-hairpin M2199/F2200 has not reached equilibrium in both the ABD and apo FVIII C2-domain (figure 3.6). For the apo FVIII C2-domain this is also apparent in figure 3.4 (c) where the angle of 4° is populated just before the end of the simulation. It is not clear when an equilibrated state would be reached in any of the simulations and statements based on the data presented here are therefore rather poorly supported.

Still, it could be observed that the twist of the β-hairpin M2199/F2200 of the ABD and apo FVIII C2-domain wild-types approach each other with average values of -45° and -34° respectively using AMD (figure 3.4 (c) and (d)). To get a sense of which hairpin conformations are most populated, a clustering using the algorithm DBSCAN of the angle values has been performed. The AMD simulations of the wild-type apo and ABD FVIII C2-domain produced three clusters each. Centroids of clusters for the apo FVIII C2-domain were -48°, -27° and 3°, where the clusters contain around 85%, 11% and 4% of the data points respectively. The simulation of the ABD FVIII C2-domain produced clusters with values and populations of -40° (86%), 9° (9%) and 19° (5%). The fact that highly populated states approach each other is an indicator that AMD simulations can produce meaningful results with the system under investigation.

(a)



(b)

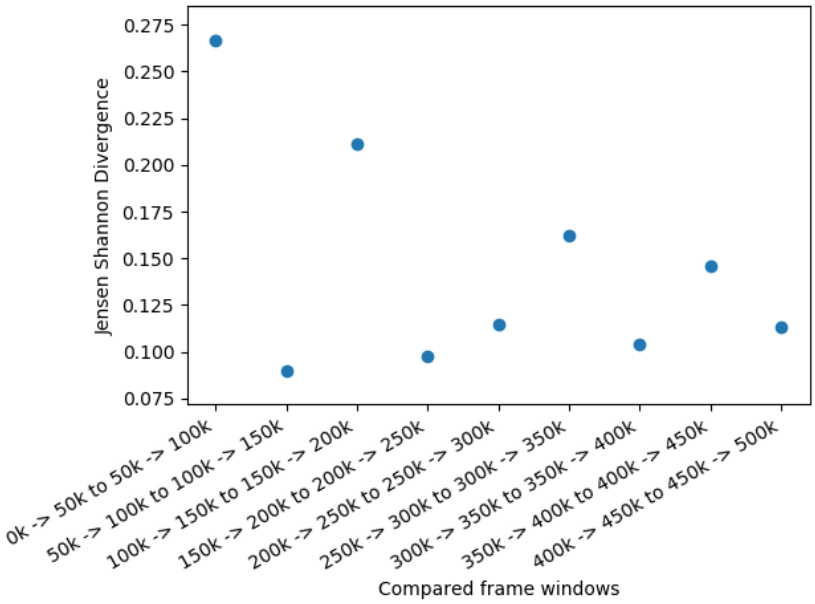


Figure 3.6: Windowed Jensen-Shannon divergence of AMD simulations: The twist of the β-hairpin is calculated in windows of 50,000 frames where successive windows are compared (e.g. '0k->50k to 50k->100K' is the Jensen-Shannon divergence of hairpin angles between frames 0 to 50,000 and frames 50,000 to 100,000) (a) For the AMD simulation of the apo FVIII C2-domain wild-type it showed that the angle population around 4° just before the end of the simulation produces a high divergence which indicates insufficient equilibration; (b) for the ABD FVIII C2-domain wild-type a trend can be observed but bearing in mind the sudden change of the apo FVIII C2-domain it might be too optimistic to assume an equilibration time indicated by this trend.

An additional observation is that highly populated states concerning the β-hairpin M2199/F2200 (with angles of -48° and 40°) are differing from the conformation of the apo FVIII C2-domain crystal-structure where an angle of around -80° was measured. This discrepancy might be due to solvation and/or the influence of crystal contacts (figure 3.7).

3.3 Substitution R2220A might influence hairpin conformation

An investigation of the twist angle of β-hairpin M2199/F2200 with AMD was also carried out for the non-binding substitution R2220A (figure 3.8). For the simulation of the apo structure with R2220A the clusters -45° (65%) and -34° (35%) were determined whereas the simulation of R2220A introduced to the ABD FVIII C2-domain produced clusters of -44° (75%) and 2° (25%). A comparison between the clusters produced in CMD simulations of the apo FVIII C2-domain WT (-48°, -27°, 3°) and the apo FVIII C2-domain with substitution R2220A (-45°, -34°) showed that the cluster with a 3° centroid (figure 3.6) is missing in the latter simulation. On the basis of these results, a cautious hypothesis for the phenomenon of non-binding is that the mutation R2220A influences the conformation of the β-hairpin M2199/F2200 in such a manner that a stable state with an angle of around 3°, which resembles the holo conformation, is no longer accessible and causes the antibody to not recognize the epitope. However, there is reasonable doubt if the simulation of the apo FVIII C2-domain with substitution R2220A is converged, even more so in the case of the ABD FVIII C2-domain with R2220A.

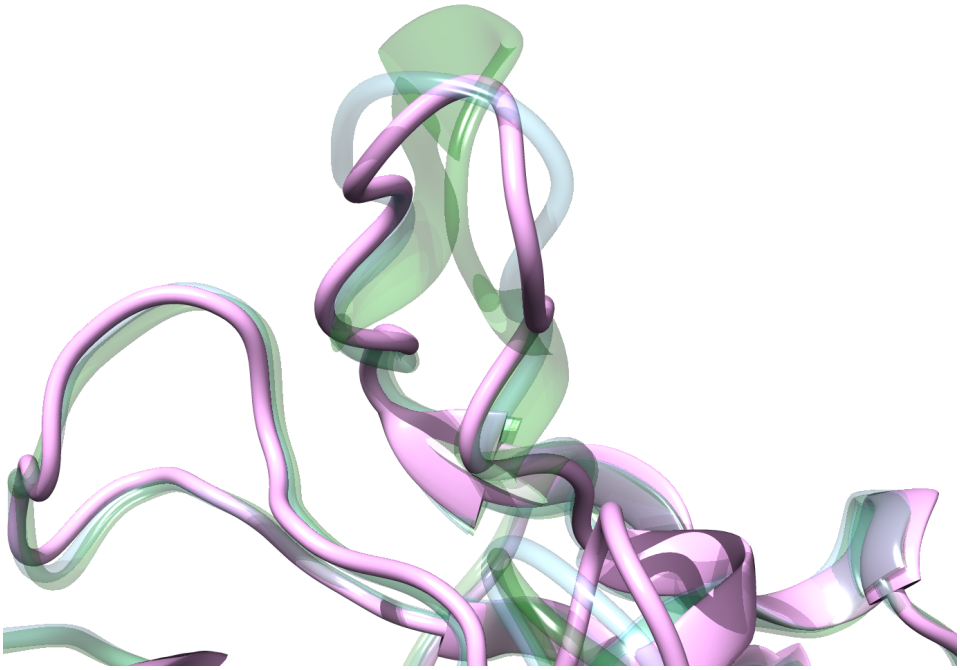


Figure 3.7: Structural view of the twist of β-hairpin M2199/F2200: The structure coloured in purple represents a frame of the AMD simulation of the apo FVIII C2-domain where the β-hairpin samples an angle of -48°, which resembles the most populated state. Blue and green depict the conformation of the holo and apo FVIII C2-domain crystal-structure respectively. It showed that the hairpin conformation of the apo FVIII C2-domain simulated in explicit solvent produces a different hairpin twist than the crystal-structures. This might be due to crystal contacts and/or effects of the solvent.

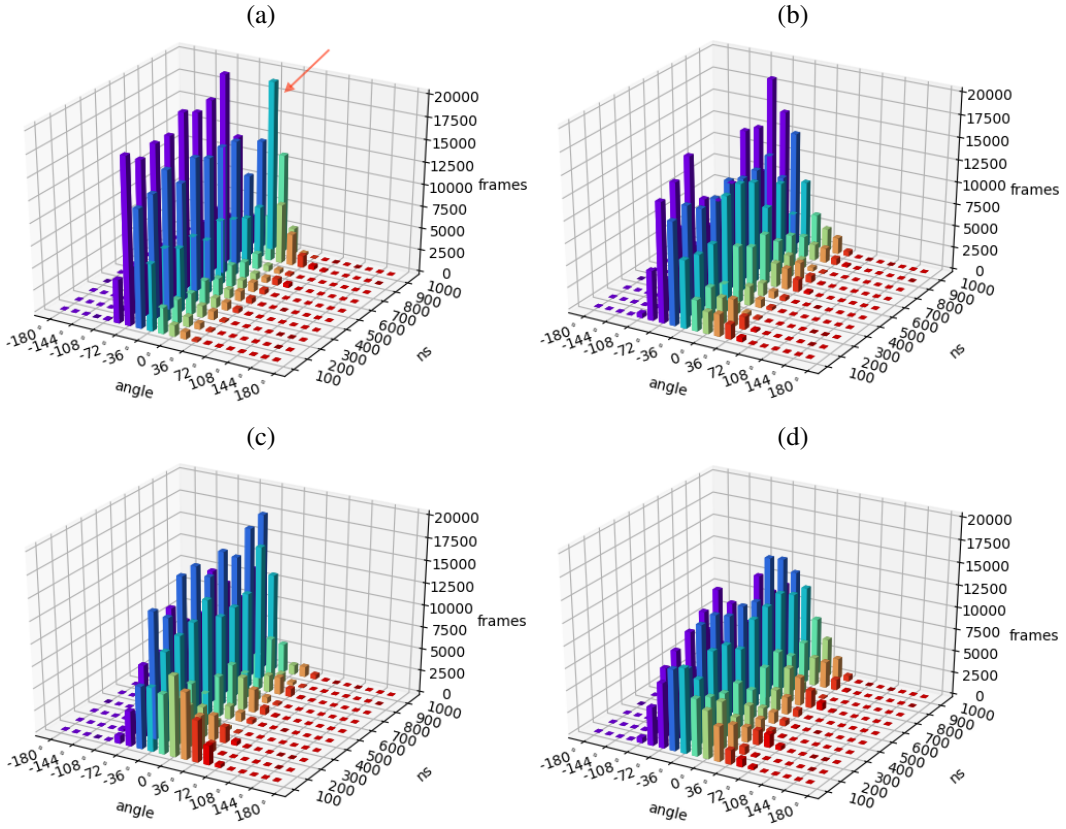


Figure 3.8: 3D histograms of the β-hairpin M2199/F2200 twist in AMD simulations: (a) The angle of the simulation of the apo FVIII C2-domain deviates after 900 ns (indicated by the red arrow); (b) It showed that the distribution of the hairpin twist angle in the case of the apo FVIII C2-domain with the substitution R2220A is much more variable and did not produce a cluster of values around 4°; (c) and (d) show the ABD FVIII C2-domain wild-type and with substitution R2220A respectively. These were the least converged AMD simulations. An explanation could be that the removal of the antibody perturbs the epitope and raises the need for longer simulations.

3.4 Conclusion

AMD simulations gave an additional view to the motion of the β-hairpin M2199/F2200 in the epitope between BO2C11 and the FVIII C2-domain. Upon removal of the antibody from the holo structure (ABD FVIII C2-domain), it appeared that the β-hairpin M2199/F2200 underwent a twist of about 50°. AMD simulations of the apo FVIII C2-domain sampled a similar hairpin twist, which was in contrast to CMD simulations and to the crystal-structure of the apo FVIII C2-domain. However, for larger systems, as the one under investigation here, a reconstruction of the energy landscape is no longer possible and could therefore not be used to investigate equilibration [170]. An analysis of β-hairpin states using the Jensen-Shannon divergence showed that equilibration has not been reached in a 1 μs run of the apo FVIII C2-domain which devalues definitive statements based on the data presented. A mere observation is that during the time of simulation, substitution R2220A introduced to the apo FVIII C2-domain did not sample the β-hairpin state around 3° which was populated in the wild-type simulation. This could point to a rarely-visited state that is needed for antibody binding. Further investigations employing for example molecular docking simulations and/or experimental techniques like nuclear magnetic resonance would be needed to support this hypothesis. Owing to the difficulties involved in assessing equilibrium of that have been mentioned, I refrained from running prolonged AMD simulations. In hindsight, the enhanced sampling technique Gaussian accelerated molecular dynamics [171] or a meta dynamics approach [124], both superior in reconstructing energy landscapes, might have been a better choice to investigate conformational states of the β-hairpin M2199/F2200. Other alterations to the protocol, such as probing different boost potentials or putting restraints on residues in the protein body to enhance the boosting of hairpin motion, might have contributed to a better understanding of the capabilities of the simulation method and/or of preferred and transient states of the hairpin. In the context of transient states and in preparation of the NMR experiments outlined in chapter 6 the hybrid method developed by Juárez-Jiménez et al. [172] would have been a sensible addition. They com-

bined the advantage of AMD in sampling conformational space with the ability of Markov State Modelling (MSM) to deduce thermodynamic and kinetic properties of the system using independent short MD simulations. In detail, system states generated by AMD are the starting point for those short MD simulations which are then consolidated using MSM to create a picture of the energy landscape.

Chapter 4

Investigating the β-hairpin energy landscape using Umbrella Sampling

Investigations using accelerated molecular dynamics showed that the twist angle of the β-hairpin M2199/F2200 of the apo and the antibody-removed FVIII C2-domain did not converge sufficiently to allow for definite statements. However, it could be hypothesised that the twist angle of the hairpin of around -80° found in the apo crystal-structure changes to -45° in solution which is suggested by AMD simulations of the apo and antibody removed holo FVIII C2-domain structure in explicit solvent.

To further investigate how the conformation of the β-hairpin containing the residues M2199/F2200 influences the free energy landscape of the apo FVIII C2-domain structure, Umbrella sampling (US) simulations were performed. US is a technique that is routinely used to improve the sampling along a reaction coordinate, typically with the aim to investigate the free energy of two states separated by an energetic barrier [173, 174]. The method is outlined in detail in section 1.4.

4.1 Evaluating simulation configurations

The reaction coordinate should optimally be defined to reproduce the apo conformation of the β-hairpin after gradually twisting the β-hairpin of the holo conformation and vice versa. To this end, two reaction coordinates have been evaluated: A) the hairpin twist defined by the torsion angle of α-carbons of residues N2198, A2201,

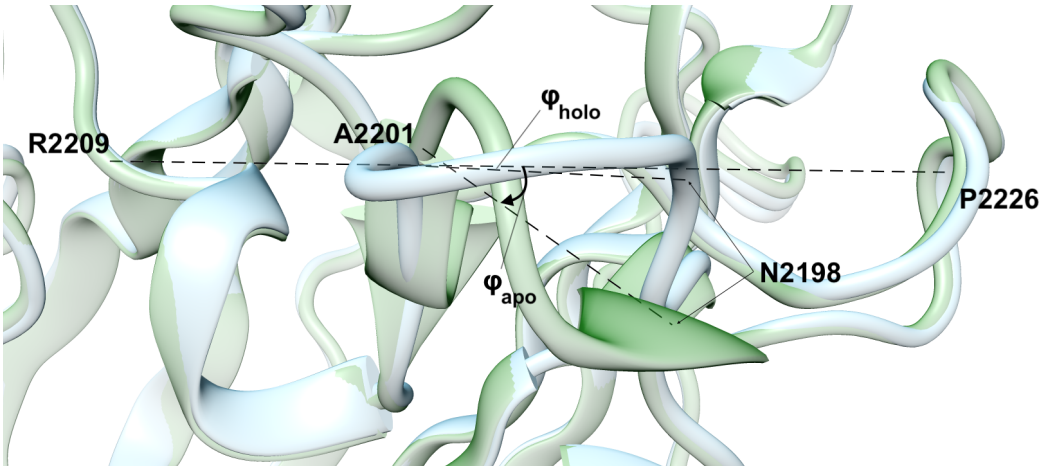


Figure 4.1: Torsion angles of reaction coordinate A: α-carbons of residues N2198, A2201, R2209, P22269 define the torsion angle with values $\varphi_{holo} = -3°$ and $\varphi_{apo} = -44°$

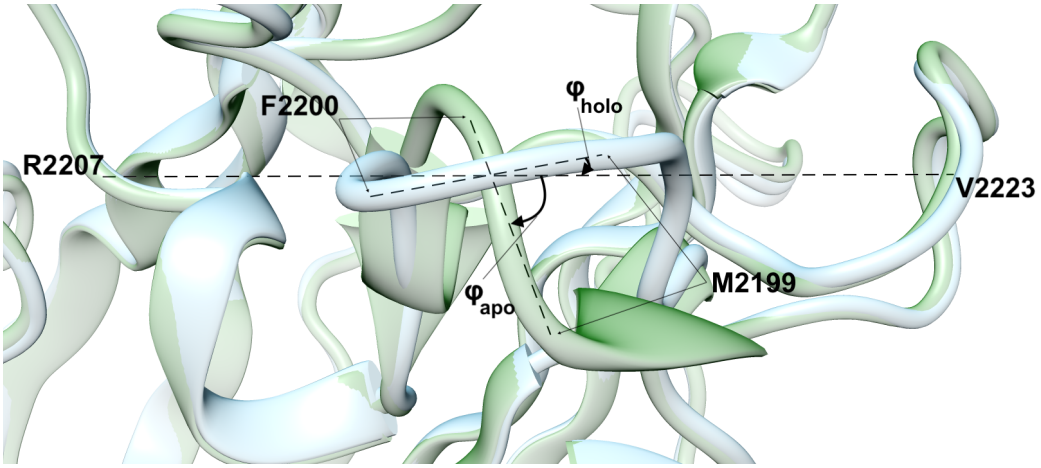


Figure 4.2: Torsion angles of reaction coordinate B: α-carbons of residues R2207, V2223, F2200, M2199 define the torsion angle with values $\varphi_{holo} = 4°$ and $\varphi_{apo} = -80°$

R2209, P2226 producing a value of $\varphi_{holo} = -3°$ for the holo and $\varphi_{apo} = -44°$ for the apo hairpin conformation (figure 4.1); B) the torsion angle of α-carbons of residues M2199, F2200, R2207, V2223 resulting in angles of $\varphi_{holo} = 4°$ and $\varphi_{apo} = -80°$ (figure 4.2). The calculation of the torsion angle follows the IUPAC recommendation [175]. US simulations were configured to map additional 50° at both ends of the reaction coordinate (a range from 130° to 50° in the case of reaction coordinate B) so that energy basins should clearly be recognizable.

Another consideration when running Umbrella sampling simulations is the length of each individual Umbrella simulation and also the amount of step-wise

change of the reaction coordinate which both determine to what degree the sampling of the reaction coordinates (each Umbrella) overlaps and in turn how accurately the potential of mean force can be reconstructed (see section 1.4). Typically, the change of the reaction coordinate (twist of the β-hairpin) increases with each individual simulation, which means that the sudden change introduced by the reaction coordinate increases as well. To dampen the effect of applied forces when twisting the hairpin a conservative step size of 2° and a successive sequence of Umbrella simulations was implemented. The term 'successive simulations' should convey that the last frame of each individual Umbrella simulation is used as the starting point for the next one (figure 4.3). This approach reduces constraining forces and the possibility for unphysical conformations that could tear the structure apart which caused problems in some simulation runs.

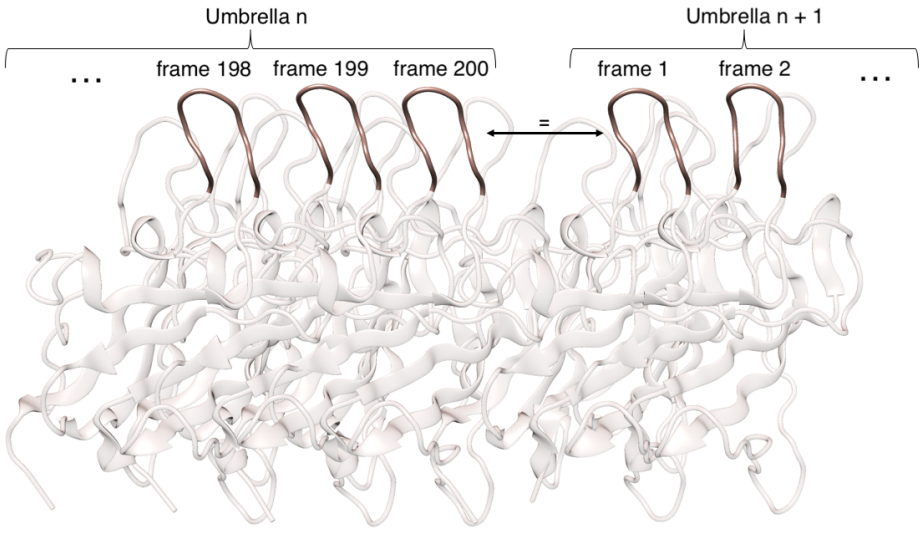


Figure 4.3: Setup of successive Umbrella simulations: Conformations of the FVIII C2-domain as found in a selection of frames of two individual Umbrella simulations (*Umbrella n* and *Umbrella n+1*) are visualized next to each other. The focus is on the highlighted β-hairpin, overlapping parts of the protein body are greyed out. The last frame of the individual Umbrella sampling simulation (*Umbrella n, frame 200*) is used as initial conformation for the next individual simulation (*Umbrella n+1, frame 1*). By that, the change of the reaction coordinate and thereby the twist of the β-hairpin from one individual simulation to the next is reduced to 2°. This gradual, 'soft' twist of the β-hairpin reduces the possibility of unphysical states like steric clashes which might arise from large changes of the reaction coordinate.

Further, simulation lengths of 1 ns, 4 ns and 8 ns for individual Umbrella simu-

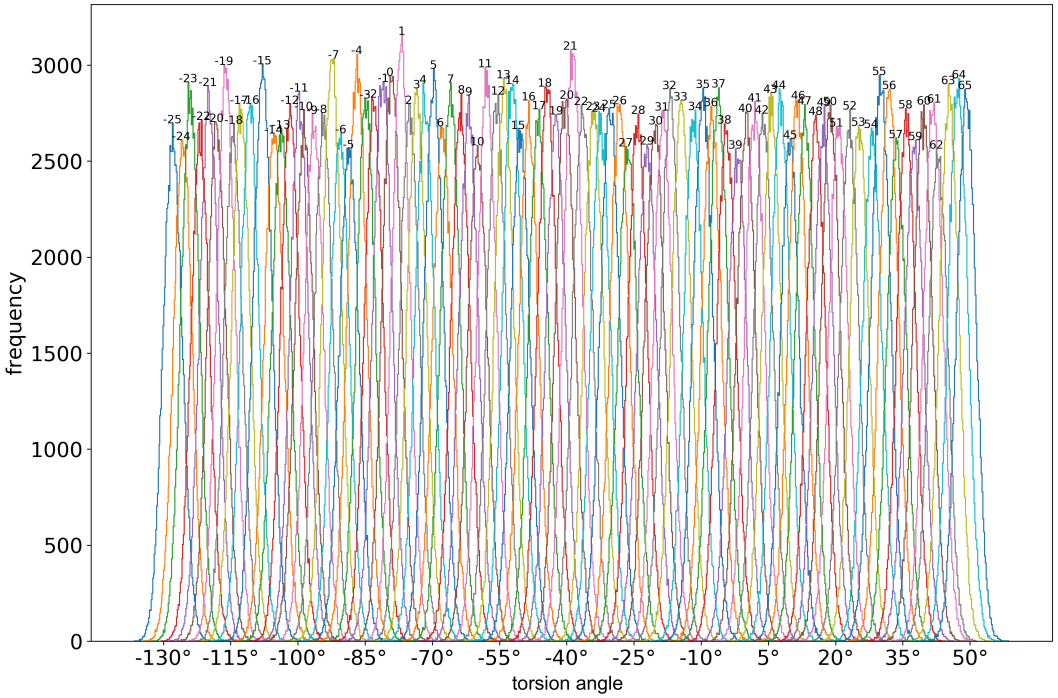


Figure 4.4: Torsion angle value distributions of individual Umbrella runs of the apo FVIII C2-domain: Above every Gaussian-like curve is the number of its individual Umbrella simulation run, where *run 0* is the simulation of the initial structure. Simulations decreasing the torsion angle have a minus sign in front. An inspection by eye proved that Umbrellas overlap sufficiently.

lations have been evaluated where a sufficient overlap of Umbrella simulations was confirmed by a visual inspection of data as illustrated in figure 4.4 for the case of the apo FVIII C2-domain.

Results illustrated in figure 4.5 showed that with a simulation length of 4 ns the estimated potential of mean force in simulations of the apo and ABD FVIII C2-domain are most similar. Ideally all curves in figure 4.5 should be akin except for statistical noise. However, the deletion of the antibody most probably introduced sources of variation non-existent in simulations of the apo FVIII C2-domain. This may indicate that the removal of parts of a crystal-structure needs special considerations and precautions.

Consulting the results of figure 4.6 and table 4.1 it showed that the reaction coordinate defined by α-carbons of residues 2199, 2200, 2207 and 2223 best resembles the β-hairpin conformations of crystal-structures. Findings outlined from now on use the following setup for Umbrella sampling simulations: Each individual

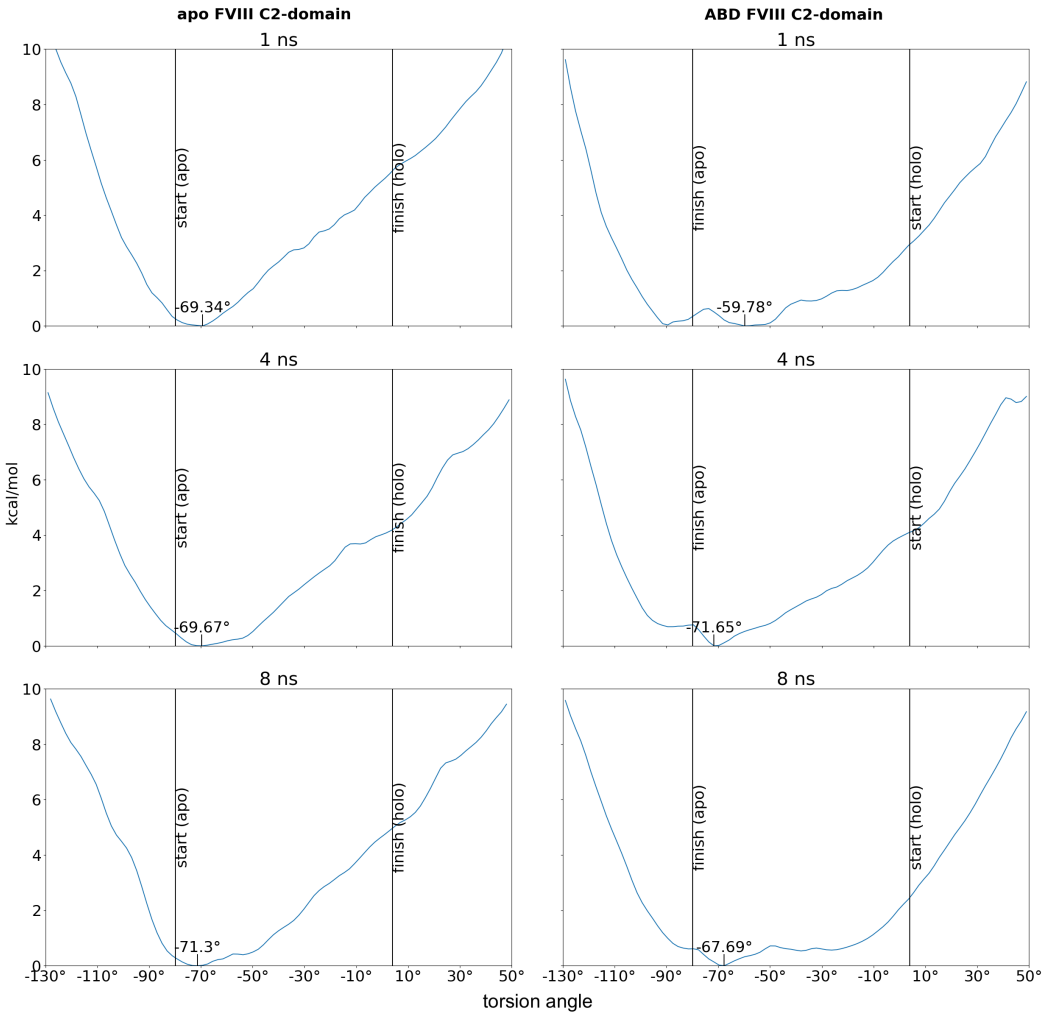


Figure 4.5: Potential of mean force of different simulation configurations: 'start' indicates the angle of the crystal-structure that was used for Umbrella simulations. For simulations of the apo FVIII C2-domain, shown in the left column, this was the angle -80°, where 'finish' is the torsion angle of the hairpin in complex (4°). It showed that for the apo FVIII C2-domain the proposed angle of US simulations carried out in explicit solvent differs from its crystal-structure value. Overall, the simulations of the antibody removed FVIII C2-domain were less consistent than simulations of the apo FVIII C2-domain.

Change of hairpin conformation from ...	2199/2200	2198/2201
apo to holo	0.399 Å	0.625 Å
holo to apo	0.774 Å	1.081 Å

Table 4.1: RMSD values after the twist of the β-hairpin: RMSD values were calculated between the average β-hairpin conformation in 200 frames where the reaction coordinate was set to the same torsion angle as found in the crystal-structures and the crystal-structure itself (figure 4.6). As an example: *apo to holo, 2199/2200*: The torsion angle is -80° in the apo conformation. After twisting the hairpin to 4° (which resembles the value of the holo conformation) the RMSD was 0.399 Å. RMSD values were calculated for the backbone of residues 2196 to 2203 (as highlighted in figure 4.3). Twisting the hairpin using the torsion angle comprised of α-carbons of 2199, 2200, 2207 and 2223 (column 2199/2200) as reaction coordinate gave the best results.

Umbrella simulation has been set to simulate a time span of 4 ns at a temperature of 25°C using explicit solvent. The reaction coordinate has been defined by the torsion angle determined by the α-carbon atoms of residues 2199, 2200, 2207 and 2223 with successive, 'soft' step-wise changes of 2°, covering a range of -130° to 50°.

4.2 Comparing the potential of mean force of non-binders

Umbrella sampling simulations of the wild-type apo and antibody-deleted holo FVIII C2-domain structure (ABD) as well as of mutants R2220A and R2220Q have been conducted and the potential of mean force has been reconstructed using the weighted histogram analysis method (WHAM) [132]. Results illustrated in figure 4.7 show, that even though trajectories are differing between the simulations of the apo and ABD FVIII C2-domain, estimated minima are in good agreement. Curves of WTs as well as R2220A mutants are fairly similar and prepossess almost the same minima around -75° to -70°. This value is broadly in line with the minimum of -80° proposed by the crystal-structure of the apo FVIII C2-domain. However, AMD simulations carried out in chapter 1.3 concluded a highly populated state around -50° which was not picked up by US simulations. Interestingly only substitution R2220Q, that was estimated to have a lower effect on total binding free

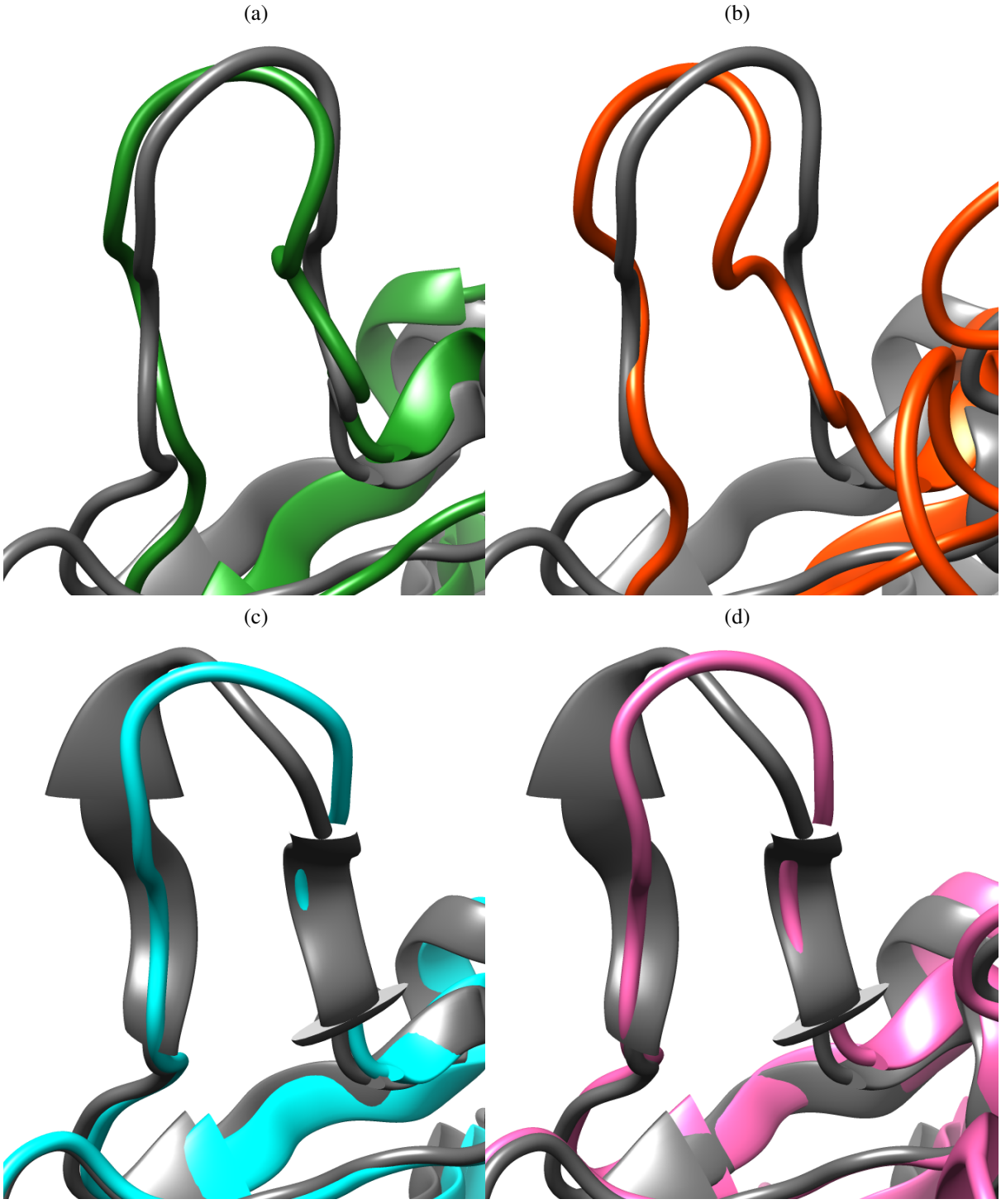


Figure 4.6: Conformations after the twist of the hairpin: (a) and (b) holo to apo conformation: The apo FVIII C2-domain crystal-structure (dark grey) with the average holo FVIII C2-domain of the Umbrella simulation having the same torsion angle as the crystal-structure using (a) reaction coordinate A and (b) reaction coordinate B; (c) and (d) apo to holo conformation: The holo FVIII C2-domain crystal-structure (dark grey) with the average apo FVIII C2-domain of the Umbrella simulation having the same torsion angle as the crystal-structure using (a) reaction coordinate A and (b) reaction coordinate B; RMSDs of hairpin backbone (residues 2196-2203): (a) 0.625 Å; (b) 0.399 Å; (c) 1.081 Å; (d) 0.774 Å

energy than R2220A in most simulation runs, shifts the minimum to an area around -55°. However, assuming thermodynamic noise of 1 kcal/mol, this difference is not significant.

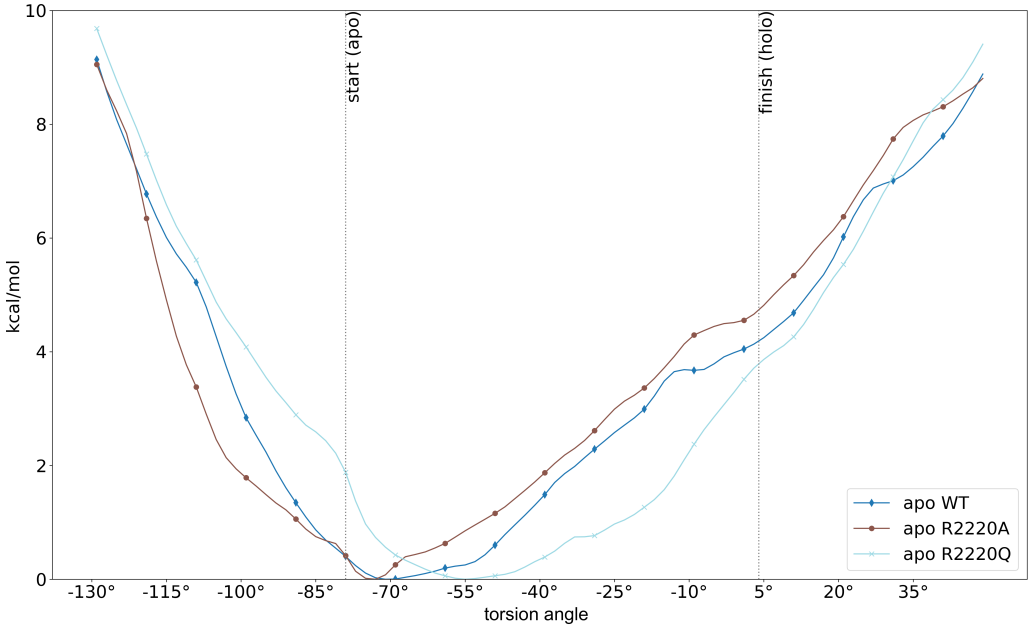
A calculation of the free energy difference (as outlined in equation 1.11) attributable to the β-hairpin between the holo and apo conformation resulted in a value of 4 kcal/mol (+/- 1 kcal/mol). Since this value was consistent in all US simulations, a correction of binding free energies (as reported in chapter 2) is unnecessary since conclusions are based on relative differences.

4.3 Conclusion

Umbrella sampling simulations were used here to estimate the impact of non-binding mutations R2220A and R2220Q on the β-hairpin M2199/F2200 which shares contacts to the side-chain of R2220. It was shown that minima of the potential of mean force of the apo FVIII C2-domain wild-type were reproducible using a range of simulation configurations. Results of the ABD FVIII C2-domain were more variable with differing minima and trajectories but non-the-less overall in line with conclusions of the apo FVIII C2-domain simulations. This was continued with the introduction of non-binding substitutions. Minima and free energy differences attributable to the conformation of the β-hairpin were comparable between both simulation sets. Presuming an error of 1 kcal/mol in calculations due to thermodynamic noise, US simulations of non-binders R2220A and R2220Q did not introduce a significant difference to the potential of mean force. The findings outlined in this chapter indicate that the β-hairpin M2199/F2200 is not confined, more flexible or largely influenced by the substitutions R2220A/R2220Q.

Still, starting structures used in these simulations might not reflect the conformation of the FVIII C2-domain with introduced non-binding mutations and therefore US simulations might be flawed. Further, as outlined in chapter 1.3, transient states might play an important role in the process of binding also with respect to the β-hairpin M2199/F2200. The influence of non-binding substitutions on transient states might be missed using US simulations even more so if these include confor-

(a)



(b)

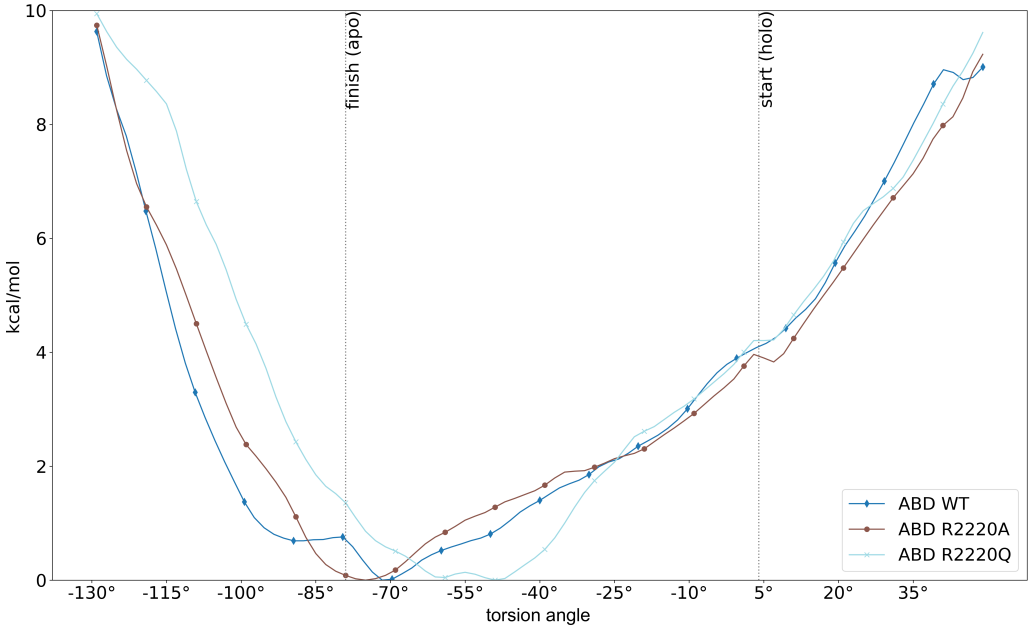


Figure 4.7: Potentials of mean force of wild-types and non-binders: (a) Potential of mean force retrieved from US simulations of the apo FVIII C2-domain and (b) of the ABD FVIII C2-domain. US simulations suggest a preferred torsion angle value of approximately -70°, although curves show some variability for different simulations setups. Most prominently, substitution R2220Q favours a conformation with a torsion angle value around -55°. It can be observed that the twist of the hairpin with substitution R2220A prefers a torsion angle of -75°. It should be noted that assuming an error of 1 kcal/mol, these differences are within the noise.

mational changes apart from the torsion angle of the hairpin. Results from chapter 1.3 support the hypothesis that conformational changes comprising multiple regions are missed with US using a singular reaction coordinate. Exclusively twisting the β-hairpin in US simulations did not result in a minimum in the potential of mean force around -50° for wild-type structures which would be expected from results of AMD simulations (chapter 1.3).

Despite these concerns, the investigation of the influence of non-binding substitutions on hairpin dynamics with US was legitimate and ruled out the possibility of a simplistic mechanism based on the torsion angle of the β-hairpin M2199/F2200 to explain the abrogation of binding.

Methods that enhance the expressiveness of US by combining it with Hamiltonian Replica Exchange MD (US/H-REMD) [176] or AMD (US/AMD) [177] might add to the understanding of hairpin motion. US/H-REMD is especially well suited to explore two reactions coordinates and enhances the sampling near energy barriers. Using this technique, multiple reaction coordinates could be explored that might improve the description of the apo and holo conformation of the hairpin. In the case of the study employing US/AMD [176], different conformations were created in AMD simulations where restraints were put on the core domain. The so created conformations of the region of interest increased the understanding of its different states and suggested two reaction coordinates to define the transition from one state over to the other. Such a sophisticated approach for deducing reaction coordinates might produce more accurate results in the case of the non-binding mutations. As noted, it is unclear how the hairpin conformation changes upon introduction of non-binding mutations and an aggressive boosting might indicate that a different reaction coordinate is more appropriate. Enhancing US simulations or the use of comparable methods like Metadynamics [124] which is also available in combination with REMD [178] might get even more important in light of results from NMR experiments that were attempted in the course of this work and which are outlined in chapter 6.

Chapter 5

Analysing the impact of binding site dynamics on binding free energy with interpretable machine learning

As outlined in section 1.6 the spatio-temporal nature of MD simulations combined with a typically extensive number of frames and a huge number of atoms complicates the detection of conformational changes that affect the measure under investigation, in this case the binding free energy. As an example, consider the calculation of interatomic distances of α-carbons atoms in the binding site and their correlation with calculated binding free energy values. This can be set up quite easily and would work well where a linear relationship is assumed (figure 5.1 (a)). However, as the formulation of force fields typically describes van der Waals interactions using a non-linear relationship (Lennard-Jones potential) a correlation coefficient would only suboptimally capture the different states of atoms concerning their interatomic distance. An interatomic distance that mainly populates two states will result in a less expressive correlation coefficient even though a relationship might be apparent (figure 5.1 (b)). In such a case a clustering algorithm could identify the different states. In a subsequent step the identified clusters could be correlated with binding free energies to understand their impact. However, clusters might be ill defined in a case where there is a linear relationship (figure 5.1a) and a normally time-consuming fine-tuning (e.g. probing for the optimal number of clusters in a

k-means clustering [179] or for the parameter cluster-density in a DBSCAN clustering [180]) would be needed. Another approach, principal component analysis [181], has other drawbacks which are discussed in section 1.6.

Tools like WISP [182] focus on long range conformational changes but lack the ability to analyse the influence of these on binding free energy. So far, establishing a rationale based on conformational changes at the atomic level to explain their impact on binding free energy is laborious and trust in the technique has to be established by manual evaluation of the outcomes. Hence, structural design that takes full advantage of the atomic resolution of MD simulations is a complicated process.

Potentially, a machine learning model that is able to encapsulate both linear and non-linear relationships could be trained to predict binding free energy from motion in the binding site or beyond, assuming that changes in distance between binding site residues as well as backbone and side chain angles impact binding free energy, which is a reasonable assumption. Such a model would possess the information how even small conformational changes in the epitope, e.g. interatomic distances, weaken or strengthen the bond. If this information would be accessible to a researcher it would take away the laborious if not impossible task of assembling this information at this level of detail himself or herself.

With the advent of deep neural networks, machine learning has become increasingly popular. The predictive power of such networks is superior and has many success stories in a range of fields, the most prominent being computer vision [183]. However, deep neural networks, with the exception of a few architectures like autoencoders, are considered 'black boxes' and methods like 'Layer-wise Relevance Propagation', that aim to elucidate which features of the data contribute to the network's decisions, are not usable out of the box and have yet to become standard tools [184, 185, 186]. Another disadvantage of deep neural networks is the lengthy process of finding the best architecture for a given task and the tuning of model parameters.

Explainable AI (XAI) is an umbrella term for machine learning techniques

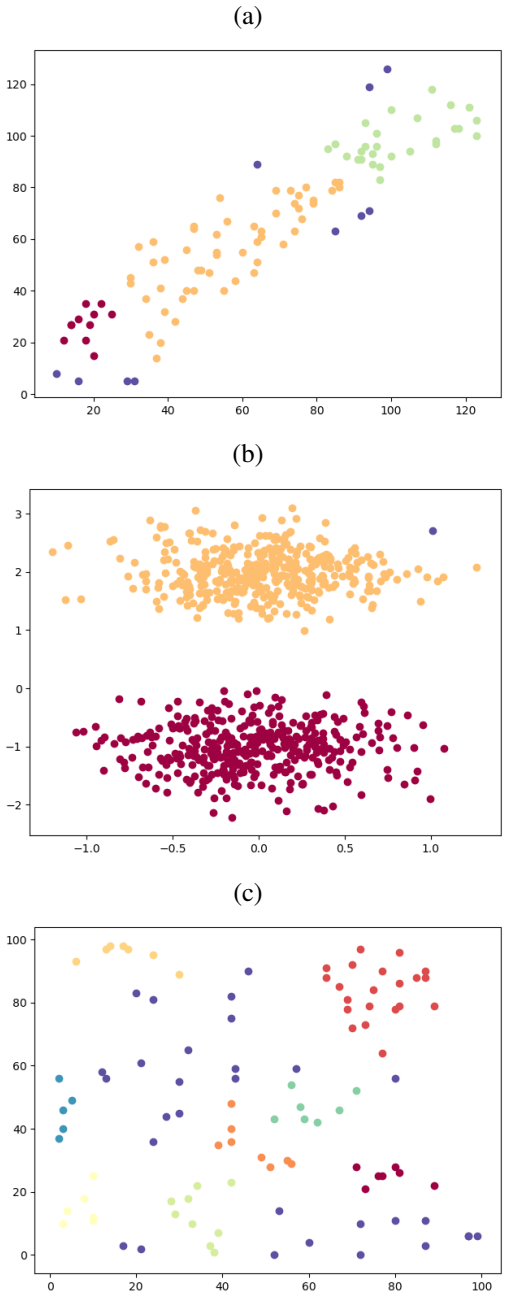


Figure 5.1: Linear and non-linear data: (a) A linear relationship between x and y values is illustrated. Clusters, as determined by DBSCAN, are indicated by colour. It shows that the calculated Pearson correlation coefficient of 0.92 with a very low p-value indicates that there is a strong linear relationship whereas a clustering is not well suited to highlight this relationship. (b) A non-linear two-state relationship is illustrated. The clustering algorithm is picking up the two states but a Pearson correlation coefficient 0.07 would naively lead to the conclusion that there is no relationship between x and y values. (c) A random distribution with a Pearson correlation of 0.05 and a non-expressive clustering cannot be distinguished from a non-linear relationship and a linear relationship respectively.

that train models that can be analysed in such a way that humans can gain intuitions about how a given model generates its outputs. This contradicts the commonly accepted idea that machine learning methods produce 'black box' predictors with little or no options to understand their inner workings. Tree models are a branch of machine learning methods that are interpretable and do not share the issue of time-consuming optimizations. Other than network architectures, tree models favour interpretability over prediction accuracy which makes them attractive for tasks where decisions have to be comprehensible [187]. Tree models have modest requirements when it comes to compute resources which is why they have been used extensively in the past. Probably because of this, the implementation, training as well as analysis tools for tree models are mature and usable in an out-of-the-box fashion.

5.1 Methods

5.1.1 Training of meaningful tree models using XGBOOST

Gradient boosted trees are a set of supervised learning algorithms that can be used for classification as well as regression problems. XGBOOST [140], standing for eXtreme Gradient Boosting, is an implementation of gradient boosted decision trees that has superior computational speed and greatly reduced resource needs over other implementations of boosted tree algorithms [188].

Boosting refers to combining weak learners that have been build in a stage-wise fashion. A weak learner is a model predicting better than chance but not by much. Training of a gradient boosted tree is initiated by a predefined tree (a weak learner) that splits the data based on a default value. In the successive stages weak learners are added that are trained on residual errors of the combined previous weak learners. Note, that the notion of combining weak learners is similar to Random Forest [189]. However, gradient boosted trees are build in an additive manner which means that a final prediction is the combination of the prediction of a sequence of weak decision trees rather than a majority vote of a "forest" of weak decision trees, which is the case for Random Forest. In general, models employing weak learners are better in modelling data with high dimensionality over extensive singular trees [190]. This is

because when a new weak decision tree is added in gradient boosting it accounts for details of the data that are not reflected by the tree so far and therefore each added tree improves the model fit. In regression with XGBOOST, errors are reflected by the residuals of the combined previous models. Following is an outline of the steps to train a Gradient boost model with XGBOOST:

1. Use one leaf tree with value of 0.5[1] and calculate residuals.

2. For a tree $m \in \{1, M\}$

   (a) Put all the residuals in the root leaf of a new tree.

   (b) find optimal output value of leaf by minimizing:

   $$\underset{O_{value}}{argmin} \sum_{i=1}^{n} L(y_i, p_i + O_{value}) + \gamma T + \frac{1}{2}\lambda O_{value}^2$$

   where y_is are the residuals, p_is are the predictions and O_{value} is the output value. The loss functions has to be differentiable since XGBOOST uses a Second Order Taylor Expansion to approximate it. λO_{value} as well as γT are regularization terms.

   (c) Having found an optimal output value we can then calculate a so called similarity score for that leaf:

   $$similarity\ score = \frac{(\sum_{i=1}^{n} g_i)^2}{[\sum_{i=1}^{n} H_i] + \lambda}$$

   where g_i are the gradients and H_i are the Hessians of the loss function.

   (d) Iteratively split the feature space[2], build a tree for each split and calculate optimal output values and similarity scores for the leaves[3].

   (e) Calculate the gain for each tree by:
   $Gain = Similarity\ Score_{left} + Similarity\ Score_{right} - Similarity\ Score_{root}$

   (f) Continue with tree that has the highest *Gain*.

[1] this is the default value in XGBOOST, other approaches use e.g. the mean of the target values
[2] XGBOOST splits at predefined quantiles to reduce the number of splits to try out
[3] these trees are trained in a greedy manner as they only provide locally optimal solutions

(g) Grow branches (sub-trees) from leaves (step (b) to (f)) until a termination criterion is met (multiple criteria can be defined, e.g. total training time, tree depth, minimum number of values in leaves).

(h) Prune the final tree by removing branches if their *Gain* is below a predefined threshold.

(i) Add remaining tree after pruning k_m to the model:

$$F_m(x) = F_{m-1}(x) + \nu k_m(x)$$

To prevent overfitting a learning rate ν reduces the influence of individual trees.

(j) Calculate residuals.

(k) Repeat (a) to (j) until a termination criterion is met.

As can be seen in step (i) the final model sums up all the predictions of the successively trained weak learners, which is why gradient boosting is a type of stage-wise additive modelling.

Gradient boosting with XGBOOST was chosen in this work because of the possibility to analyse the impact of features, which is straightforward with the Python package SHAP [191]. Further considerations included the availability of a Python package, manual labour to set up the training and analysis of the model and computational requirements.

XGBOOST has been used in the field of MD before to train a model that predicts meta-stable states and identifies essential internal coordinates of these states [192]. The training data of Brandt *et al.* are distances and angles retrieved from previously clustered meta-stable state coordinates. To extract the most important features (distances/angles) they successively remove features that are deemed unimportant, based on their impact on accuracy.

In this work, a tree model was trained on distances and motion of binding site residues to predict the change in Gibbs free energy (ΔG values). Using machine learning terminology the former would be called 'the features' and the latter 'the

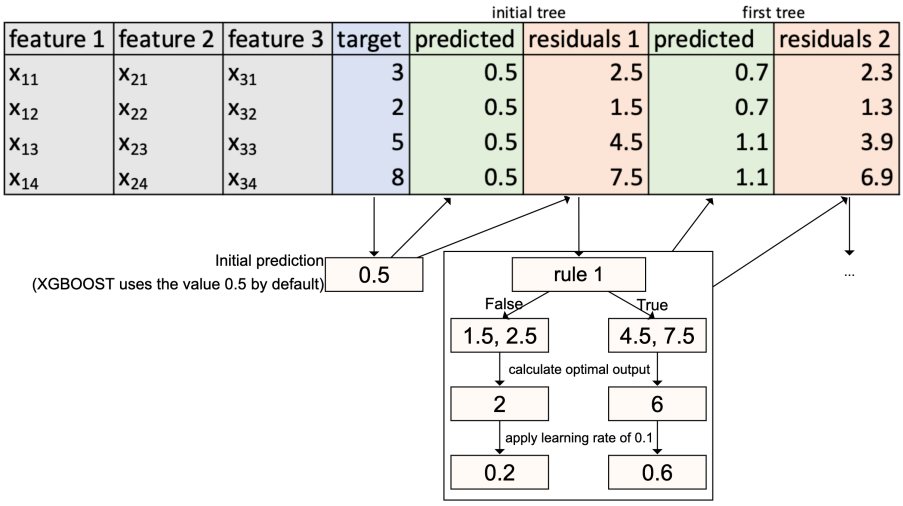


Figure 5.2: Simplified training of a gradient boosted regression tree model with XGBOOST: The initial tree is one node with the default value 0.5. In the next step, residuals from this single node tree are modelled using a weak learner (a decision tree with reduced depth). Rules, like rule 1, are determined by the split of the feature space that produces a tree with the highest 'gain'. There are optimizations in the case of larger datasets, but in general trees (and thereby rules) are build for all possible splits (not shown here). The output values of the leaves depend on the chosen loss function. With the loss function $\frac{1}{2}(target - prediction)^2$ the optimal output for leaves turns out to be the mean of its values. To prevent overfitting, a learning rate of 0.1 is applied when a tree gets added to the model.

target' or 'target variable'. In detail, features included all interatomic distances of α-carbons and distances of centres of mass of side chains of residues in the binding site not more than 3.9 Å apart (excluding hydrogens) as well as ϕ, ψ and χ_1 angles of binding site residues. An interatomic distance threshold of 3.9 Å should comfortably take account of most atomic interactions, especially hydrogen bonds which are expected to not exceed 3.1 Å and included all residues defined as part of the binding site as specified in the experimental work of Spiegel *et al.* [61]. Further, the β-hairpin torsion angle as discussed in the chapter 1.3 and chapter 4 was included.

ΔG values were calculated on a frame-wise basis for an extended simulation (1,1 µs) of the structure of the antibody BO2C11 bound to the wild-type FVIII C2-domain. The calculation of these ΔG values was done using MM/GBSA but was much less rigorous than the one employed for binding free energies as in chapter

2. Since each ΔG value that is used for training of the tree model here is based on a single frame (in chapter 2 150 frames were used for the calculation of the binding free energy) no dynamic effects, like differing side-chain conformations, are reflected in these ΔG values. As in chapter 2 conformational entropy is not considered in the calculation of ΔG values. Because of the lack of dynamics and conformational entropy, these ΔG values should not be understood as measurements that should be compared to experimentally determined binding free energy values. Still, if a machine learning model could be trained to predict these less rigorous ΔG values from distances and angles of binding site residues a successive analysis of the machine learning model can give information about the importance of distances and angles which is exactly the motivation here.

The training data contained 550,000 lines, each line comprised of 284 feature values and the ΔG values as the target and added up to about 2 GB of data. It is common practice to train machine learning models only on a fraction of the data. A so called validation data set is retained for an estimation of the model performance at each training iteration. This is also important for the task of hyper parameter tuning, which is the process of finding optimal values of model parameters, e.g. maximum tree-depth. To assess the final performance across tuned models yet another separate portion, the so called test data set is retained. According to this, the data was randomly shuffled and split into a training data set of 70%, a validation data set of 15% and a test data set of 15%. A k-fold cross validation has not been implemented in the course of this work because an exploration of the suitability of analysing gradient boosted trees to interpret MD trajectories is not primarily dependent on the robustness of the model.

The data sets were uploaded to the cloud computing provider Amazon Web Services (AWS) [193] and an XGBOOST model was trained as well as hyper parameters fine-tuned using the built-in algorithm and functionality of the AWS machine learning service SageMaker. Using a 16-core server, the hyper parameter tuning and training of the model took less than 10 hours. Binding free energy was predicted with a root-mean squared error of 5 kcal/mol on test data set. Since build-

ing a predictive model that could be used with other structures is not the aim here, the model is build merely for the analysis of the model itself, overfitting or loss of generality is of no concern. Given the range of calculated binding free energy values from around -56 kcal/mol up to -1 kcal/mol, an error of 5 kcal/mol represents an error of 10% and demonstrated that the selected features can be used to predict binding free energy with a reasonable degree of accuracy.

5.1.2 Understanding feature impact with Shapley Values

Lundberg *et al.* introduced the use of SHAP (SHapley Additive exPlanations) values to machine learning as a unified measure to correctly and consistently interpret predictions of models [194]. XGBOOST in combination with the Python package SHAP is now widely employed to analyse and raise trust in trained models in many different fields, including medical research [195], chemistry [196] and road traffic [197] to name a few. The SHAP library implemented in Python makes it easy to calculate Shapley values for a given model and is also equipped with expressive visualization techniques. In figure 5.4 features are ordered by their impact on model predictions alongside the nominal impact. For a more detailed view that incorporates the distribution of feature values the visualization technique of figure 5.5 can be used. Every feature value is represented by a dot where clusters of dots pinpoint areas of great similarity. The ordering is the same as in figure 5.4. Colour indicates whether a feature value, in our case an interatomic or side chain distance, a ϕ, ψ or a χ_1 angle, is high or low. Other visualization techniques are implemented by the Python package SHAP that break down an investigated prediction into feature contributions or display the correlation between two features.

The calculation of a Shapley value for a feature j is as follows:

$$\varphi_i(\upsilon) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [\upsilon(S \cup \{i\}) - \upsilon(S)] \tag{5.1}$$

where N is a set of features, n the total number of features and υ is the value or 'worth' of the machine learning model. S is a set of permutations of set N where feature i is excluded. The marginal contribution of i is then calculated by subtracting

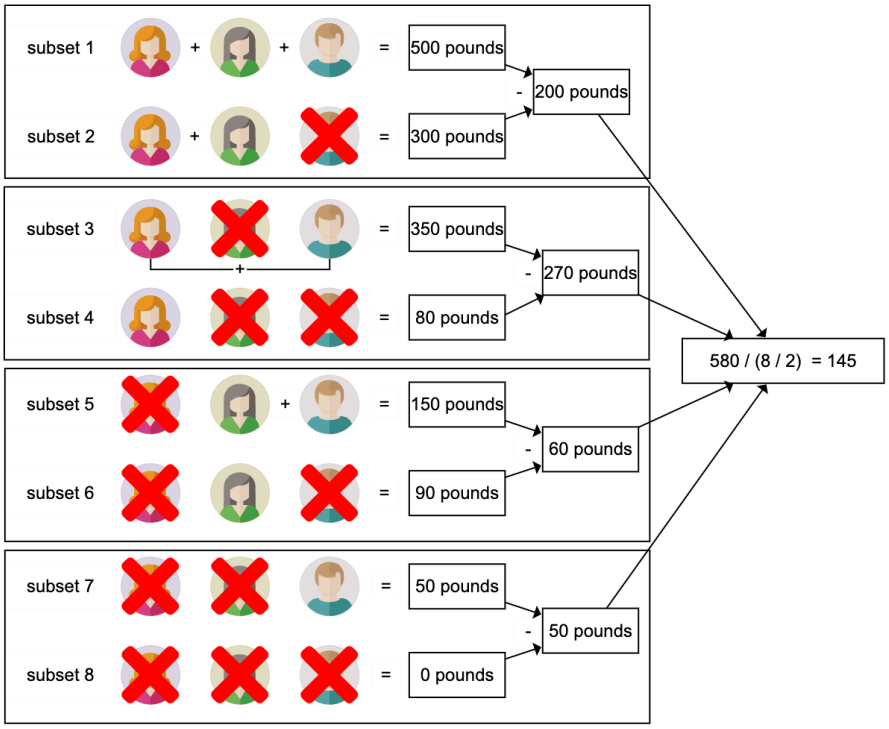


Figure 5.3: Intuitive calculation of the Shapley value: Three contributors (or features) have gained a payout of 500 pounds. To find out about the contribution of the third person, the outputs of all possible subsets of the set of contributors with and without the third person are subtracted. These are then summed and divided by $\frac{number\ of\ subsets}{2}$ which results in the Shapley value of 145 in this case.

the 'worth' of S with and without i $(\upsilon(S \cup \{i\}) - \upsilon(S))$. The term $\frac{|S|!(n-|S|-1)!}{n!}$ scales the contribution relative to the number of feature permutations. An intuition how the Shapley value is calculated is given in figure 5.3 (permutations are not considered).

Generally, the order in which a model reads in the features can make a difference in the prediction and the Shapley value is therefore calculated using all permutations of features to remove bias. Since going through all possible combinations quickly gets computationally intractable the SHAP library makes use of the hierarchical structure of tree based models to speed up the calculation.

5.2 Discussion of the two most impactful distances

Using the techniques discussed above, the most influential feature was determined as the distance of the centre of mass of the arginine 2215 side chain on the C2-

domain to the side chain of aspartic acid 32 on the antibody heavy chain. This finding is somewhat in line with results from binding free energy calculations from chapter 2 where the artificial increase of this distance, by a reduction of arginine 2215 to alanine, resulted in the strongest change in binding free energy by far.

An interesting finding is that the side chain distance between the C2-domain residue 2253 which is a threonine and the aspartic acid 97 on the antibody heavy chain is the second most influential feature. This contradicts the prediction for the substitution T2253A to a certain extend which was examined in the course of chapter 2. The alteration hardly made any difference to the binding free energy or to the experimentally determined binding affinity as can be seen in figure 2.7 (a). The take-away message from section 2.3 was that the energetic loss due to the substitution T2253A at the site itself is compensated by stronger bonds at other locations. This finding is once more presented in table 5.1.

From figure 5.5 it can further be seen that the side chain distance T2253 to D97 increases binding free energy by up to 8 kcal/mol in some frames (the red line goes up to about 8 kcal/mol). The red coloring indicates that this happens when the distance is high.

Another point mutation at the location T2253 which has been reported on the CDC Hemophilia Mutation Project data base is the one to proline [198]. The severity of this mutation has been classified as mild haemophilia A. A mapping of the FVIII C2-domain epitope including the residue T2253 by Pellequer and co-workers [199] points out that the mutation to proline most probably causes a conformational change that results in a modified FVIII activity. Further, it has been found that a mutation to alanine at this location did not cause major changes in activity. Pellequer and co-workers attribute this finding to the insufficiency of alanine to represent charged and structurally constraining amino acids like proline which has also been shown in other cases of their investigation.

A CMD simulation of T2253P conducted here using the same configuration as in chapter 2 produced binding free energy values that were indeed about 8 kcal/mol higher than for the wild-type and T2253A simulations. A pairwise decomposition

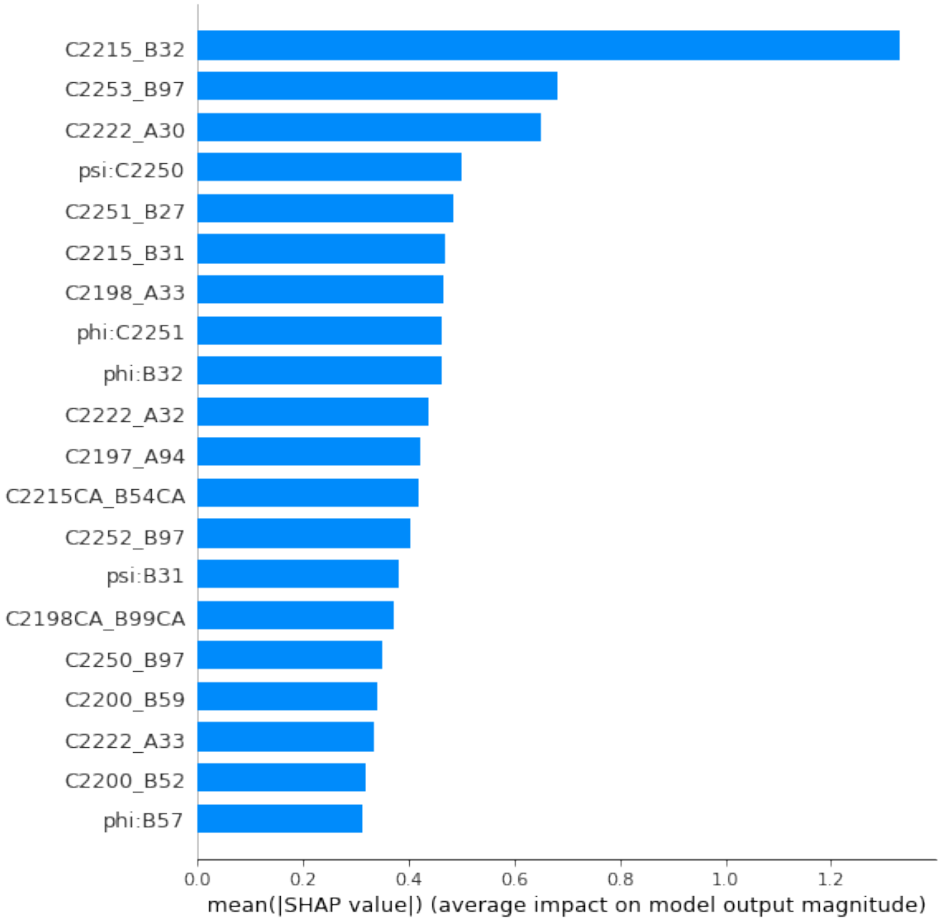


Figure 5.4: Plot of feature importance and averaged impact as by the Python package SHAP: Here the most important feature is C2215_B32 which is the distance of the side chain of the C2-domain (chain identifier C) residue 2215 to the antibody heavy chain (chain identifier B) residue 32. In average this feature influences the model outcome by 1.2 kcal/mol

of binding free energies and a comparison with the substitution to alanine showed that the substitution to proline has a stronger effect at the site itself (2.41 kcal/mol) and moreover influences neighbouring residues L2251 and L2252 that sit at the tip of a β-hairpin important for binding with increases in pairwise energy by 1.06 kcal/mol and 3.04 kcal/mol respectively (figure 5.2). These major differences between T2253P and T2253A mainly explain the difference in calculated binding free energy and possibly experimental binding free energy.

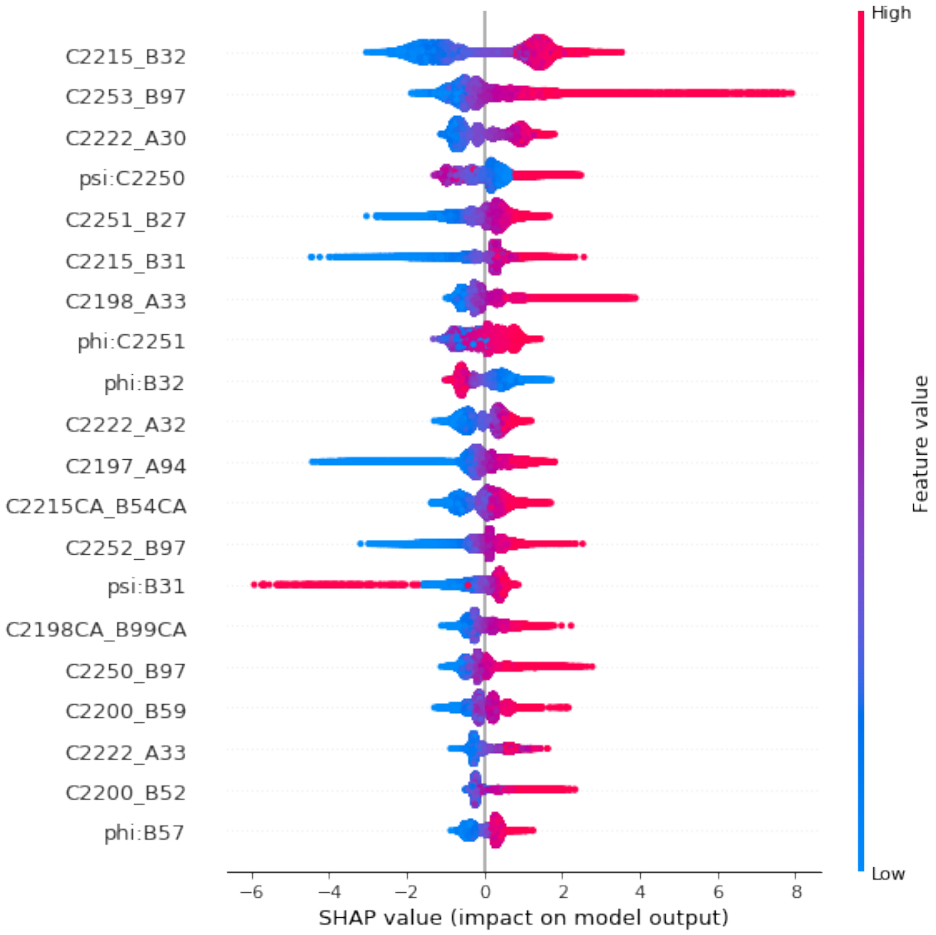


Figure 5.5: Plot of feature value and impact as by the Python package SHAP. As in figure 5.4 the most important feature is C2215_B32, which is the distance of centres of mass between the side chains of the residue R2215 (C2215, where C is the chain identifier that specifies that this residue is in the C2-domain) and the aspartic acid 32 (chain identifier B assign this residue to the antibody heavy chain). Most of the distance values can be found in two clusters. The one in blue indicates that a low distance has a stabilizing effect on binding of about 2 kcal/mol whereas distances in the red cluster are high and result in the opposite.

	T2253A	WT	T2253A - WT
F2196	-3.67	-3.18	-0.49
T2197	-4.49	-4.83	0.34
N2198	-7.33	-7.03	-0.3
M2199	-8.19	-7.58	-0.61
F2200	-7.92	-8.61	0.69
W2203	-1.08	<1	-1.08
G2214	-1.58	<1	-1.58
R2215	-25.16	-23.89	-1.27
R2220	-16.27	-15.81	-0.46
Q2222	-6.16	-6.29	0.13
V2223	-1.38	-1.32	-0.06
S2250	-10.21	-9.93	-0.28
L2251	-9.64	-10.53	0.89
L2252	-7.45	-8.08	0.63
T2253(A)	-4.6	-12.14	7.54
M2255	-1.04	-1.05	0.01
H2315	-1.32	<1	-1.32
Q2316	-3.86	-3.21	-0.65

Table 5.1: Gain in binding free energy of T2253A compensated by stronger bonds at other locations: All values in kcal/mol; First column: interaction energies of binding site residues of T2253A contributing more then 1 kcal/mol; second column: interaction energies of these residues in the wild-type; third column: difference between the first and second column; green line: the substitution T2253A increases binding free energy by around 8 kcal/mol compared to the wild-type; red lines: interactions at these locations strengthen as a cause of the substitution and diminish the effect of the point substitution T2253A.

5.3 Conclusion

An interpretable machine learning model was trained using the implementation of gradient boosted decision trees XGBOOST employing the cloud service provider Amazon Web Service. The data used for training was an extended simulation (1.1 µs in total) of the wild-type structure of the C2-domain in complex with the antibody BO2C11.

The model predicted binding free energies from distances and angles of epitope residues with a root mean squared error of 5 kcal/mol which translates to an error of just under 10%. This is far from the optimal error value of 1 kcal/mol which would be an exact prediction given thermodynamic noise but was enough in this scenario to highlight interesting interactions in the binding site.

F2196	-0.60
T2197	0.60
N2198	0.48
M2199	1.00
F2200	-0.42
W2203	1.08
G2214	0.33
R2215	-1.50
R2220	-0.06
Q2222	-1.67
V2223	-1.22
N2224	-1.60
S2250	0.31
L2251	1.06
L2252	3.04
T2253(P/A)	2.41
M2255	1.04
H2315	-0.78
Q2316	0.56

Table 5.2: Difference of pairwise interaction of substitutions T2253A and T2253P; green lines: T2253P produces a 2.41 kcal/mol higher pairwise energy than T2253A at the site of the substitution and further increases values for neighbouring residues L2251 and L2252 by 1.06 kcal/mol and 3.04 kcal/mol respectively over T2253A.

As was expected from previously conducted simulations of the substitution R2215A, distances and angles of this residue determined the binding free energy to a high degree. It further showed that the distance between this residue and residue 32 on the antibody heavy chain has a non-linear relationship and mainly populates two states (figure 5.5).

A potentially important location for binding affinity that was not highlighted in the alanine scan conducted by Lin and co-workers [44] is residue T2253. The artificial increase in side chain distance between T2253 to D97 by the reduction of threonine to alanine did also not produce a different binding free energy than the one of the wild type when calculated employing MM/GBSA (see section 2.3). Replacements to alanine did not indicate that the side chain distance to residue 97 on the antibody heavy chain is one of the major determinants for binding free energy as was proposed by the analysis of the trained tree model (figure 5.4).

A computational investigation using CMD of yet another substitution at this location T2253P was found to have a pronounced effect on binding free energy. T2253P has not been evaluated experimentally but is identified as a cause of mild

haemophilia A within the CDC Hemophilia Mutation Project data base [198]. Proline forms a pyrrolidine loop connecting the α-carbon with the main chain nitrogen atom and substitutions with proline are typically used to rigidify flexible regions [200]. The substitution T2253P may change the conformation and/or flexibility of the β-hairpin so that it interacts in a different way with phospholipids, hence it gives rise to mild haemophilia A. Proline substitutions often have a profound effect on function and may not be substitutions of choice in designing functioning FVIII molecules. However, the loss in function may also reduce the affinity to antibodies like BO2C11 that overlap with the phospholipid epitope, containing the β-hairpin with residue T2253 [63], which has been suggested by the results presented in this chapter.

Popular clustering algorithms like k-means [179] or DBSCAN [180] have not been evaluated against the machine learning model used in this work. Because of the apparent drawbacks of clustering outlined in the introduction of this chapter, it would only suboptimally support structural investigations of MD trajectories. Same can be said about analyses using correlation coefficients.

The undertaking of a PCA analysis is rather different from the latter approaches. Its data is a set of simulations where amino acid substitutions have already been introduced. Each of these simulations is then analysed with PCA to draw conclusions about how conformational changes induced by amino acid substitutions influence binding free energy by dynamics. Computational methods to compare principal components exist but are not routinely used nor easy to implement [201]. A manual inspection and comparison of components by eye could potentially highlight important conformational changes. Still, such an undertaking would not be practical for a set of 18 simulations. Another approach would be to calculate components of a global covariance matrix consisting of the combined trajectories of all 18 simulations. This would ease the process of finding meaningful conformational changes. Identified changes could then be correlated with calculated binding free energies.

The subjective experience in this work was that the training of an XGBOOST

decision tree and the interpretation of it using the Python package SHAP was straightforward (for a comparison see table 5.3). The fine-tuning of parameters of the decision tree has been automated by utilizing built-in functionality of Amazon Web Services.

It can be argued that techniques that use a user defined set of features, which has been the case for training the of the XGBOOST decision tree, are biased by the expectations and preferences of the experimenter. However, including all features would mean including all possible interatomic distances and angles found in a protein structure. For a complex containing approximately 10,000 atoms this would not be feasible. The issue can be mitigated by using default thresholds; nevertheless, important interactions, especially those that arise after large scale conformational changes, might be missed. PCA does not share this shortcoming since it can be set up to capture all relevant motion. Still, a subsequent selection of important motion involves the experimenter and introduces some subjectivity to the analysis.

Concerning training data, reproducibility has not been investigated in this work by, for example, using a repeated simulation run and/or using fractions of simulated frames. It would be interesting to quantify the variability of predicted outcomes or in other words if two relatively lengthy MD simulations could diverge to a degree that influences the predictions of the machine learning model.

Except for clustering, each of the techniques highlighted in this chapter provides a built-in measurement of accuracy which conveys a sense of expressiveness and trustworthiness of conclusions. Calculations of correlation coefficients are typically accompanied by p-values and each calculated PCA component has a magnitude that quantifies how much of overall motion it represents. Machine learning approaches try to build a model that optimizes the prediction of a target variable and report the accuracy of the model. Different quantities and methods exist for clustering and a whole field of research is dedicated to this topic [202]. A familiarization with the subject of cluster validation was beyond the scope of this chapter and would most probably be for most researchers that want to use MD simulations as an inspiration for experimental work.

The biggest disadvantage of XGBOOST decisions trees with an analysis employing Shapley values is that the data, as it has been used for the training of the machine learning model here, cannot be compared to experimental results. The confidence in conclusions of the technique is dependent on the trustworthiness of underlying MD simulations whose accuracy is not an assumption that can be acted on. An evaluation is further hindered by the fact that ΔG values as calculated by MM/GBSA have many shortcomings, even more so, when only a single frame is considered as has been the case in the generation of data for training the tree model. Also, MM/GBSA calculations are suited for the calculation of relative binding free energies but not of absolute binding free energies, as has been discussed in chapter 2. Since the data used here is a single prolonged simulation of the wild-type structure and ΔG values are calculated by MM/GBSA there is no approach for a comparison of these to experimentally determined binding affinities. A set of simulations of structures containing amino acid substitutions could be run to obtain relative binding free energies which could then be compared to experimental binding affinities as has been done in chapter 2. This is how trust in MD simulations was established in the first place, which verified its use for the training of a machine learning model. However, such an evaluation decreases the usefulness of the machine learning approach because detailed insight could already be gained by investigating the effect of amino acid substitution in existing MD simulations.

Another approach for evaluating the data and the machine learning approach could be by comparison of the ranking of the importance of residues as in figure 5.4 with a ranking of binding affinity of amino acid changes as determined in experiments. However, as has been shown in the case of T2253, a reduction to alanine resulted in a low ranking of the residue, in a sense that it has a weak impact on binding affinity in experiments, whereas an interaction of this residue was ranked the second most influential feature in the XGBOOST decision tree.

Because of issues with evaluating the data, the use of a trained interpretable machine learning model to investigate binding free energies is not optimal. A better use case for interpretable machine learning approaches could be the investigation

	XGBOOST with SHAP	Correlation Coefficients	Clustering	PCA*
documentation and project execution	8	7	2	6
practicable on commodity hardware	8	10	3	9
visualization of results	10	2	2	8
atomic resolution	✓	✓	✓	-
linear relationships	✓	✓	-	✓
non-linear relationships	✓	-	✓	-
built-in measurement of accuracy	✓	✓	-	✓
comparison to experimental results	-	-	-	✓
non-user specific selection of data	-	-	-	✓

Table 5.3: Strength and weaknesses of analysis techniques of MD simulations with added binding free energy calculations: This table represents a subjective perception of data science projects where the different methods have been applied. In the case of correlation coefficients and clustering, results are typically visualized with an additional Python package or any other data visualization software which distinguishes these techniques from the others, that provide built-in visualization capabilities. Computational resources vary to a fair degree with chosen clustering algorithm and it is tough to pin down a number here. A PCA analysis using the package AMBERTOOLS18 [88] was experienced as more difficult to set up than XGBOOST or correlation coefficients. The choice of an appropriate clustering algorithm and its optimization make such projects more complex.

of an allosteric pathway that is linked to protein activity. An example is the protein HDAC8 whose 'in' and 'out' states determine its activity and that are defined by the side chain orientation of its residue Y100. Different disease causing mutations remote from this residue have been shown to influence the occurrence of 'in' and 'out' states in experiments and in MD simulations [75]. With an interpretable machine learning model, like the one discussed here, the nature of the influence of 'in' and 'out' state impacting residues might be better understood from MD simulations and might give rise to new targeted drugs to treat certain forms of cancer that are caused by altered HDAC8 activity.

Yet another use case that won't need a validation would be as an inspiration for NMR experiments. Different states of side chains can easily be spotted by the analysis presented here and by further quantifying these states by e.g. hydrogen bonds these could inform NMR experiments focussing on these states. If the findings of the machine learning model are confirmed in NMR experiments this could possibly lead to sophisticated and targeted alterations of the structure to promote one over another state.

Chapter 6

FVIII C2-domain Expression and Purification

In previous chapters it was proposed that non-binding mutations R2220A and R2220Q introduce conformational changes in the FVIII C2-domain epitope to antibody BO2C11 and thus affect the binding. However, molecular dynamics simulations carried out in this work did not show large scale conformational changes upon introduction of these mutations. This might be due to steric hindrance of epitope motion in the case of the holo FVIII C2-domain or insufficient sampling time in the case of the apo FVIII C2-domain structure.

With the aim of investigating conformational changes of non-binding mutations R2220A and R2220Q in the FVIII C2-domain epitope to BO2C11, an attempt was made to establish a protocol for wild-type FVIII C2-domain purification with the goal to conduct nuclear magnetic resonance spectroscopy (NMR) experiments. NMR was chosen over other structural biology methods like X-ray crystallography and cryo-electron microscopy because of its ability to capture protein dynamics in solution [203]. It would also be possible to compare dynamic behaviour captured in NMR experiments to computationally created trajectories. A disadvantage of NMR is the lengthy, labour intensive spectral post-processing of data. A big part of that is the protein backbone assignment, where spin systems are identified and linked to the protein sequence. Thankfully, Nuzzio et al. report ^{1}H, ^{13}C, and ^{15}N backbone chemical shift assignments for the FVIII C2-domain under investigation here [204].

After the wild-type FVIII C2-domain has been purified and NMR experiments have been conducted, it was envisaged to repeat the protocol with mutation R2220A introduced to the FVIII C2-domain. By comparing the NMR spectra of the wild-type FVIII C2-domain and R2220A FVIII C2-domain conformational changes (if any) should become evident. In any case, this would narrow down the explanations for the phenomenon of non-binding associated with mutations to R2220.

6.1 Attempts using plasmid pET-32b(+)

As a first attempt, the FVIII C2-domain sequence was introduced into a pET-32b(+) vector designed for expression in Escherichia coli bacteria (*E. coli*). *E. coli* have also been used in all the following purification attempts because of its ability to digest labelled glucose, fast growth, good protein expression and easy handling as well as costs and availability of growth and purification media. Further, the use of *E. coli* is preferred because ^{13}C-glucose and ^{15}N-ammonium chloride can be used as the sole carbon and nitrogen source, thereby leading to uniformly ^{13}C, ^{15}N isotropically labelled protein. The pET-32b(+) vector which was ordered from GenScript Biotech contains besides the sequence of the FVIII C2-domain a thioredoxin (TrxA) tag alongside a 49 residue linker region containing a tag of six histidines (His_6) and an enterokinase cleavage site. Thioredoxin has the ability to promote disulfide bridges which may be advantageous in forming the link between the thiol groups of the FVIII C2-domain cysteines 2,174 and 2,326, which connect the the N- and C-terminal regions [60]. Further, the correct formation of this disulfide bond was promoted by conducting, in parallel, expression using both SHuffle cells (NEB) in addition to expression with Novagen BL21 (DE3) *E. coli* cells. SHuffle cells have been developed to enhance the correct folding of proteins and promote the formation of disulfide bonds [205]. An enterokinase cleavage site was chosen over other options like FXa or thrombin sites because of its highly specific Asp-Asp-Asp-Asp-Lys cleavage site. A linker region between the cleavage site and the FVIII C2-domain was introduced comprising the last three residues of the neighbouring C1-domain. The His_6 tag was used in Ni-NTA chromatography to extract

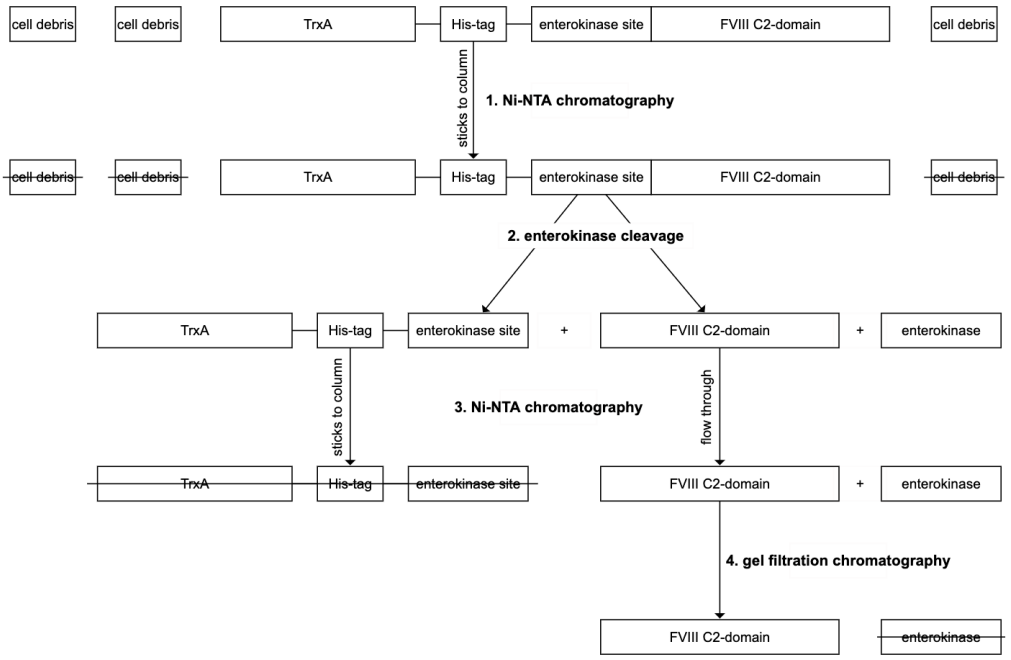


Figure 6.1: Purification of the FVIII C2-domain from the pet32b(+) construct: 1. The cell lysate of *E. coli* is ran over a nickel column (Ni-NTA chromatography) where uncleaved protein would stick to the column because of the high affinity of the His_6 tag; 2. The protein gets eluded from the column (not shown here) and is then cleaved by enterokinase, separating the section containing thioredoxin, His_6, and the enterokinase site from the FVIII C2-domain part; 3. In a second Ni-NTA chromatography the former part, containing the His_6 tag sticks to the column and the FVIII C2-domain alongside enterokinase is found in the flow through; 4. In a size exclusion remaining impurities, including enterokinase get removed from the sample, leaving pure FVIII C2-domain.

the protein from *E. coli* debris. The concept of the purification protocol is illustrated in figure 6.1.

The plasmid pET-32b(+) and Novablue cells were thawed at 4°C. 1μl DNA was added to 50μl cells, put on ice for 30 minutes and heat-shocked at 42°C for 45 seconds with a subsequent 10 minutes on ice. This process ensures that DNA is directly taken up by the Novablue cells which are genetically modified in that manner. Novablue cells were then put in 750μl of autoclaved 25 mM LB medium and incubated for 30-40 minutes at 37°C. 20μl of grown Novablue cells were put on a Petri dish prepared with autoclaved LB media containing 2% agar and 1 mM ampicillin, which was added after the media cooled down to skin temperature, and grown

overnight in an incubator at 37°C. One colony was mounted on a pipette tip and put into 1 ml autoclaved 25mM LB medium and incubated at 37°C overnight. The DNA was extracted from this culture using the protocol of Invitrogen™ PureLink™ Quick Plasmid MiniPrep-Kit [206]. A DNA sample was sent off for sequencing and results confirmed the sequence of the ordered plasmid.

The confirmed DNA was then transformed into BL21(DE3) *E. coli* cells following the same steps as above up to the incubation of the 1 ml culture. This time, a single colony was put into 5 ml LB medium containing 1 mM ampicillin and incubated over the day at 37°C, shaking at 200 rpm. 1 ml of this culture got added to 100 ml M9 medium which contained 75 mM PO4, 8.5 mM NaCl, 0.1 ml of 1 M MgSO4, 0.1 ml of 0.1 M CaCl2, 0.1 ml of Micronutrient solution 1, 0.1 ml of Thymine/Biotin solution, 0.1 ml of ampicillin, 1g glucose, 0.1 ^{1}5N and the pH was adjusted to 7.4. A pH of 7.4 was chosen to resemble conditions of the FVIII protein environment which is within human blood that spans a pH of 7.36 to 7.44. The sample was then incubated overnight at 37°C, shaking at 200 rpm. The Thymine/Biotin solution was taken from a prepared stock containing 0.1 g of each component in 100 ml distilled water. The next day, two 1 l M9 cultures were prepared with the composition as the 100 ml culture with 45 ml of the 100 ml overnight culture got added to each liter. The two cultures were grown at 37°C, shaking at 200 rpm over the day. IPTG was added and the temperature reduced to 18°C when each culture reached an optical density of 0.65A and left overnight for expression of the protein. The next day, the 2 liter cultures as well as all following samples were put on ice. The 2 liter cultures were transferred into centrifuge beakers and spun down at 4000 rpm at 4°C for 20 minutes. Resulting pellets were scarped into a Falcon tube and 40 ml lysis buffer was added. The lysis buffer consisted of 50 ml M9 base buffer, two protease tablets, 1 small spatula tip DNAse as well as lysosyme and 150μl 3 mM $MgCl_2$. The sample was then sonicated 4 times for 30 seconds with 30 second intervals. The lysed sample was distributed in two centrifuge flasks and spun down for one hour at 18,000 rpm at 4°C. Subsequent, the supernatant was loaded onto a Ni-NTA column that got eluded using an AKTA system. Wells exhibiting a high

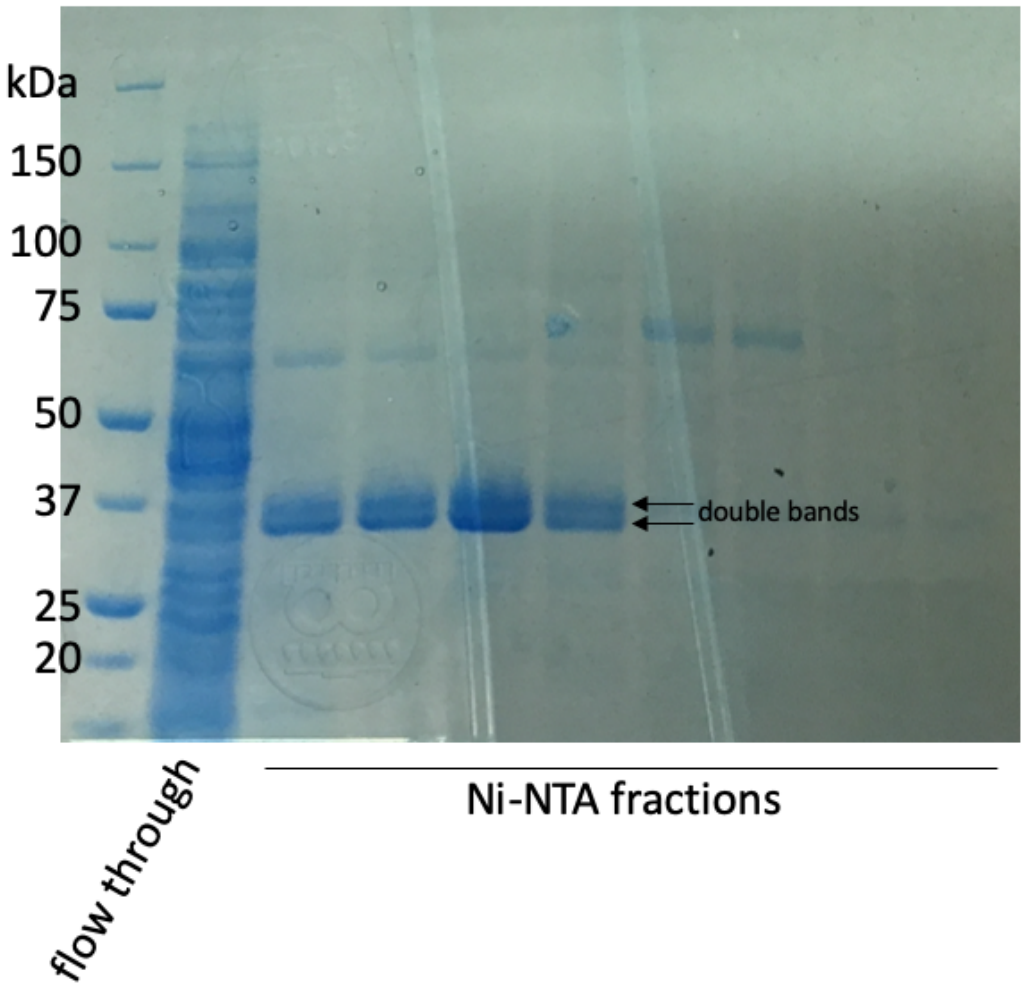


Figure 6.2: Gel of Ni-NTA elution: It showed that protein accumulated just below 37 kDa, which is the size of the FVIII C2-domain that is interlinked with the tags that have been introduced due to purification reasons, the biggest part being thioredoxin. It is not clear why double bands appear at this position.

UV (280nM) value were analysed using reducing SDS-PAGE. As can be seen in figure 6.2 the gel showed large quantities of protein just below 37 kDa which is in line with a theoretical molecular weight value of around 35 kDa. The reason for an occurrence of a double band that showed on the gel is not clear. However, if one band is an impurity this would be accounted for in the cleavage since only the FVIII C2-domain construct gets cleaved by enterokinase and could be discriminated by its reduction in size from 35 kDa to around 17 kDa.

Wells exhibiting protein of 35 kDa size were then transferred to a 15 kDA membrane for an overnight dialysis in a beaker containing two liter of dialysis buffer

with 100 mM NaCl and 20 mM PO4 at pH 7.4 and 4°C. The amount of protein in the dialysed sample was then measured using NanoDrop and enterokinase levels calculated accordingly for the process of cleavage. Initially, cleavage was very slow due to a lack of calcium but after this was fixed, cleaved protein precipitated out of solution. For a better solubility a higher salt concentration would be favourable but this hindered protease activity and did not result in the desired effect. Attempts to prevent precipitation by adding different levels of glycerol, of up to 5%, to the dialysis buffer were unsuccessful. Spinning down or filtering the sample resulted in loss of the protein to the extent that there was insufficient for NMR spectroscopy. The purification with plasmid pET-32b(+) never got past this stage. Following steps of the purification envisaged a reverse column, where the cleaved TrxA, His_6 and enterokinase part would stick to the Ni-NTA column and the FVIII C2-domain plus the enterokinase enzyme would be found in the flow through. Lastly, a size exclusion should remove remaining impurities like the enterokinase enzyme which has a molecular weight of 31 kDa over a 17 kDa of the FVIII C2-domain.

Upon cleavage the protein undergoes a drastic change in its isometric point (pI) from 6.21 for the uncleaved state to 8.98 for the FVIII C2-domain and 5.48 for the TrxA, His_6, enterokinase part which might be the cause for the precipitation.

6.2 Attempts using plasmid pET-28a(+)

Because of the high theoretical pI value of 8.98 of the FVIII C2-domain a new purification protocol was developed using a pET-28a(+) plasmid. The construct did not include any tags or thioredoxin and was directed to express the sequence FVIII C2-domain in *E. coli* alone. Since only very few of the other components found in the *E. coli* cytoplasm should have a similar high pI value, the expressed FVIII C2-domain could theoretically be purified directly from lysed cells by cation exchange chromatography. Transformation, expression and harvesting of FVIII C2-domain protein followed the same steps as in the pet32b(+) protocol, except for the use of kanamycin instead of ampicillin since pET-28a(+) has a resistance to this antibiotic.

The first step of purification was a cation exchange chromatography column

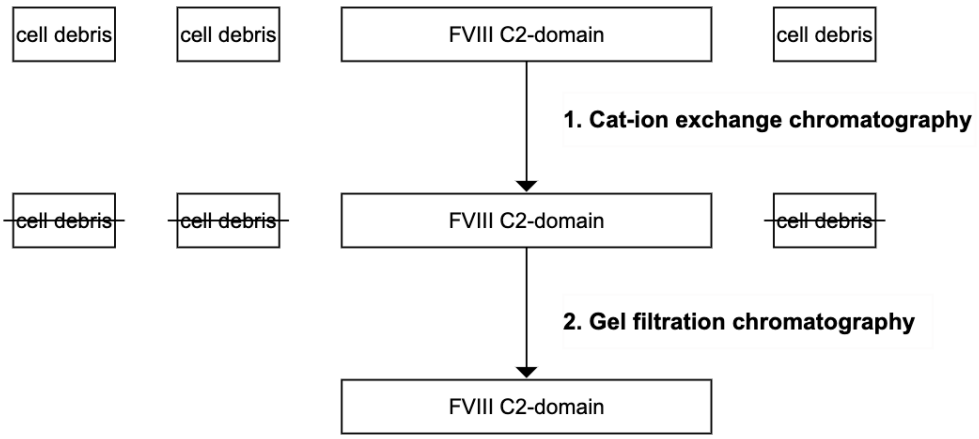


Figure 6.3: Purification of the FVIII C2-domain from the pET-28a(+) construct: 1. After lysing the *E. coli* cells the protein is found in solution along the cell debris. Making use of its high net positive charge FVIII C2-domain is loaded on a cation exchange column; 2. FVIII C2-domain is eluded from the column and remaining impurities are removed by gel filtration chromatography

that captured FVIII C2-domain because of its high net positive charge (high pI). The following elution of the column showed that not only FVIII C2-domain bound to the column. However, since expressed protein was found in wells with relatively little impurities (figure 6.4, wells 1 to 3) a successive size exclusion gel filtration should result in NMR purity grade protein.

A NanoDrop protein concentration measurement resulted in 0.057 mg protein per ml in well 2 of figure 6.4. Well 2 and its neighbouring well were extracted for gel filtration. Unfortunately, the protein concentration was further reduced during the execution of the gel filtration (Figure 6.5). Protein concentration was measured to 0.00124 mM using NanoDrop which is too low for NMR experiments. It is likely that more protein precipitated out of solution. This however would not be visible on the UV curve produced in the gel filtration.

Due to coronavirus shut down of the laboratory it was not possible to refine the protocol further. Refinements envisaged buffer solutions with pH values further from the pI value of the protein to prevent precipitation.

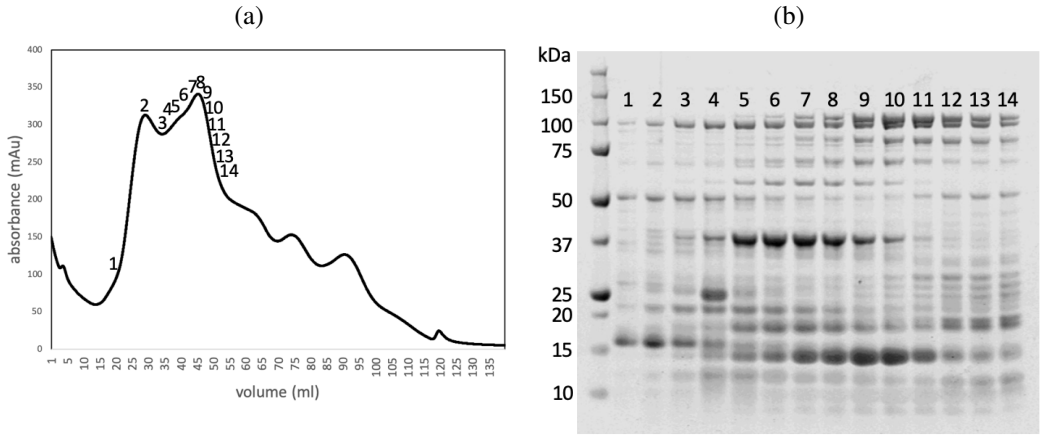


Figure 6.4: Cation exchange chromatography of FVIII C2-domain; (a) trace of lysed *E. coli* cells: It showed that not only FVIII C2-domain stuck to the column which caused the two peaks of well 2 and 8; (b) elution fractions analysed with SDS-PAGE reducing agent: Expressed FVIII C2-domain having a molecular weight of around 17 kDa is most probably found around well 2. Remaining impurities in these wells should well be removed in a size exclusion chromatography using gel filtration.

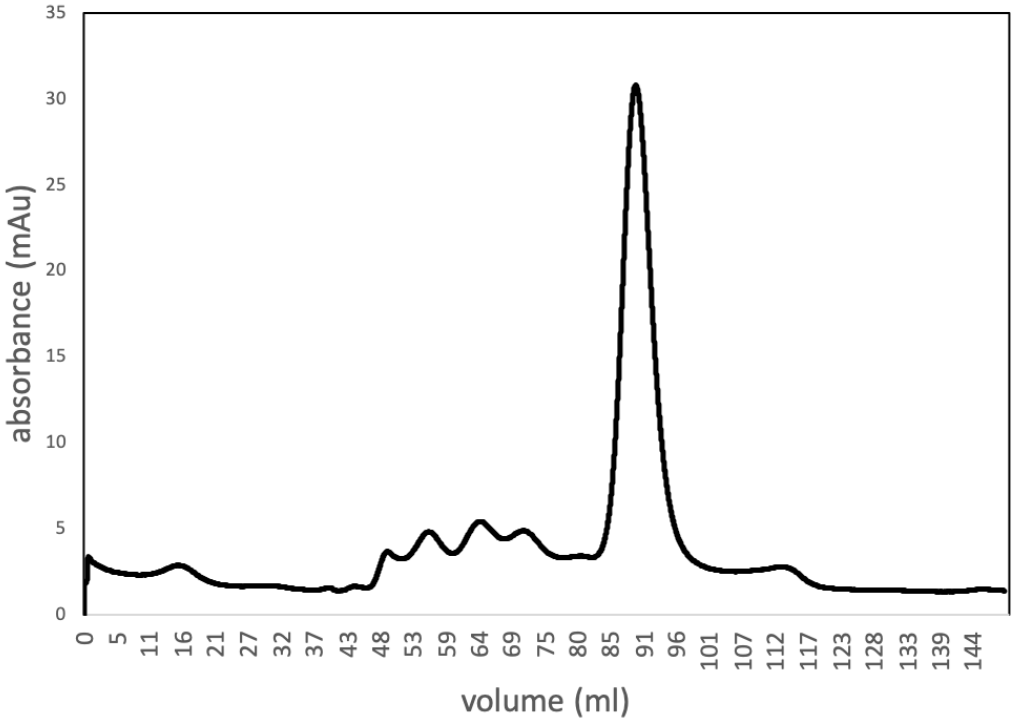


Figure 6.5: Gel filtration: The peak around 90 ml might be FVIII C2-domain protein but the resulting protein concentration of 1.24 μM was too low to conduct NMR experiments.

Chapter 7

Concluding remarks

The molecular structure of biomolecules in conjunction with their dynamics represent a major contributor to their function [207]. The formation of a complex is often initiated and/or followed by a change of protein conformation and accompanied by a change in function. Yet, experimental methods to investigate these phenomena are not routinely accessible [208]. Although experimental methods undoubtedly have advantages and are considered the 'Gold standard', they are typically expensive and labour intensive, and often involve error-prone preparatory steps such as protein expression and purification where unexpected issues my arise. As one small piece of evidence to support the latter point, the experimental work conducted for this thesis and discussed in chapter 6 required the adaptation of a purification protocol that led to numerous, time-consuming downstream issues.

Computational methods such as MD simulations represent supplementary approaches for investigating protein dynamics that do not share most of the shortcomings associated with experimental work. MD simulations afford a means of probing the energetic landscape that is shaped by preferred conformational states and transitions between these states [209]. Expressive simulations can be run on commodity hardware with an initial outlay of below £2000 with negligible running costs. The setup and analysis of simulations can be highly automated. When issues arise, comprehensive log files mean that the information for resolving them is potentially available, although the sheer volume of data becomes a challenge in its own right. (This contrasts with experimental work, where the problem is often the

paucity of information.)

In this thesis, it has been shown that MD simulations combined with MM/GBSA binding free energy calculations can produce good correlations with experimentally-determined binding affinities. Five of the six residues considered to be major components of the functional epitope of antibody BO2C11 were identified, with experimental evidence for the importance of the sixth residue that was "missed" (N2198) being comparatively weak [44] (as discussed in section 2.2).

To investigate the different conformations of a β-hairpin that is part of the binding site, both AMD and US simulations were conducted. AMD simulations are designed to quickly overcome energetic barriers in the energy landscape and therefore can provide a more complete picture of conformational states. The experience in the context of the apo FVIII C2-domain was that AMD simulations are hard to equilibrate and that a sensible reconstruction of the energy landscape for atomic structures the size of the FVIII C2-domain cannot be achieved, in line with previous findings [170].

US simulations modify a predefined reaction coordinate and can be used to calculate the potential of mean force along the coordinate. In this research, US indicated that the twisting of the β-hairpin M2199/F2200 that is part of the binding site of the holo FVIII C2-domain is associated with similar potential of mean force curves for both the apo and antibody-removed holo FVIII C2-domain structure. More interestingly, simulations suggested that the conformation of the β-hairpin M2199/F2200 favours a twist angle that is approximately 10° away from the one measured in the apo FVIII C2-domain crystal-structure.

Upon the introduction of mutations that were described as non-binding by Lin *et al.* [44] the potential of mean force did not differ largely from the one of the wild-type. From the view of US simulations this indicates, that these mutations do not influence the β-hairpin M2199/F2200 to a degree that would explain the abrogation of binding.

However, the challenge of interpreting MD simulations soon becomes insurmountable at the atomic or even at a more coarse grained level. A number of tech-

niques exist that reduce the sheer number of atomic coordinates to a comparatively small number of meaningful quantities. In the case of AMD simulations, conformational states can be highlighted using clustering of RMSD values or via the reconstruction of the energy landscape. US simulations are by design optimized to elucidate the forces arising when changing one or two pre-selected reaction coordinates. To characterise the binding between mutants of the C2-domain of the blood protein FVIII and the human antibody BO2C11, the binding free energy of MD simulations of the complex was calculated using the technique MM/GBSA. Important pairwise interactions in the binding site were then highlighted by a decomposition of the energy values and by focused visualization. Nonetheless, a manual analysis of pairwise interaction energies of residues in the binding site of a set of 17 simulations is labour-intensive and has a tendency to become subjective.

Bearing these challenges in mind, the usefulness of employing a machine learning approach to aid the analysis of simulation data was investigated. It showed that it is feasible both to generate suitable training data and to develop an appropriate machine learning model to address this task. Specifically, a decision tree-based model was used to estimate binding free energies based on interatomic distances, and side chain and backbone angles. The estimates were of reasonable accuracy, and the subsequent analysis of the decision tree highlighted potential combinations of interatomic distances and angles that have a strong impact on binding free energy. These results were promising, but in the absence of experimental validation, they remain speculative.

It was hoped that the necessary validation would be attainable via NMR experiments involving the FVIII C2-domain. However, attempts to express and purify the recombinant apo FVIII C2-domain were unsuccessful; expressed protein precipitated out of solution and the remaining protein concentration proved insufficient for NMR. The next step would have been to adjust the pH of the buffer solutions, but was unfortunately not possible owing to the closure of the lab in response to the corona virus pandemic. This not only curtailed an investigation into the potential usefulness of the new machine learning method, and a means of benchmarking the

MD simulations, there was an additional motivation for conducting NMR experiments: to introduce the non-binding mutation R2220A and observe the impact (if any) on the conformation of the molecule, and thereby gain insights into the causes underlying the abrogation of BO2C11 binding observed in R2220A muteins.

A key emphasis of this work was to provide results in a human interpretable format without losing too much information, i.e. without losing the advantage of the atomic granularity and high temporal resolution of MD simulations. To this end, a novel approach for finding sufficient equilibration time was developed that describes the dynamics of each residue in an area of interest by the Jensen-Shannon distance of its backbone angles. MM/GBSA proved to be a potent method for the ranking of binding free energies but an analysis of pairwise interaction energies of residues in the binding site turned out to be cumbersome for two contrasting reasons. Firstly, there was no visualization tools for comparing the interaction energies of single residues to generate plots like that in figure 2.14. Secondly, to build an understanding of how changes in pairwise interactions influence binding free energy via a comparison of differences in interaction energies of binding site residues in a set of 17 simulations would have been a very labour-intensive endeavour. A reduction of the dimensionality to faciltate human interpretability was then pursued by employing a machine learning approach. Using visualization techniques provided by the Python package SHAP it was shown that important interaction could be identified that might inspire future research [191].

Progress concerning the structure-to-function relationship of the FVIII C2-domain has been made in multiple ways. CMD simulations with MM/GBSA binding free energy calculations successfully ranked experimental binding affinities and could thereby be used to investigate muteins where no experimental binding affinity measurements exist, as was the case with T2253P in this work. AMD simulations proposed that the preferred conformation of the beta-hairpin M2199/F2200 deviates from the one reported in the crystal-structure. However, it was not clear if AMD simulations reached equilibrium, which would have permits a stronger conclusion to be draw. US simulations suggested that the non-binding mutations R2220A and

R2220Q do not influence the β-hairpin M2199/F2200 to a degree that would explain the abrogation of binding. Finally, a machine learning approach highlighted impactful motion in the binding site which could be used to introduce knowledge-drive mutations to the FVIII C2-domain.

Overall, the computational as well as experimental methods used in this work have been found to complement each other, even though the yield of experiments was rather poor. Knowledge spanning both domains makes one appreciate the strengths of each method but also the potential weaknesses. This provides the mindset and tools to tackle ever more challenging questions in the field of structural biology.

Bibliography

[1] Adeel S Yousphi, Ayesha Bakhtiar, Muhammad Arslan Cheema, Syed Nasim, and Waqas Ullah. Acquired Hemophilia A: A Rare but Potentially Fatal Bleeding Disorder. *Cureus*, 11(8), aug 2019.

[2] Teja Thorat, Peter J. Neumann, and James D. Chambers. Hemophilia burden of disease: A systematic review of the cost-utility literature for hemophilia. *Journal of Managed Care and Specialty Pharmacy*, 24(7):632–642, jul 2018.

[3] Jamie O'Hara, David Hughes, Charlotte Camp, Tom Burke, Liz Carroll, and Daniel Anibal Garcia Diego. The cost of severe haemophilia in Europe: the CHESS study. *Orphanet Journal of Rare Diseases*, 12(1):106, may 2017.

[4] Megan M. Ullman and W. K. Hoots. Assessing the costs for clinical care of patients with high-responding factor VIII and IX inhibitors. *Haemophilia*, 12(SUPPL. 6):74–80, dec 2006.

[5] Ronald D. Barr, Mahassen Saleh, William Furlong, John Horsman, Julia Sek, Mohan Pai, and Irwin Walker. Health status and health-related quality of life associated with hemophilia. *American Journal of Hematology*, 71(3):152–160, nov 2002.

[6] Miwa Goto, Hideyuki Takedani, Kazuhiko Yokota, and Nobuhiko Haga. Strategies to encourage physical activity in patients with hemophilia to improve quality of life. *Journal of Blood Medicine*, 7:85–98, may 2016.

[7] Ronald Hoffman, Edward J. Benz, Leslie E. Silberstein, Helen Heslop, Jeffrey I. Weitz, John Anastasi, Mohamed E. Salama, and Syed A. Abutalib. *Hematology : basic principles and practice*. 2018.

[8] Marc Jacquemin, Arne Neyrinck, Maria Iris Hermanns, Renaud Lavend'homme, Filip Rega, Jean-Marie Saint-Remy, Kathelijne Peerlinck, Dirk Van Raemdonck, Charles James Kirkpatrick, J G Gilles, J Vermylen, and J M Saint-Remy. FVIII production by human lung microvascular endothelial cells. *Blood*, 108(2):515–7, jul 2006.

[9] Junliang Pan, Thanh Theresa Dinh, Anusha Rajaraman, Mike Lee, Alexander Scholz, Cathrin J Czupalla, Helena Kiefel, Li Zhu, Lijun Xia, John Morser, Haiyan Jiang, Laura Santambrogio, and Eugene C Butcher. Patterns of expression of factor VIII and von Willebrand factor by endothelial cell subsets in vivo. *Blood*, 128(1):104–9, 2016.

[10] J. Evan Sadler. BIOCHEMISTRY AND GENETICS OF VON WILLEBRAND FACTOR. *Annual Review of Biochemistry*, 67(1):395–424, jun 1998.

[11] Anna Mazurkiewicz-Pisarek, Grazyna Płucienniczak, Tomasz Ciach, and Andrzej Plucienniczak. The factor VIII protein and its function. *Acta Biochimica Polonica*, 63(1):11–16, 2016.

[12] Evgueni L Saenko, Midori Shima, and Andrey G Sarafanov. Role of Activation of the Coagulation Factor VIII in Interaction with vWf, Phospholipid, and Functioning within the Factor Xase Complex. *Trends in Cardiovascular Medicine*, 9(7):185–192, oct 1999.

[13] Peter J Lenting, Jan A. Van Mourik, and Koen Mertens. The life cycle of coagulation factor VIII in view of its structure and function. *Blood*, 92(11):3983–3996, dec 1998.

[14] Peter J. Lenting, Cécile V. Denis, and Olivier D. Christophe. Emicizumab, a bispecific antibody recognizing coagulation factors IX and X: How does it actually compare to factor VIII? *Blood*, 130(23):2463–2468, 2017.

[15] https://commons.wikimedia.org/wiki/File:Classical_blood_coagulation_pathway.png.

[16] Joseph Gish. Structural Studies of Complexes of Blood Coagulation Factor VIII. *WWU Graduate School Collection*, jan 2019.

[17] Massimo Franchini and Pier Mannucci. The History of Hemophilia. *Seminars in Thrombosis and Hemostasis*, 40(05):571–576, jun 2014.

[18] Courtney D. Thornburg. How I approach: Previously untreated patients with severe congenital hemophilia A. *Pediatric Blood & Cancer*, 65(12):e27466, dec 2018.

[19] E. Berntorp, G. Dolan, C. Hay, S. Linari, E. Santagostino, A. Tosetto, G. Castaman, M. T. Álvarez-Román, R. Parra Lopez, J. Oldenburg, T. Albert, U. Scholz, M. Holmström, J. F. Schved, M. Trossaërt, C. Hermans, A. Boban, C. Ludlam, and S. Lethagen. European retrospective study of real-life haemophilia treatment. *Haemophilia*, 23(1), 2017.

[20] A. V. Hoffbrand and P. A. H. Moss. *Hoffbrand's essential haematology*.

[21] NHS England. Clinical Commissioning Policy Proposition: Immune Tolerance Induction (ITI) for haemophilia A (all ages). Technical report, 2016.

[22] Johannes Oldenburg, Johnny N. Mahlangu, Benjamin Kim, Christophe Schmitt, Michael U. Callaghan, Guy Young, Elena Santagostino, Rebecca Krusejarres, Claude Negrier, Craig Kessler, Nancy Valente, Elina Asikanius, Gallia G. Levy, Jerzy Windyga, Midori Shima, Rebecca Kruse-Jarres, Claude Negrier, Craig Kessler, Nancy Valente, Elina Asikanius, Gallia G. Levy, Jerzy Windyga, and Midori Shima. Emicizumab prophylaxis in hemophilia A with inhibitors. *New England Journal of Medicine*, 377(9):809–818, aug 2017.

[23] Amit C. Nathwani, Andrew M. Davidoff, and Edward G.D. Tuddenham. Gene Therapy for Hemophilia. *Hematology/Oncology Clinics of North America*, 31(5):853–868, 2017.

[24] H. Marijke Van Den Berg. A cure for hemophilia within reach. *New England Journal of Medicine*, 377(26):2592–2593, dec 2017.

[25] S.J. Schep, R.E.G. Schutgens, K. Fischer, and M.L. Boes. Review of immune tolerance induction in hemophilia A. *Blood Reviews*, 32(4):326–338, jul 2018.

[26] S. R. Earnshaw, C. N. Graham, C. L. McDade, J. B. Spears, and C. M. Kessler. Factor VIII alloantibody inhibitors: cost analysis of immune tolerance induction vs. prophylaxis and on-demand with bypass treatment. *Haemophilia*, 21(3):310–319, may 2015.

[27] Samantha C. Gouw, Johanna G. van der Bom, Rolf Ljung, Carmen Escuriola, Ana R. Cid, Ségolène Claeyssens-Donadel, Christel van Geet, Gili Kenet, Anne Mäkipernaa, Angelo Claudio Molinari, Wolfgang Muntean, Rainer Kobelt, George Rivard, Elena Santagostino, Angela Thomas, and H. Marijke van den Berg. Factor VIII Products and Inhibitor Development in Severe Hemophilia A. *New England Journal of Medicine*, 368(3):231–239, jan 2013.

[28] Jan Astermark. FVIII inhibitors: Pathogenesis and avoidance. *Blood*, 125(13):2045–2051, mar 2015.

[29] Mathias Behrmann, John Pasi, Jean Marie R. Saint-Remy, Ronald Kotitschke, and Michael Kloft. Von Willebrand factor modulates factor VIII immunogenicity: Comparative study of different factor VIII concentrates in a haemophilia A mouse model. *Thrombosis and Haemostasis*, 88(2):221–229, dec 2002.

[30] Djuro Josić, Andrea Buchacher, Christoph Kannicht, Yow Pin Lim, Klemens Löster, Katharina Pock, Stephen Robinson, Horst Schwinn, and Monika

Stadler. Degradation products of factor VIII which can lead to increased immunogenicity. In *Vox Sanguinis*, volume 77, pages 90–99. Karger Publishers, oct 1999.

[31] Sébastien Lacroix-Desmazes, Ana-Maria Navarrete, Sébastien André, Jagadeesh Bayry, Srinivas V. Kaveri, and Suryasarathi Dasgupta. Dynamics of factor VIII interactions determine its immunologic fate in hemophilia A. *Blood*, 112(2):240–249, jul 2008.

[32] C. L. Kempton and S. L. Meeks. Toward optimal therapy for inhibitors in hemophilia. *Blood*, 124(23):3365–3372, nov 2014.

[33] Christine L Kempton and Gilbert C White. How we treat a hemophilia A patient with a factor VIII inhibitor. *Blood*, 113(1):11–7, jan 2009.

[34] L. A. Valentino, C. L. Kempton, R. Kruse-Jarres, P. Mathew, S. L. Meeks, and U. M. Reiss. US Guidelines for immune tolerance induction in patients with haemophilia a and inhibitors. *Haemophilia*, 21(5):559–567, 2015.

[35] Courtney D. Thornburg and Jonathan Ducore. A novel approach to immune tolerance induction in haemophilia A with factor VIII inhibitor. *Haemophilia*, 25(1):e48–e50, jan 2019.

[36] J F Healey, I M Lubin, and P Lollar. The cDNA and derived amino acid sequence of porcine factor VIII. *Blood*, 88(11):4209–14, dec 1996.

[37] C R Hay, J N Lozier, C A Lee, M Laffan, F Tradati, E Santagostino, N Ciavarella, M Schiavoni, H Fukui, A Yoshioka, J Teitel, P M Mannucci, and C K Kasper. Safety profile of porcine factor VIII and its use as hospital and home-therapy for patients with haemophilia-A and inhibitors: the results of an international survey. *Thrombosis and haemostasis*, 75(1):25–9, jan 1996.

[38] D. Lillicrap, A. Schiviz, C. Apostol, P. Wojciechowski, F. Horling, C. K. Lai, C. Piskernik, W. Hoellriegl, and P. Lollar. Porcine recombinant factor VIII

(Obizur; OBI-1; BAX801): product characteristics and preclinical profile. *Haemophilia*, 22(2):308–317, mar 2016.

[39] Pier Mannuccio Mannucci and Massimo Franchini. Porcine recombinant factor VIII: An additional weapon to handle anti-factor VIII antibodies. *Blood Transfusion*, 15(4):365–368, 2017.

[40] P Lollar, E T Parker, and P J Fay. Coagulant properties of hybrid human/porcine factor VIII molecules. *The Journal of biological chemistry*, 267(33):23652–7, nov 1992.

[41] R T Barrow, J F Healey, D Gailani, D Scandella, and P Lollar. Reduction of the antigenicity of factor VIII toward complex inhibitory antibody plasmas using multiply-substituted hybrid human/porcine factor VIII molecules. *Blood*, 95(2):564–8, jan 2000.

[42] P M Zakas, K Vanijcharoenkarn, R C Markovitz, S L Meeks, and C B Doering. Expanding the ortholog approach for hemophilia treatment complicated by factor VIII inhibitors. *Journal of thrombosis and haemostasis : JTH*, 13(1):72–81, jan 2015.

[43] E. T. Parker, John F Healey, Rachel T Barrow, Heather N Craddock, and Pete Lollar. Reduction of the inhibitory antibody response to human factor VIII in hemophilia A mice by mutagenesis of the A2 domain B-cell epitope. *Blood*, 104(3):704–710, apr 2004.

[44] Jasper C Lin, Ruth A Ettinger, Jason T Schuman, Ai-Hong Zhang, Muhammad Wamiq-Adhami, Phuong-Cac T Nguyen, Shelley M Nakaya-Fletcher, Komal Puranik, Arthur R Thompson, and Kathleen P Pratt. Six amino acid residues in a 1200 Å2 interface mediate binding of factor VIII to an IgG4κ inhibitory antibody. *PloS one*, 10(1):e0116577, 2015.

[45] Kathleen P. Pratt. Engineering less immunogenic and antigenic FVIII proteins. *Cellular immunology*, 301:12, 2016.

[46] B. Nolan, J. Mahlangu, D. Perry, G. Young, R. Liesner, B. Konkle, S. Rangarajan, S. Brown, H. Hanabusa, K. J. Pasi, I. Pabinger, S. Jackson, L. M. Cristiano, X. Li, G. F. Pierce, and G. Allen. Long-term safety and efficacy of recombinant factor VIII Fc fusion protein (rFVIIIFc) in subjects with haemophilia A. *Haemophilia*, 22(1):72–80, jan 2016.

[47] Ingrid Pabinger-Fasching. The story of a unique molecule in hemophilia A: Recombinant single-chain factor VIII. *Thrombosis Research*, 141:S2–S4, may 2016.

[48] S. W. Pipe. Bioengineered molecules for the management of haemophilia: Promise and remaining challenges. *Haemophilia*, 24:68–75, may 2018.

[49] Francis S. Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795, feb 2015.

[50] Michael Makris. Hemophilia gene therapy is effective and safe. *Blood*, 131(9):952–953, mar 2018.

[51] Savita Rangarajan, Liron Walsh, Will Lester, David Perry, Bella Madan, Michael Laffan, Hua Yu, Christian Vettermann, Glenn F. Pierce, Wing Y. Wong, and K. John Pasi. AAV5–Factor VIII Gene Transfer in Severe Hemophilia A. *New England Journal of Medicine*, 377(26):2519–2530, dec 2017.

[52] Shannon L. Meeks, Courtney Cox, and W. Hunter Baldwin. Epitope Mapping of Inhibitor Patient Plasmas during Immune Tolerance Induction. *Blood*, 130(Suppl 1), 2017.

[53] Richard Prescott, Hiroaki Nakai, Evgueni L Saenko, Inge Scharrer, Inga Marie Nilsson, JE Humphries, Deborah Hurst, Gordon Bray, and Dorothea Scandella. The inhibitor antibody response is more complex in hemophilia A patients than in most nonhemophiliacs with factor VIII autoantibodies. *Blood*, 89(10):3663–71, 1997.

[54] M Shima, D Scandella, A Yoshioka, H Nakai, I Tanaka, S Kamisue, S Terada, and H Fukui. A factor VIII neutralizing monoclonal antibody and a human inhibitor alloantibody recognizing epitopes in the C2 domain inhibit factor VIII binding to von Willebrand factor and to phosphatidylserine. *Thrombosis and Haemostasis*, 69(3):240–246, mar 1993.

[55] By Dorothea Scandella, Gary E Gilbert, Midori Shima, Hiroaki Nakai, Christine Eagleson, Matthew Felch, Richard Prescott, K J Rajalakshmi, Leon W Hoyer, and Evgueni Saenko. Some Factor VIII Inhibitor Antibodies Recognize a Common Epitope Corresponding to. *Blood*, 86(5):1811–1819, 1995.

[56] Marc G. Jacquemin, Benoıt G. Desqueper, Abdellah Benhida, Luc Vander Elst, Marc F. Hoylaerts, Marleen Bakkus, Kris Thielemans, Jef Arnout, Kathelijne Peerlinck, Jean Guy G. Gilles, Jos Vermylen, and Jean-Marie R. Saint-Remy. Mechanism and Kinetics of Factor VIII Inactivation: Study With an IgG4 Monoclonal Antibody Derived From a Hemophilia A Patient With Inhibitor. *Blood*, 92(2), 1998.

[57] By John F Healey, Rachel T Barrow, Hiba M Tamim, Ira M Lubin, Midori Shima, Dorothea Scandella, and Pete Lollar. Residues Glu2181-Val2243 Contain a Major Determinant of the Inhibitory Epitope in the C2 Domain of Human Factor VIII. *Blood*, 92(10):3701–3710, 1998.

[58] R. Laub, M. Di Giambattista, P. Fondu, H. H. Brackmann, H. Lenk, E. L. Saenko, M. Felch, and D. Scandella. Inhibitors in German hemophilia A patients treated with a double virus inactivated factor VIII concentrate bind to the C2 domain of FVIII light chain. *Thrombosis and Haemostasis*, 81(1):39–44, 1999.

[59] Kathelijne Peerlinck, Marc G. Jacquemin, Jef Arnout, Marc F. Hoylaerts, Jean Guy G. Gilles, Renaud Lavend'homme, Karen M. Johnson, Kathleen Freson, Dorothea Scandella, Jean-Marie R. Saint-Remy, and Jos Vermylen.

Antifactor VIII Antibody Inhibiting Allogeneic but not Autologous Factor VIII in Patients With Mild Hemophilia A. *Blood*, 93(7), 1999.

[60] Kathleen P. Pratt, Betty W. Shen, Kazuya Takeshima, Earl W. Davie, Kazuo Fujikawa, and Barry L. Stoddard. Structure of the C2 domain of human factor VIII at 1.5 Å resolution. *Nature*, 402(6760):439–442, nov 1999.

[61] P C Spiegel, M Jacquemin, J M Saint-Remy, B L Stoddard, and K P Pratt. Structure of a factor VIII C2 domain-immunoglobulin G4kappa Fab complex: identification of an inhibitory antibody epitope on the surface of factor VIII. *Blood*, 98(1):13–9, jul 2001.

[62] R T Barrow, J F Healey, M G Jacquemin, J M Saint-Remy, and P Lollar. Antigenicity of putative phospholipid membrane-binding residues in factor VIII. *Blood*, 97(1):169–74, jan 2001.

[63] Shannon L Meeks, John F Healey, Ernest T Parker, Rachel T Barrow, and Pete Lollar. Antihuman factor VIII C2 domain antibodies in hemophilia A mice recognize a functionally complex continuous spectrum of epitopes dominated by inhibitors of factor VIII activation. *Blood*, 110(13):4234–42, dec 2007.

[64] Edward N. Van den Brink, Wendy S. Bril, Ellen A.M. Turenhout, Marleen Zuurveld, Niels Bovenschen, Marjolein Peters, Thynn Thynn Yee, Koen Mertens, Deborah A. Lewis, Thomas L. Ortel, Pete Lollar, Dorothea Scandella, and Jan Voorberg. Two classes of germline genes both derived from the VH1 family direct the formation of human antibodies that recognize distinct antigenic sites in the C2 domain of factor VIII. *Blood*, 99(8):2828–2834, apr 2002.

[65] A. E. Griffiths, W. Wang, F. K. Hagen, and Philip J. Fay. Use of affinity-directed liquid chromatography-mass spectrometry to map the epitopes of a factor VIII inhibitor antibody fraction. *Journal of Thrombosis and Haemostasis*, 9(8):1534–1540, aug 2011.

[66] Divi Venkateswarlu. Structural investigation of zymogenic and activated forms of human blood coagulation factor VIII: A computational molecular dynamics study. *BMC Structural Biology*, 10(1):1–20, feb 2010.

[67] Luca Mollica, Franca Fraternali, and Giovanna Musco. Interactions of the C2 domain of human factor V with a model membrane. *Proteins: Structure, Function and Genetics*, 64(2):363–375, may 2006.

[68] Jiangfeng Du, Kanin Wichapong, Tilman M. Hackeng, and Gerry A. F Nicolaes. Molecular simulation studies of human coagulation factor VIII C domain-mediated membrane binding. *Thrombosis and Haemostasis*, 113(02):373–384, mar 2015.

[69] Kenneth Segers, Olivier Sperandio, Markus Sack, Rainer Fischer, Maria A. Miteva, Jan Rosing, Gerry A.F. Nicolaes, and Bruno O. Villoutreix. Design of protein-membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12697–12702, jul 2007.

[70] Phuong Cac T. Nguyen, Kenneth B. Lewis, Ruth A. Ettinger, Jason T. Schuman, Jasper C. Lin, John F. Healey, Shannon L. Meeks, Pete Lollar, and Kathleen P. Pratt. High-resolution mapping of epitopes on the C2 domain of factor VIII by analysis of point mutants using surface plasmon resonance. *Blood*, 123(17):2732–2739, 2014.

[71] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, nov 1957.

[72] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[73] Adam Hospital, Josep Ramon Goñi, Modesto Orozco, and Josep L. Gelpí. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8(1):37–47, 2015.

[74] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, oct 2011.

[75] Micha B.A. Kunze, David W. Wright, Nicolas D. Werbeck, John Kirkpatrick, Peter V. Coveney, and D. Flemming Hansen. Loop interactions and dynamics tune the enzymatic activity of the human histone deacetylase 8. *Journal of the American Chemical Society*, 135(47):17862–17868, nov 2013.

[76] Lim Heo and Michael Feig. Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 115(52):13276–13281, dec 2018.

[77] Guillaume Bouvignies, Pramodh Vallurupalli, D. Flemming Hansen, Bruno E. Correia, Oliver Lange, Alaji Bah, Robert M. Vernon, Frederick W. Dahlquist, David Baker, and Lewis E. Kay. Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature*, 477(7362):111–117, sep 2011.

[78] S. Flores. The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Research*, 34(90001):D296–D301, jan 2006.

[79] Gira Bhabha, Justin T. Biel, and James S. Fraser. Keep on moving: Discovering and perturbing the conformational dynamics of enzymes. *Accounts of Chemical Research*, 48(2):423–430, feb 2015.

[80] Gregory M. Cockrell and Evan R. Kantrowitz. ViewMotions Rainbow: A new method to illustrate molecular motions in proteins. *Journal of Molecular Graphics and Modelling*, 40:48–53, mar 2013.

[81] Cornerstone Advisors Inc. https://www.crnrstone.com/case-studies/significant-improvement-opportunities-exist-credit-unions-benchmarking-study-reveals/, 2006.

[82] Mark Gerstein and Werner Krebs. A database of macromolecular motions. *Nucleic Acids Research*, 26(18):4280–4290, sep 1998.

[83] M. Kokkinidis, N. M. Glykos, and V. E. Fadouloglou. Protein flexibility and enzymatic catalysis. In *Advances in Protein Chemistry and Structural Biology*, volume 87, pages 181–218. Academic Press Inc., 2012.

[84] Martin Karplus. Role of conformation transitions in adenylate kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17), apr 2010.

[85] Rajesh K. Karmani, Gul Agha, Mark S. Squillante, Joel Seiferas, Marian Brezina, Jonathan Hu, Ray Tuminaro, Peter Sanders, Jesper Larsson Träffe, Robert A. Geijn, Jesper Larsson Träff, Robert A. Geijn, MS Benjamin Sander, John L. Gustafson, Ron O. Dror, Cliff Young, David E. Shaw, Calvin Lin, Jenq-Kuen Lee, Rong-Guey Chang, Chi-Bang Kuan, Giorgos Kollias, Ananth Y. Grama, Zhiyuan Li, R. Clint Whaley, and Richard W. Vuduc. Anton, A Special-Purpose Molecular Simulation Machine. In *Encyclopedia of Parallel Computing*, pages 60–71. Springer US, 2011.

[86] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, jan 2000.

[87] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[88] D.M. York D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGra and P.A. Kollman. AMBER 2018. *University of California, San Francisco*, 2018.

[89] William D. Lees, Lenka Stejskal, David S. Moss, and Adrian J. Shepherd. Investigating substitutions in antibody-antigen complexes using molecular dynamics: A case study with broad-spectrum, influenza A antibodies. *Frontiers in Immunology*, 8(FEB), feb 2017.

[90] Xintian Xu, Ping Chen, Jingfang Wang, Jiannan Feng, Hui Zhou, Xuan Li, Wu Zhong, and Pei Hao. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life Sciences*, 63(3):457–460, 2020.

[91] Paweł Śledź and Amedeo Caflisch. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology*, 48:93–102, feb 2018.

[92] Wendy D Cornell, Piotr Cieplak, / Christopher, I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc*, 117:5179–5197, 1995.

[93] T. Shimanouchi. Tables of molecular vibrational frequencies. Consolidated volume II. *Journal of Physical and Chemical Reference Data*, 6(3):993–1102, 1977.

[94] Jean-Paul Paul Ryckaert, Giovanni Ciccotti, and Herman J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, mar 1977.

[95] R. Elber, A. P. Ruymgaart, and B. Hess. SHAKE parallelization, nov 2011.

[96] Berk Hess, Henk Bekker, Herman J.C. Berendsen, and Johannes G.E.M. Fraaije. LINCS: A Linear Constraint Solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, sep 1997.

[97] K. Anton Feenstra, Berk Hess, and Herman J.C. Berendsen. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of Computational Chemistry*, 20(8):786–798, 1999.

[98] Chad W Hopkins, Scott Le Grand, Ross C Walker, and Adrian E Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. 2015.

[99] R. Bryn Fenwick, Santi Esteban-Martín, and Xavier Salvatella. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *European Biophysics Journal*, 40(12):1339–1355, 2011.

[100] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations. *Biophysical Journal*, 108(5):1153–1164, mar 2015.

[101] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, jul 1983.

[102] Alexey V. Onufriev and Saeed Izadi. Water models for biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(2):e1347, mar 2018.

[103] J. D. Bernal and R. H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *The Journal of Chemical Physics*, 1(8):515–548, aug 1933.

[104] Robert B. Best, Wenwei Zheng, and Jeetain Mittal. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124, nov 2014.

[105] Gül H. Zerze, Wenwei Zheng, Robert B Best, and Jeetain Mittal. Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *Journal of Physical Chemistry Letters*, 10(9):2227–2234, 2019.

[106] In Chul Yeh, Michael S. Lee, and Mark A. Olson. Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. *Journal of Physical Chemistry B*, 112(47):15064–15073, nov 2008.

[107] Patrice Koehl. Electrostatics calculations: latest methodological advances. *Current Opinion in Structural Biology*, 16(2):142–151, apr 2006.

[108] Alexey Onufriev, Donald Bashford, and David A. Case. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Structure, Function and Genetics*, 55(2):383–394, may 2004.

[109] Nicolas Calimet, Michael Schaefer, and Thomas Simonson. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins: Structure, Function and Genetics*, 45(2):144–158, nov 2001.

[110] Vickie Tsui and David A. Case. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *Journal of the American Chemical Society*, 122(11):2489–2498, mar 2000.

[111] David A. Onufriev, Alexey and Bashford, Donald and Case. Modification of the Generalized Born Model Suitable for Macromolecules. *The Journal of Physical Chemistry B*, 104(15):3712–3720, 2000.

[112] J. Weiser, Peter S. Shenkin, and W. Clark Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *Journal of Computational Chemistry*, 20(2):217–230, jan 1999.

[113] D.A. Case, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Ko-

valenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2017. Technical report, University of California, San Francisco, 2017.

[114] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, 10(5):449–61, may 2015.

[115] http://www.gromacs.org/.

[116] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD, 2005.

[117] https://www.charmm.org/.

[118] https://www.schrodinger.com/desmond.

[119] Michael R. Shirts, Christoph Klein, Jason M. Swails, Jian Yin, Michael K. Gilson, David L. Mobley, David A. Case, and Ellen D. Zhong. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *bioRxiv*, page 077248, sep 2016.

[120] http://ambermd.org/GPUPerformance.php#RCWBenchmarks.

[121] Susanna Hug. Classical Molecular Dynamics in a Nutshell. In *Biomolecular Simulations*, pages 127–152. Humana Press, Totowa, NJ, 2013.

[122] Meyer B. Jackson. On the time scale and time course of protein conformational changes. *The Journal of Chemical Physics*, 99(9):7253–7259, 1993.

[123] Levi C.T. Pierce, Romelia Salomon-Ferrer, Cesar Augusto F. De Oliveira, J. Andrew McCammon, and Ross C. Walker. Routine access to millisecond

time scale events with accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 8(9):2997–3002, sep 2012.

[124] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–12566, oct 2002.

[125] Michael K. Gilson and Huan Xiang Zhou. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36:21–42, 2007.

[126] R. G. Palmer. Broken ergodicity. *Advances in Physics*, 31(6):669–735, jan 1982.

[127] Arthur F. Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Physical Review Letters*, 78(20):3908–3911, 1997.

[128] Donald Hamelberg, John Mongan, and J. Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120(24):11919–11929, 2004.

[129] Romelia Salomon. http://ambermd.org/tutorials/advanced/tutorial22/, 2020.

[130] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, feb 1977.

[131] Glenn M. Torrie and John P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters*, 28(4):578–581, oct 1974.

[132] Benoît Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 1995.

[133] J. Kottalam and David A. Case. Dynamics of ligand escape from the heme pocket of myoglobin. *Journal of the American Chemical Society*, 110(23):7690–7697, nov 1988.

[134] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, oct 1992.

[135] Hemant Kumar Srivastava and G. Narahari Sastry. Molecular dynamics investigation on a series of HIV protease inhibitors: Assessing the performance of MM-PBSA and MM-GBSA approaches. *Journal of Chemical Information and Modeling*, 52(11):3088–3098, 2012.

[136] Kangpeng Xiao, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie Jian Zou, Na Li, Yaqiong Guo, Xiaobing Li, Xuejuan Shen, Zhipeng Zhang, Fanfan Shu, Wanyi Huang, Yu Li, Ziding Zhang, Rui Ai Chen, Ya Jiang Wu, Shi Ming Peng, Mian Huang, Wei Jun Xie, Qin Hui Cai, Fang Hui Hou, Wu Chen, Lihua Xiao, and Yongyi Shen. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, 583(7815):286–289, 2020.

[137] Salman Ali Khan, Komal Zia, Sajda Ashraf, Reaz Uddin, and Zaheer Ul-Haq. Identification of chymotrypsin-like protease inhibitors of SARS-CoV-2 via integrated computational approach. *Journal of Biomolecular Structure and Dynamics*, 2020.

[138] Panagiotis L. Kastritis and Alexandre M.J.J. Bonvin. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *Journal of Proteome Research*, 9(5):2216–2225, 2010.

[139] Chunlan Pu, Guoyi Yan, Jianyou Shi, and Rui Li. Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. *MedChemComm*, 8(7):1452–1458, 2017.

[140] Fu Chen, Hui Liu, Huiyong Sun, Peichen Pan, Youyong Li, Dan Li, and Tingjun Hou. Assessing the performance of the MM/PBSA and MM/GBSA

methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. *Physical Chemistry Chemical Physics*, 18(32):22129–22139, aug 2016.

[141] Tatu Pantsar and Antti Poso. Binding affinity via docking: Fact and fiction. *Molecules*, 23(8):1899, 2018.

[142] Conor D. Parks, Zied Gaieb, Michael Chiu, Huanwang Yang, Chenghua Shao, W. Patrick Walters, Johanna M. Jansen, Georgia McGaughey, Richard A. Lewis, Scott D. Bembenek, Michael K. Ameriks, Tara Mirzadegan, Stephen K. Burley, Rommie E. Amaro, and Michael K. Gilson. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 34(2):99–119, 2020.

[143] Haruto Ishikawa, Kyungwon Kwak, Jean K. Chung, Seongheun Kim, and Michael D. Fayer. Direct observation of fast protein conformational switching. *Proceedings of the National Academy of Sciences of the United States of America*, 105(25):8619–8624, jun 2008.

[144] Chia En Chang, Wei Chen, and Michael K. Gilson. Evaluating the accuracy of the quasiharmonic approximation. *Journal of Chemical Theory and Computation*, 1(5):1017–1028, 2005.

[145] Samuel Genheden and Ulf Ryde. Will molecular dynamics simulations of proteins ever reach equilibrium? *Physical Chemistry Chemical Physics*, 14(24):8662–8677, 2012.

[146] Jacob Kongsted and Ulf Ryde. An improved method to predict the entropy term with the MM/PBSA approach. *Journal of Computer-Aided Molecular Design*, 23(2):63–71, 2009.

[147] Huiyong Sun, Lili Duan, Fu Chen, Hui Liu, Zhe Wang, Peichen Pan, Feng Zhu, John Z. H. Zhang, and Tingjun Hou. Assessing the performance of

MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Physical Chemistry Chemical Physics*, 20(21):14450–14460, may 2018.

[148] Martha S. Head, James A. Given, and Michael K. Gilson. "Mining minima": Direct computation of conformational free energy. *Journal of Physical Chemistry A*, 101(8):1609–1618, 1997.

[149] Chia En A. Chang, Wei Chen, and Michael K Gilson. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5):1534–1539, jan 2007.

[150] Sailu Sarvagalla, Chun Hei Antonio Cheung, Ju Ya Tsai, Hsing Pang Hsieh, and Mohane Selvaraj Coumar. Disruption of protein-protein interactions: Hot spot detection, structure-based virtual screening and: In vitro testing for the anti-cancer drug target-survivin. *RSC Advances*, 6(38):31947–31959, 2016.

[151] Gaetano Calabrò, Christopher J. Woods, Francis Powlesland, Antonia S.J.S. Mey, Adrian J. Mulholland, and Julien Michel. Elucidation of Nonadditive Effects in Protein-Ligand Binding Energies: Thrombin as a Case Study. *Journal of Physical Chemistry B*, 120(24):5340–5350, 2016.

[152] Alan E. Mark and Wilfred F. Van Gunsteren. Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *Journal of Molecular Biology*, 240(2):167–176, 1994.

[153] Andrej Šali and Tom L. Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815, dec 1993.

[154] Swagata Dasgupta, Ganesh H. Iyer, Stephen H. Bryant, Charles E. Lawrence, and Jeffrey A. Bell. Extent and nature of contacts between protein molecules

in crystal lattices and between subunits of protein oligomers. *Proteins: Structure, Function and Genetics*, 28(4):494–514, 1997.

[155] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, David C Richardson, and David C. Richardson. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 1):12–21, jan 2010.

[156] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[157] Boris Aguilar and Alexey V. Onufriev. Efficient computation of the total solvation energy of small molecules via the r6 generalized born model. *Journal of Chemical Theory and Computation*, 8(7):2404–2411, jul 2012.

[158] D. L. Dorset. X-ray Diffraction: A Practical Approach. *Microscopy and Microanalysis*, 4(5):513–515, 1998.

[159] B. Knapp, S. Frantal, M. Cibena, W. Schreiner, and P. Bauer. Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible? *Journal of Computational Biology*, 18(8):997–1005, aug 2011.

[160] Samuel Genheden, Tyler Luchko, Sergey Gusarov, Andriy Kovalenko, and Ulf Ryde. An MM/3D-RISM approach for ligand binding affinities. *Journal of Physical Chemistry B*, 2010.

[161] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, 2011.

[162] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, 8(9):3314–3321, sep 2012.

[163] Andrew C. Wallace, Roman A. Laskowski, and Janet M. Thornton. Ligplot: A program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering, Design and Selection*, 8(2):127–134, 1995.

[164] Roman A. Laskowski and Mark B. Swindells. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10):2778–2786, 2011.

[165] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z.H. Zhang, and Tingjun Hou. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16):9478–9508, 2019.

[166] Lili Duan, Xiao Liu, and John Z.H. Zhang. Interaction entropy: A new paradigm for highly efficient and reliable computation of protein-ligand binding free energy. *Journal of the American Chemical Society*, 138(17):5722–5728, 2016.

[167] Holger Gohlke, David A Case, Molecular Biology, The Scripps, and N Torrey Pines Rd. Converging Free Energy Estimates : MM-PB (GB) SA Studies on the Protein – Protein Complex Ras – Raf. pages 238–250, 2003.

[168] Donald Hamelberg and J. Andrew McCammon. Fast peptidyl cis-trans isomerization within the flexible gly-rich flaps of HIV-1 protease. *Journal of the American Chemical Society*, 127(40):13778–13779, 2005.

[169] Jiří Šponer, Judit E. Šponer, Arnošt Mládek, Petr Jurečka, Pavel Banáš, and Michal Otyepka. Nature and magnitude of aromatic base stacking in DNA and RNA: Quantum chemistry, molecular mechanics, and experiment. *Biopolymers*, 99(12):978–988, 2013.

[170] Yinglong Miao, William Sinko, Levi Pierce, Denis Bucher, Ross C. Walker, and J. Andrew McCammon. Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *Journal of Chemical Theory and Computation*, 10(7):2677–2689, 2014.

[171] Yinglong Miao, Victoria A. Feher, and J. Andrew McCammon. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation*, 11(8):3584–3595, aug 2015.

[172] Jordi Juárez-Jiménez, Arun A. Gupta, Gogulan Karunanithy, Antonia S.J.S. Mey, Charis Georgiou, Harris Ioannidis, Alessio De Simone, Paul N. Barlow, Alison N. Hulme, Malcolm D. Walkinshaw, Andrew J. Baldwin, and Julien Michel. Dynamic design: Manipulation of millisecond timescale motions on the energy landscape of cyclophilin A. *Chemical Science*, 11(10):2670–2680, 2020.

[173] Ilyas Yildirim, Hajeung Park, Matthew D. Disney, and George C. Schatz. A dynamic structural model of expanded RNA CAG repeats: A refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *Journal of the American Chemical Society*, 135(9):3528–3538, mar 2013.

[174] Hengameh Shams, Javad Golji, and Mohammad R.K. Mofrad. A Molecular Trajectory of α-Actinin Activation. *Biophysical Journal*, 103(10):2050–2059, nov 2012.

[175] G. P. Moss. Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*, 68(12):2193–2222, jan 1996.

[176] Wei Jiang, Yun Luo, Luca Maragliano, and Benoît Roux. Calculation of free energy landscape in multi-dimensions with hamiltonian-exchange umbrella sampling on petascale supercomputer. *Journal of Chemical Theory and Computation*, 8(11):4672–4680, 2012.

[177] Juan A. Bueren-Calabuig and Julien Michel. Elucidation of Ligand-Dependent Modulation of Disorder-Order Transitions in the Oncoprotein MDM2. *PLoS Computational Biology*, 11(6):1–27, 2015.

[178] Alejandro Gil-Ley and Giovanni Bussi. Enhanced conformational sampling using replica exchange with collective-variable tempering. *Journal of Chemical Theory and Computation*, 11(3):1077–1085, 2015.

[179] Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[180] Jing Jin, Hao Wang, Zhongjun Shu, and Lingang Lu. Effect of Two Maleic Anhydride Grafted Polymers as Modifier on Intumescent Flame Retardancy and Mechanical Property of Polypropylene Based Composites. *Cailiao Yanjiu Xuebao/Chinese Journal of Materials Research*, 31(3):219–225, 2017.

[181] Ian T Jolliffe, Jorge Cadima, and Jorge Cadima. Principal component analysis : a review and recent developments Subject Areas. *Phil.Trans.R.Soc.A*, 374(2065):1–16, 2016.

[182] Adam T. Van Wart, Jacob Durrant, Lane Votapka, and Rommie E. Amaro. Weighted implementation of suboptimal paths (WISP): An optimized algorithm and tool for dynamical network analysis. *Journal of Chemical Theory and Computation*, 10(2):511–517, feb 2014.

[183] Jianqing Fan, Cong Ma, and Yiqiao Zhong. A Selective Overview of Deep Learning, apr 2019.

[184] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[185] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv*, (1708.08296), 2017.

[186] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus Robert Müller. Layer-Wise Relevance Propagation: An Overview. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11700 LNCS, pages 193–209. Springer Verlag, 2019.

[187] Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39, sep 2018.

[188] Szilard Pafka. Benchmarking Random Forest Implementations — Data Science Los Angeles. http://datascience.la/benchmarking-random-forest-implementations/, 2015.

[189] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.

[190] Harris Drucker. Improving regressors using boosting techniques. *14th International Conference on Machine Learning*, (June):107–115, 1997.

[191] Scott Lundberg. Explainers — SHAP latest documentation. https://shap.readthedocs.io/en/latest/, 2020.

[192] Simon Brandt, Florian Sittel, Matthias Ernst, and Gerhard Stock. Machine Learning of Biomolecular Reaction Coordinates. *Journal of Physical Chemistry Letters*, 9(9):2144–2150, 2018.

[193] Amazon Web Services (AWS) - Cloud Computing Services. https://aws.amazon.com/, 2020.

[194] Scott M Lundberg and Su In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 4766–4775, 2017.

[195] Daping Yu, Zhidong Liu, Chongyu Su, Yi Han, XinChun Duan, Rui Zhang, Xiaoshuang Liu, Yang Yang, and Shaofa Xu. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. *Thoracic Cancer*, 11(1):95–102, jan 2020.

[196] Amir Bahador Parsa, Ali Movahedi, Homa Taghipour, Sybil Derrible, and Abolfazl (Kouros) Mohammadian. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis and Prevention*, 136, 2020.

[197] Andreja Stojić, Nenad Stanić, Gordana Vuković, Svetlana Stanišić, Mirjana Perišić, Andrej Šoštarić, and Lazar Lazić. Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. *Science of the Total Environment*, 653:140–147, feb 2019.

[198] CDC Hemophilia Mutation Project (CHAMP & CHBMP). https://www.cdc.gov/ncbddd/hemophilia/champs.html.

[199] J L Pellequer, A J Gale, J H Griffin, and E D Getzoff. Homology models of the C domains of blood coagulation factors V and VIII: a proposed membrane binding mode for FV and FVIII C2 domains. *Blood cells, molecules & diseases*, 24(4):448–61, dec 1998.

[200] Haoran Yu, Yang Zhao, Chao Guo, Yiru Gan, and He Huang. The role of proline substitutions within flexible regions on thermostability of luciferase. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854(1):65–72, 2015.

[201] W J Krzanowski. Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, 74(367):703–707, 1979.

[202] A. D. Gordon. Cluster Validation. In *Data Science, Classification, and Related Methods*, number 1971, pages 22–39. 1998.

[203] James Keeler. *Understanding NMR spectroscopy*. John Wiley & Sons, 2011.

[204] Kristin M. Nuzzio, David B. Cullinan, Valerie A. Novakovic, John M. Boettcher, Chad M. Rienstra, Gary E. Gilbert, and James D. Baleja. Backbone resonance assignments of the C2 domain of coagulation factor VIII. *Biomolecular NMR Assignments*, 7(1):31–34, apr 2013.

[205] Julie Lobstein, Charlie A. Emrich, Chris Jeans, Melinda Faulkner, Paul Riggs, and Mehmet Berkmen. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories*, 11, may 2012.

[206] Fisher Scientific. Invitrogen™ PureLink™ Quick Plasmid Miniprep Kit. https://www.fishersci.de/shop/products/invitrogen-purelink-quick-plasmid-miniprep-kit-2/p-4926496?change_lang=true, 2020.

[207] K. Linderstrom-Lang. Enzymes. *Annual Review of Biochemistry*, 6(1):43–72, jun 1937.

[208] Zimei Bu and David J.E. Callaway. *Proteins move! Protein dynamics and long-range allostery in cell signaling*, volume 83. Elsevier Inc., 1 edition, 2011.

[209] Daniel J. Rigden. *From protein structure to function with bioinformatics: Second Edition*. 2017.