

**BIOINFORMATIC APPROACHES TO
INVESTIGATE HIV CAPSID-NANOBODY
INTERACTION**

**A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
MASTER OF SCIENCE
in Bioengineering**

**by
Şeref Berk ATİK**

**July 2023
İZMİR**

We approve the thesis of **Şeref Berk ATİK**

Asst. Prof. Dr. Hümevra TAŞKENT SEZGİN

Department of Bioengineering, Izmir Institute of Technology

Asst. Prof. Dr. Arzu UYAR

Department of Bioengineering, Izmir Institute of Technology

Asst. Prof. Dr. Cihangir YANDIM

Department of Genetics and Bioengineering, Izmir University of Economics

14 July 2023

Asst. Prof. Dr. Hümevra TAŞKENT SEZGİN

Supervisor, Department of Bioengineering

Izmir Institute of Technology

Prof. Dr. Engin ÖZÇİVİCİ

Head of the Department of Bioengineering

Prof. Dr. Mehtap EANES

Head of the Graduate School of
Engineering & Science

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my dear supervisor Asst. Prof. Dr. Hümeyra TAŞKENT SEZGİN for her endless support, encouragement, motivation, and guidance throughout my thesis studies. She always shared her wisdom and knowledge with me, and having her as my mentor was a pleasure.

I am grateful to my thesis committee members Asst. Prof. Dr. Arzu UYAR and Asst. Prof. Dr. Cihangir YANDIM for their contributions and guidance to support my thesis.

I want to thank TUSEB for supporting and funding my research under Group A Urgent R&D Funding.

Finally, I am deeply grateful to my family: My mother, my brother, and my sister-in-law. They believed in me in every decision I made and always encouraged me to be a better person.

ABSTRACT

BIOINFORMATIC APPROACHES TO INVESTIGATE HIV CAPSID-NANOBODY INTERACTION

Infection with HIV is still a global pandemic. Since the discovery of this highly mutagenic virus, nearly 40 million people have passed away as a result of HIV-related health problems. Currently, 38.4 million people are HIV-positive. Following infection, the viral genome gets integrated into the host cell genome. The infected person carries the virus for the rest of their life and can spread it to others through bodily fluids. Because there is no treatment for HIV, the World Health Organization recommends that infected people be diagnosed early through comprehensive screening to restrict the virus's spread. As a result, there is still a need to create practical, sensitive diagnostic tools, particularly for use in the field of HIV infection testing. In this study, the interaction between HIV-1 capsid protein, the first antigen found in the blood during the acute phase of HIV infection, and a nanobody (Nb, a single domain antibody) known to bind to capsid is investigated at the molecular level through computational methods. Because the structure of HIV-1 CA binding-Nb is unknown, all-atom models of the Nb structure were constructed using comparative methods, deep-learning-based methods, and hybrid methods (SwissModel, trRosetta, Robetta, AlphaFold2), and promising models were chosen. In the second stage, molecular docking was used to produce HIV-1 capsid-nanobody complex structures, which were then tested for stability and native-likeness using standard molecular dynamics simulations. Understanding the molecular details of the HIV-1 capsid-nanobody complex, we believe, will provide essential data for using this antigen-antibody pair in an immunosensor system for HIV-1 infection diagnosis.

ÖZET

BİYOİNFORMATİK YÖNTEMLERLE HIV KAPSİT-NANOBADİ ANTİKOR ETKİLEŞİMİNİN İNCELEMESİ

HIV enfeksiyonu küresel bir salgın olarak devam etmektedir. Bu mutasyon geçirme kapasitesi yüksek virüsün keşfinden bu yana, yaklaşık 40 milyon insan HIV enfeksiyonu sonucunda oluşan fırsatçı enfeksiyonlar veya hastalıklar dolayısıyla hayatını kaybetmiştir. Günümüzde yaklaşık 38.4 milyon insan HIV-pozitif oldukları tahmin edilmektedir. Enfeksiyondan sonra virüs viral genomunu konak hücre genomuna entegre eder. Enfekte olan kişi, geri kalan hayatı boyunca virüsü taşır ve vücut sıvıları aracılığıyla diğer insanlara bulaştırabilir. HIV enfeksiyonu için henüz bir tedavi olmadığından, Dünya Sağlık Örgütü, virüsün yayılmasını kısıtlamak için enfekte olan insanların kapsamlı tarama yoluyla erken teşhis edilmesini önermektedir. Bu nedenle, özellikle HIV enfeksiyonu testi için sahada kullanılmak üzere pratik, ve hassas tanı araçlarına ihtiyaç duyulmaktadır. Bu çalışmada, HIV-1 kapsid proteini (HIV-1 CA) ile kapside bağlanan bir nanobadi (Nb, tek bölgeli antikor) arasındaki etkileşimin moleküler detayları incelenmektedir. Nb proteininin 3-boyutlu yapısı bilinmediği için, Nb yapısının tüm atom modelleri karşılaştırmalı yöntemler, derin öğrenme tabanlı yöntemler ve hibrit yöntemler (SwissModel, trRosetta, Robetta, AlphaFold2) kullanılarak oluşturuldu ve uygun olabilecek modeller seçildi. İkinci aşamada, moleküler yanaştırma yöntemleri kullanılarak olası HIV-1 kapsid-nanobadi kompleksi yapıları üretildi. Son aşamada, moleküler yanaştırma çalışmasında uygun bulunan HIV-1 CA-Nb kompleks yapıların standart moleküler dinamik simülasyonları ile stabiliteleri ve doğal benzerlikleri test edildi. Bu çalışmada HIV-1 kapsid-nanobadi kompleks yapısının çözümlenmesiyle, Nb proteini ile geliştirilecek bir immünosensör ile HIV'in farklı suşlarının tanısı için kullanılması mümkün olacaktır.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	xii
CHAPTER 1. INTRODUCTION	1
1.1. Human Immunodeficiency Virus (HIV).....	1
1.1.1. Viral Structure of HIV.....	3
1.1.2. HIV Viral Actions and Infection.....	5
1.1.3. Current Diagnostic Techniques for HIV	7
1.1.4. HIV-1 and HIV-2 Comparison.....	9
1.1.5. HIV-1 Capsid Protein.....	10
1.2. Nanobodies	12
1.2.1. Nanobody Structure.....	13
1.2.2. Applications of Nanobodies.....	16
1.3. Computational Methods for Understanding Protein-Protein Interactions	17
1.3.1. Protein Structure Prediction	17
1.3.2. Molecular Docking.....	18
1.3.3. Molecular Dynamics	18
1.4. Aim of the Study.....	19

CHAPTER 2. MATERIALS AND METHODS	20
2.1. Data Collection	20
2.1.1. Protein Data Bank	20
2.1.2. The Single Domain Antibody Database	21
2.1.3. Los Alamos National Laboratory	22
2.2. Multiple Sequence Alignment of HIV-1 Capsid	23
2.3. Modelling the 3D Structure of the Nanobody	23
2.3.1. Homology Modelling	24
2.3.2. Structure Assessment Test	28
2.4. Determining Protein-Protein Complex Orientation.....	35
2.4.1. Molecular Docking.....	36
2.4.2. Evaluation of the Docked Complexes	37
2.5. Molecular Dynamics Simulations	44
2.5.1. Simulation Ensembles	44
2.5.2. Periodic Boundary Condition (PBC).....	45
2.5.3. Force Fields	45
 CHAPTER 3. RESULTS AND DISCUSSION	 48
3.1. HIV-1 Capsid Protein Multiple Sequence Alignment	48
3.2. Modeling Nanobody 3D Structure	50
3.3. Structure Assessment of the Nanobody	53
3.4. HIV-1 Capsid Protein and Nanobody Interaction Analysis	57
3.4.1. ZDock Blind Docking	57
3.4.2. ClusPro Blind Docking	58
3.4.3. Haddock Guided Docking and MD Simulations.....	61
	vii

3.4.4. ClusPro Guided Docking and MD Simulations	67
CHAPTER 4. CONCLUSION.....	94
REFERENCES	96
APPENDIX A	104

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
Figure 1. UNAIDS 2021 Adults and children living with HIV	2
Figure 2. HIV transmission mechanism. (<i>HIV Infection / BioNinja</i> , n.d.)	2
Figure 3. Graphical representation of the HIV viral structure.....	4
Figure 4. Viral lifecycle of HIV (Engelman & Cherepanov, 2012).	6
Figure 5. Viral RNA, antigen and antibody response levels through.....	9
Figure 6. Maturation of HIV-1 through sequential cleavage of Gag polyprotein... ..	10
Figure 7. HIV Capsid structures... ..	11
Figure 8. Number of articles per year from “nanobody” keyword search of PubMed... ..	13
Figure 9. Comparative representation of classical IgG antibody... ..	14
Figure 10. Representation of a nanobody structure... ..	15
Figure 11. Sequence variability of 90 non-redundant, protein-binding nanobodies... ..	15
Figure 12. sdAb-DB submission and accession workflow (Wilton et al., 2018).	22
Figure 13. Evaformer block in AlphaFold2 pipeline (Jumper et al., 2021).....	26
Figure 14. Structure module in AlphaFold2 (Jumper et al., 2021).	26
Figure 15. The trRosetta protocol is visualized in a flowchart, outlining the.....	28
Figure 16. An empty general (No Proline or Glycine) Ramachandran plot... ..	29
Figure 17. Example output of a MolProbity analysis.	30
Figure 18. CASP target T0542 reference model and predicted structure model... ..	32
Figure 19. Normalized QMEAN scores are expressed as z-scores in comparison.....	33
Figure 20. An example of results of QMEANDisCo evaluation... ..	35
Figure 21. Important residues that take part in the formation of capsid core. (A)... ..	43
Figure 22. Sequence logo representation of 414 aligned HIV-1 capsid sequences... ..	49
Figure 23. HIV-1 Capsid protein (PDBID: 4XFX) highlighted according to... ..	50
Figure 24. Sequence 3D computational models of CANTDcb1... ..	51
Figure 25. Evaluation of CANTDcb1 on Ramachandran plot.....	53
Figure 26. Local QMEANDisCo charts of the CANTDcb1 models... ..	55

Figure 27. Normalized QMEAN scores of CANTDcb1models expressed as.....	56
Figure 28. Residues commonly involved in salt bridges in HIV-1 capsid....	59
Figure 29. Residues commonly involved in salt bridges in CANTDcb1....	59
Figure 30. Residues commonly involved in salt bridges in HIV-1 capsid in....	60
Figure 31. Residues commonly involved in salt bridges in CANTDcb1 in....	60
Figure 32. PDBSum results of restricted Haddock run. In both figures.....	62
Figure 33. Docked complexes of restricted Haddock run....	63
Figure 34. RMSD calculations of the structures through the simulation.....	64
Figure 35. HIV-1 CA:Glu98 and CANTDcb1:Arg44 distances per atom....	65
Figure 36. RMSD calculations of the structures through the simulation of....	66
Figure 37. HIV-1 CA:Glu98 and CANTDcb1:Arg44 distances through....	67
Figure 38. First of the docking results from the Robetta model docking with....	69
Figure 39. RMSD calculations of the first complex of the docking results from.....	70
Figure 40. HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as shown with.....	71
Figure 41. HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as shown with.....	72
Figure 42. HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as shown with....	73
Figure 43. Second run RMSD calculations of the first complex of the docking.....	74
Figure 44. Second run HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as....	75
Figure 45. Second run HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as....	76
Figure 46. Second run HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as.....	77
Figure 47. Third run RMSD calculations of the first complex of the docking.....	78
Figure 48. Third run HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as....	79
Figure 49. Third run HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as....	80
Figure 50. Third run HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as.....	81
Figure 51. Eighth Complex of the docking results from Robetta model docking....	82
Figure 52. First run RMSD calculations of the eighth complex of the docking.....	86
Figure 53. Second run RMSD calculations of the eighth complex of the....	87
Figure 54. First run point of view for the eighth complex of the docking results....	88
Figure 55. Second run point of view for the eighth complex of the docking....	89
Figure 56. 80 th percent analysis of HIV-1 CA Met10:O-CANTDcb1 LEU109:HA....	90
Figure 57. 80 th percent analysis of HIV-1 CA Met10:O-CANTDcb1 Tyr110:H.....	90
Figure 58. 80 th percent analysis of HIV-1 CA Val11:HA-CANTDcb1 Tyr110:O.....	91

Figure 59. 80th percent analysis of HIV-1 CA Val10:O-CANTDcb1 Tyr113:HZ2..... 91

Figure 60. First complex of the docking results from the trRosetta model docking... ..92

LIST OF TABLES

<u>Table</u>	<u>Page</u>
Table 1. Current Diagnostic Tests for HIV Infection... ..	7
Table 2. MolProbity results of the CANTDcb1 models.	54
Table 3. Interactions in two MD runs of model 8 complex of the docking results... ..	84

CHAPTER 1

INTRODUCTION

1.1. Human Immunodeficiency Virus (HIV)

HIV is the leading cause of Acquired Immunodeficiency Syndrome (AIDS), and approximately 38.4 million people are infected with HIV, according to 2021 Joint United Nations Programme on HIV/AIDS (UNAIDS) data. Most of these infections are in Eastern and Southern Africa, with approximately 20.6 million people living with HIV (Figure 1) (UNAIDS Fact Sheet). can be transmitted through body fluid, as represented in Figure 2.

Adults and children estimated to be living with HIV | 2021



Figure 1. UNAIDS 2021 Adults and children living with HIV (*UNAIDS 2021 Adults and Children Living with HIV*, 2021).



Figure 2. HIV transmission mechanism. (*HIV Infection* / BioNinja, n.d.)

The target of this virus is the CD4⁺ T cells, which are part of the immune system. Infection of the CD4⁺ T cells with HIV leads to a weakened immune system since they cannot fulfill their purpose. Thus, leaving the body vulnerable to other infections. Without treatment, this viral infection can lead to AIDS and maybe death due to opportunistic infections (UNAIDS).

Replicating by RNA without any proof-reading mechanism causes HIV to have a higher mutation rate, thus leading to greater genetic diversity (Rihn et al., 2013a). HIV has two types named HIV-1 and HIV-2; among them, HIV-1 has the most common occurrence of infection. HIV-1 is examined under four groups, M, N, O, and P, with several subtypes under them (Bbosa et al., 2019).

1.1.1. Viral Structure of HIV

HIV is a lentivirus which is a genus of retroviruses. The genetic material of HIV is composed of two identical copies of single-stranded RNA(ssRNA). Structural proteins (gag), viral enzymes (pol), and viral envelope proteins (env) are encoded in this genome (Wilk et al., 2001).

The structural proteins to mention include matrix (MA), capsid (CA), and nucleocapsid (NC), which are derived from the polyprotein Gag. MA protein is associated with the inner surface of the viral membrane and takes a role in viral budding. CA protein surrounds the genetic material and takes part in delivering genetic material during infection. NC protein participates in the specific encapsidation of RNA (Ganser-Pornillos et al., 2008; Mishra et al., 2020; Zhang et al., 1998). The viral enzymes include; reverse transcriptase (RT), integrase (IN), and protease PR. RT is responsible for building the DNA copy of viral RNA, and IN is responsible for incorporating this DNA copy into the host cell genome. PR takes part in the maturation of the virus; this enzyme cleaves viral polyproteins into their functional pieces (Johnston & Hoth, 1993; *The Structural Biology*

of *HIV*, n.d.). Surface and transmembrane envelop proteins gp120 and gp41 take part in the recognition and infection of the target proteins (Julien et al., 2013).

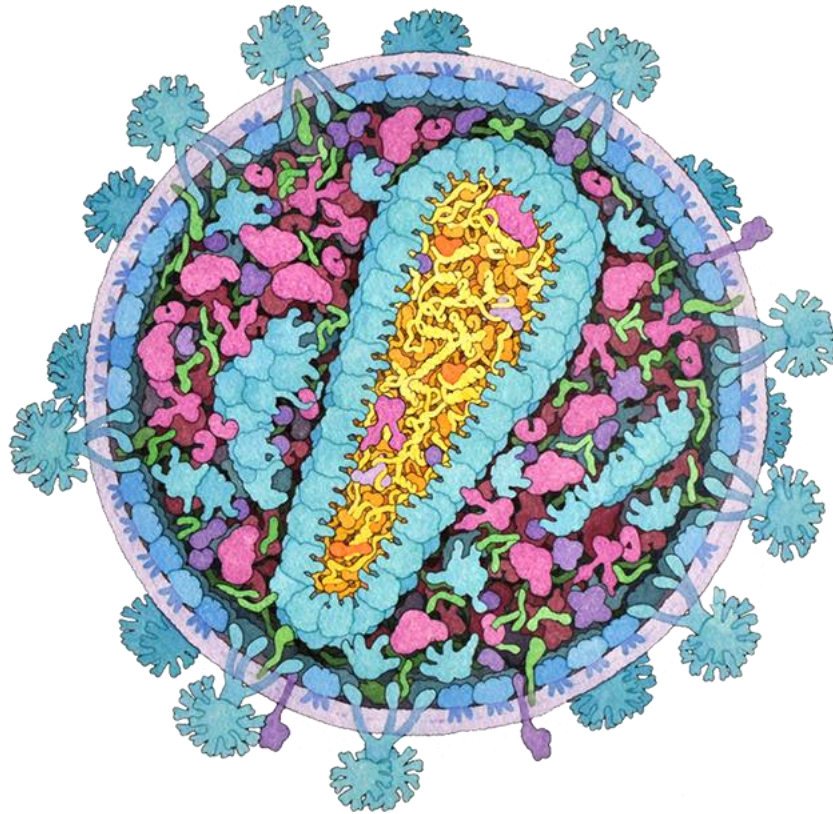


Figure 3. Graphical representation of the HIV viral structure. Structural proteins, except nucleocapsid, are represented in blue, viral enzymes are represented in pink, and accessory proteins are represented in green (*The Structural Biology of HIV*, n.d.).

1.1.2. HIV Viral Actions and Infection

Infection of HIV starts with viral entry to the host cell. Envelope proteins of HIV, surface and transmembrane glycoproteins gp120 and gp41, recognize the target protein's surface receptors and cause a cascade of conformational changes to merge the viral membrane into the target cell's membrane. Thus, the HIV core enters the target protein cytoplasm, shown in steps 1 and 2 of Figure 4. The viral core comprises RNA enclosed by capsid cage structure, RT, and IN enzymes. After entering the target cell, the capsid cage structure opens to release genomic material, as shown in step 3 of Figure 4. During the infection period, viral RT synthesizes the viral DNA from viral ssRNA, then viral IN integrates viral DNA into the host cell's DNA, shown in steps 4-6 in Figure 4 (Engelman & Cherepanov, 2012).

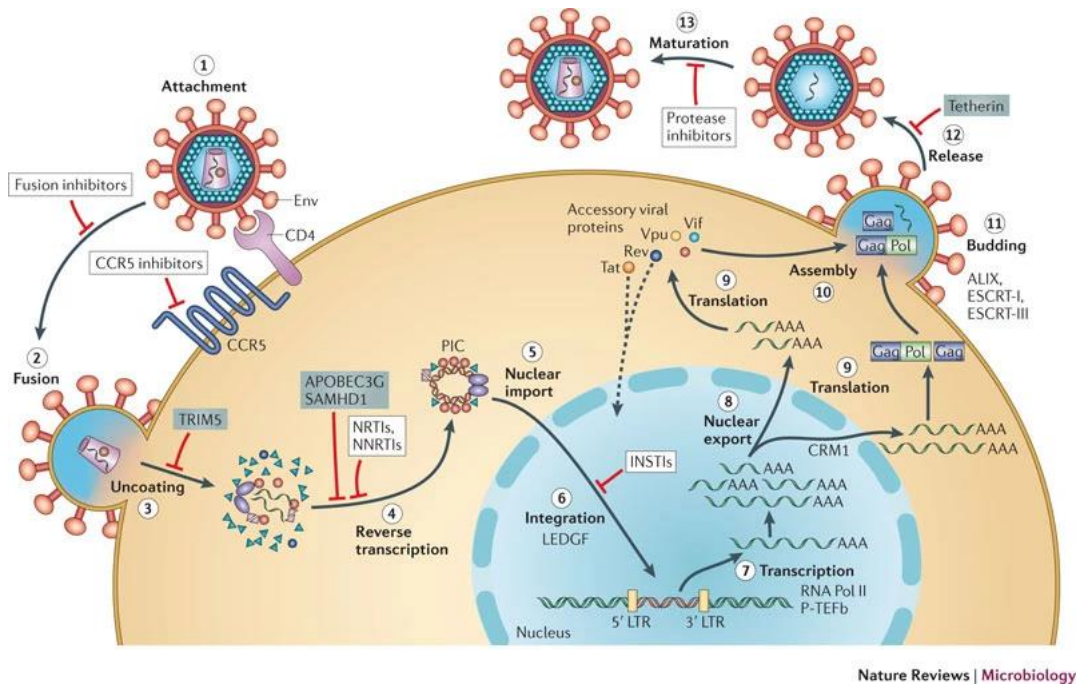


Figure 4. Viral lifecycle of HIV (Engelman & Cherepanov, 2012).

After the infection, the synthesis of the new viruses starts within the host cell. The host cell's transcription mechanism mediates this synthesis through the integrated viral DNA. Synthesized virus-like particles assemble at the plasma membrane, and viral budding occurs, resulting in an immature virus. Viral PR processes Gag polyprotein outside the host cell, leading to viral maturation. After maturation, the new virus can infect new cells (Engelman & Cherepanov, 2012; Ganser-Pornillos et al., 2008).

1.1.3. Current Diagnostic Techniques for HIV

Diagnosis of the virus is an important starting point for treatment and taking preventative measures for the spread of infection. There are several methods for the detection of HIV listed in Table 1. During the acute phase of the infection, HIV possesses high replication potential and the highest infectivity rate. In the acute period, an immune response is not yet developed, and diagnosis can only be made through high-sensitivity assays, such as nucleic acid amplification tests (NAAT). Although NAAT is extensively used in resource-rich environments, it is in limited use for resource-limited environments due to being expensive and complex. Therefore, accessible point-of-care (PoV) diagnostic tests can benefit resource-limited settings (Cornett & Kirn, 2013; Gray et al., 2018).

Table 1. Current Diagnostic Tests for HIV Infection. The table is adapted from Cornett & Kirn, 2013.

Technology	Principle	Strengths	Limitations
First- and second-generation immunoassays	Detect IgG response	Detect HIV-specific IgG	Do not detect HIV-specific IgM and antigens
Third-generation immunoassays	Detect IgG and IgM response	HIV specific IgMs are present earlier in body.	Do not detect HIV antigens

(cont. on the next page)

Cont. of Table 1

Fourth-generation immunoassays	Anti-HIV Abs are detected by recombinant antigens and p24 antigen is detected by antihuman Abs	Detect both Abs and Ags, allowing detection before immune response generated	May miss HIV infections before viral antigen reaches detectable levels
Rapid tests	Employs lateral flow, immunoconcentration, or particle agglutination technologies	Performs similar to lab-based immunoassays with <30 min of completion time	Generation dependent
NAATs	Nucleic acids are amplified with specific primers and detected with labeled probes	Detect acute HIV infection before viral antigen reaches detectable levels	Complex and expensive. Most only detect HIV-1 and can provide false-negatives in some Ab-positive cases

After the transmission, the earliest detectable virus particles are viral RNA and the p24 capsid protein of HIV. These particles reach detectable levels after the first and second week, respectively. Then, the immune response is generated, and diagnosis can proceed through antibody response. Currently, fourth-generation antibody-antigen assays are the recommended method of diagnosis. These methods detect both p24 capsid protein and antibody response to viral infection. Viral RNA, antigen, and antibody response levels according to days after transmission are shown in Figure 5, with diagnostic tests according to their earliest applicable timeframes (Gray et al., 2018).

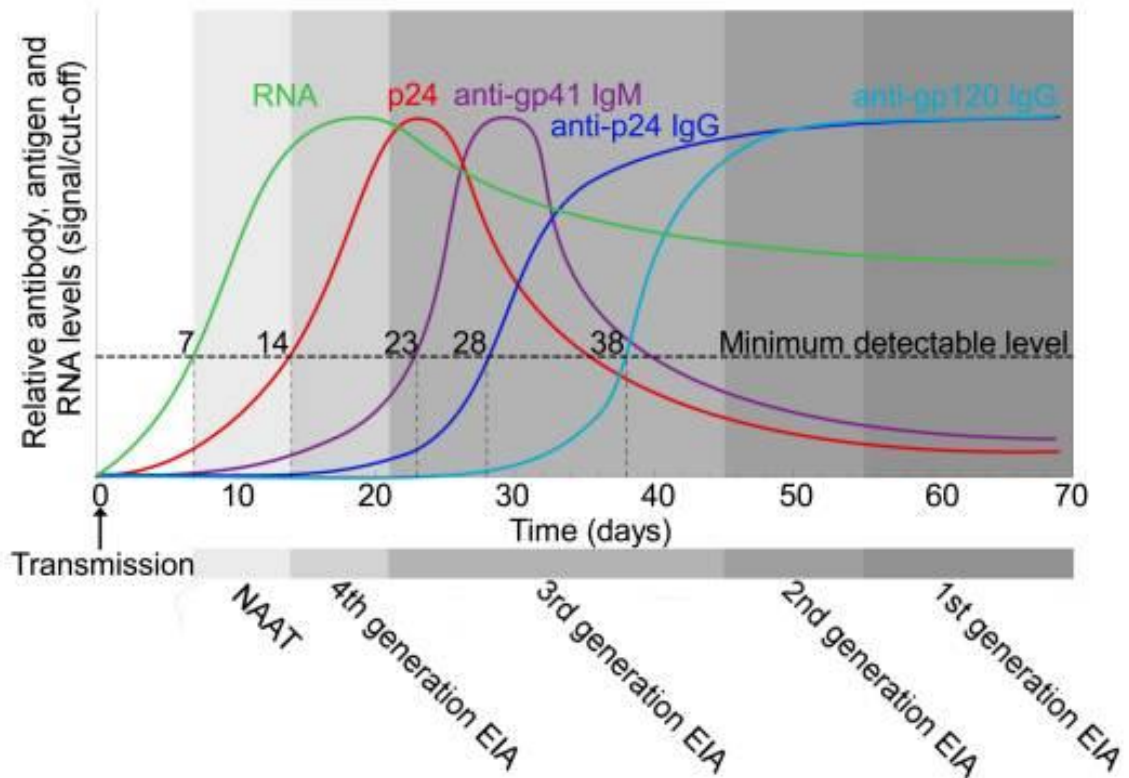


Figure 5. Viral RNA, antigen and antibody response levels through the timeframes after transmission and diagnostic test generations according to their earliest applicable timeframes (Gray et al., 2018).

1.1.4. HIV-1 and HIV-2 Comparison

From the infectivity perspective, HIV-2 is less infective and virulent than HIV-1, which can explain the low viral load of HIV-2. HIV-2 is mostly restricted to the western part of Africa. However, some cases exist in Europe, the United States of America, and India. HIV-2 has nine groups, A to I, but only groups A and B are circulating (Bbosa et al., 2019; Visseaux et al., 2016).

1.1.5. HIV-1 Capsid Protein

Capsid is a structural protein of HIV. Before maturation, it is a part of the Gag polyprotein, which upon maturation, forms a capsid cage around the viral genome (Figure 6). It participates in several critical functions during the viral lifecycle, such as reverse transcription, cytoplasmic transport, nuclear entry, and viral maturation. It interacts with more than 20 host factors for successful infection.

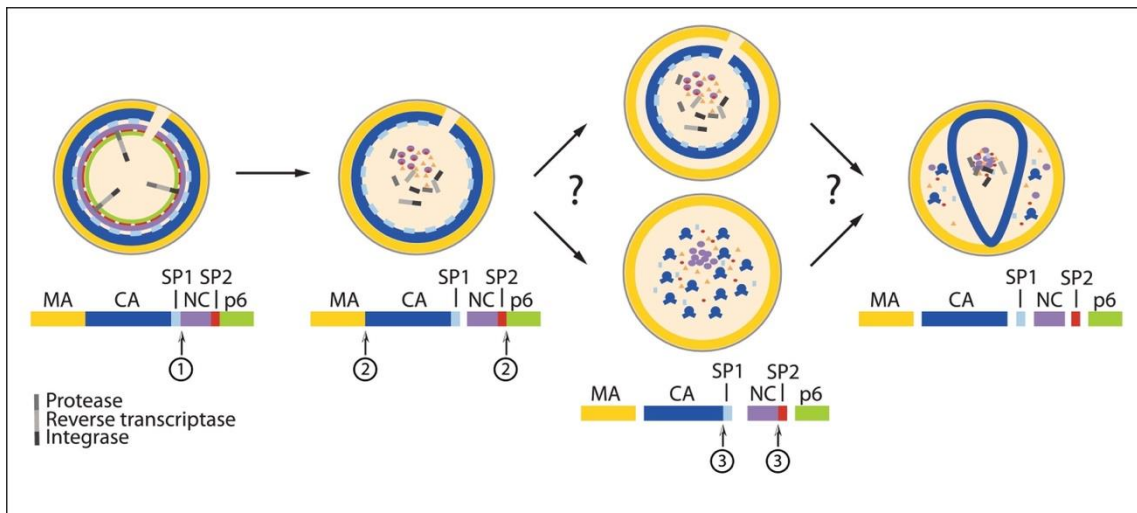


Figure 6. Maturation of HIV-1 through sequential cleavage of Gag polyprotein (Perilla et al., 2021).

HIV capsid protein is composed of two domains that fold independently, N-terminal (NTD) and C-terminal (CTD), connected with a flexible linker domain. The capsid's three-dimensional (3D) structure is dominated by α -helices, seven in NTD and

four in CTD, and a cyclophilin A (CypA)-binding loop at NTD. Capsid monomers form hexameric and pentameric structures to form an enclosed capsid core. The formation of the core is mediated by interactions at two-, three-, and sixfold symmetry regions of Capsid multimers. The stability and timing of these interactions are critical for capsid to fulfill its purpose in the viral lifecycle. An increase or decrease of stability in these interactions may res-infectious virus. A capsid cage is mostly composed of hexameric structures, while pentameric structures are located at the curvature regions to form a closed capsid core (Figure 7). HIV-2 Capsid is differentiated by its polymerization properties and thermal stability from HIV-1 Capsid. However, we will not go into detail about this comparison in this study (McFadden et al., 2021; Miyazaki et al., 2017; Perilla et al., 2021).

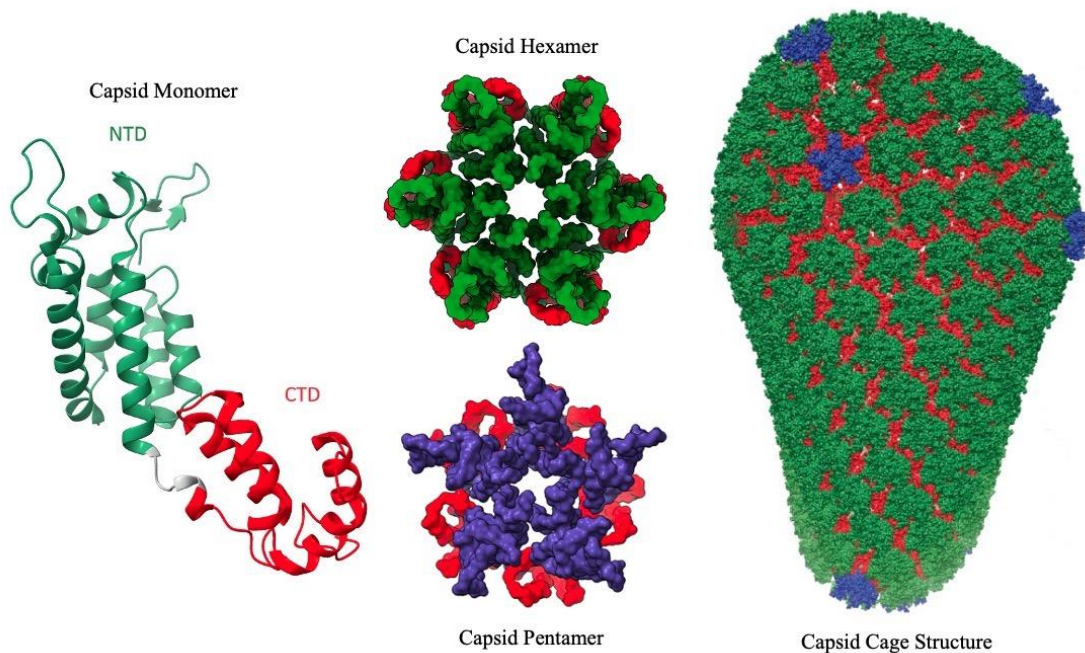


Figure 7. HIV Capsid structures. Capsid monomer PDBID: 6WAP, Capsid Hexamer PDBID: 5MCX, Capsid Pentamer PDBID: 5MCY (Deshmukh et al., 2013).

As mentioned earlier, HIV replicates via RNA, exhibiting high mutation rates. In general, high mutation rates could benefit survival against immunological pressure. However, capsid protein must maintain functional roles in viral assembly, maturation, uncoating, and nuclear import. Therefore, HIV capsid lacks the genetic robustness required to maintain a high mutation rate (Rihn et al., 2013b). This results in high amino acid sequence conservation rates in the capsid and makes this protein promising for detecting different variants of HIV. Conservation rates of the capsid protein will be further discussed in Chapter 3.1.

1.2. Nanobodies

Nanobodies are derived from heavy-chain only antibodies of camelid family animals. They are also referred to as single-domain antibodies (sdAbs) or variable heavy chain of heavy domain (VHH). Derivation of a nanobody is mostly achieved through immunized animals. However, there are also ongoing synthetic development studies (Bao et al., 2021; Mitchell & Colwell, 2018b; Valdés-Tresanco et al., 2022).

The importance of nanobodies lies not only in their smaller size but they can also achieve nanomolar affinities to their cognate antigen. Nanobodies are also easier to produce; their hydrophobic nature, stability, and resistance to reducing environments allow them to be produced in different environments, such as bacteria, yeast, or mammalian cells. With the provided benefits, nanobody-related studies are increasing every year (Figure 8) (Bao et al., 2021; Mitchell & Colwell, 2018b; Valdés-Tresanco et al., 2022).

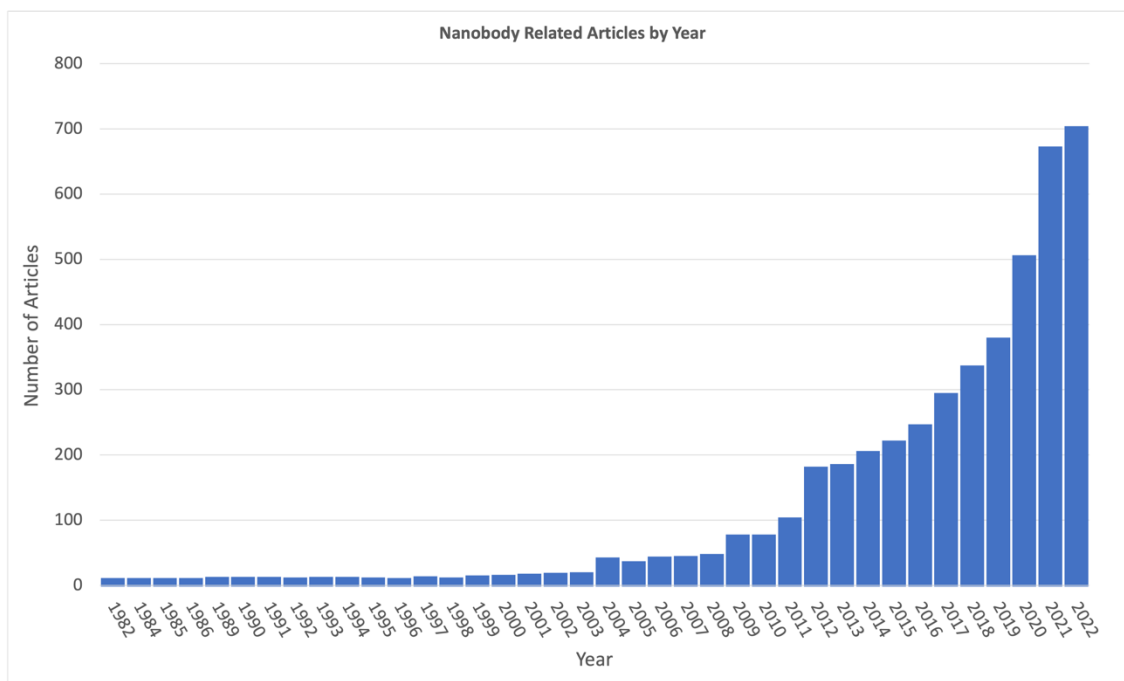


Figure 8. Number of articles per year from “nanobody” keyword search of PubMed.

1.2.1. Nanobody Structure

Nanobodies are composed of the VHH domain of Camelid heavy chain antibodies and do not contain the constant chains (Figure 9). Nanobodies have a globular fold with a framework/skeleton region and complementarity-determining regions (CDRs), the main antigen-recognition regions (Figure 10). Across the nanobody sequences, framework/skeleton regions are highly conserved, whereas variable CDR regions show less conservation, CDR3 being the least conserved and generally longest among them (Figure 11). They can preserve their affinity towards their antigen despite being much smaller, 15kDa, compared to an antibody. Being smaller in size and having longer CDR3

allows nanobodies to reach the surfaces an antibody may not reach (Mitchell & Colwell, 2018b).

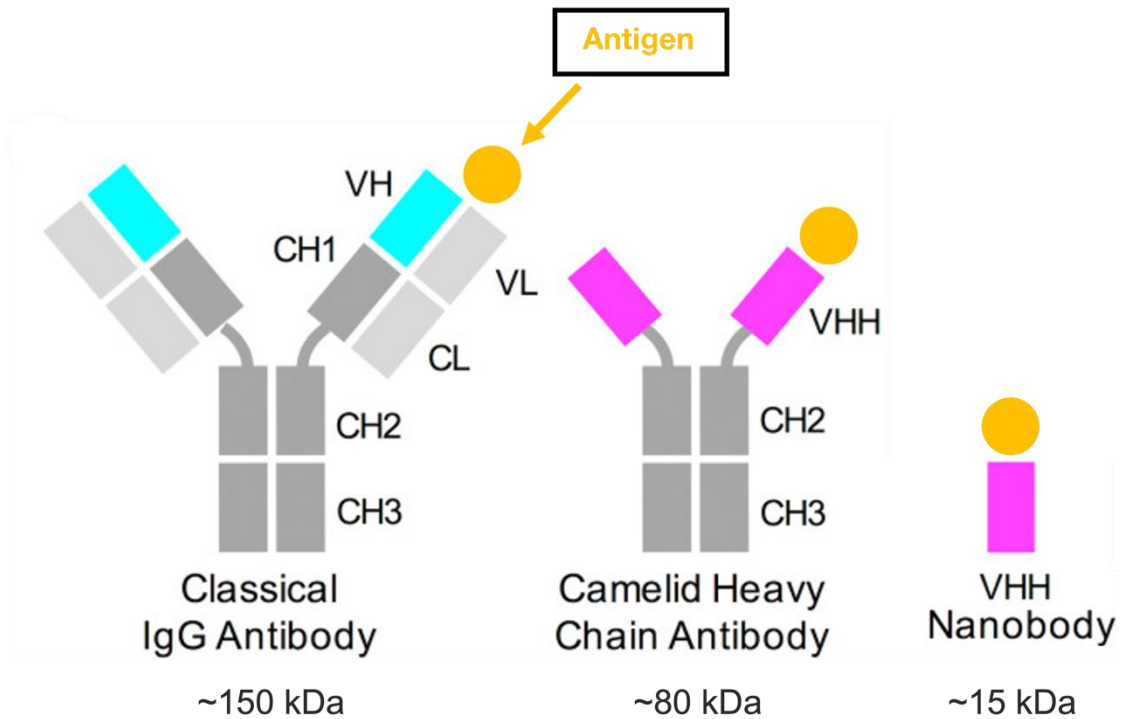


Figure 9. Comparative representation of classical IgG antibody, Camelid heavy chain antibody, and nanobody. The figure is adapted from the 2018 study by Mitchell & Colwell. Variable heavy chain (VH) is light blue, variable light chain (VL) and constant light chain (CL) is light grey, constant heavy chains (CH) are dark grey, variable heavy chain of heavy domains are magenta, and antigen is yellow.

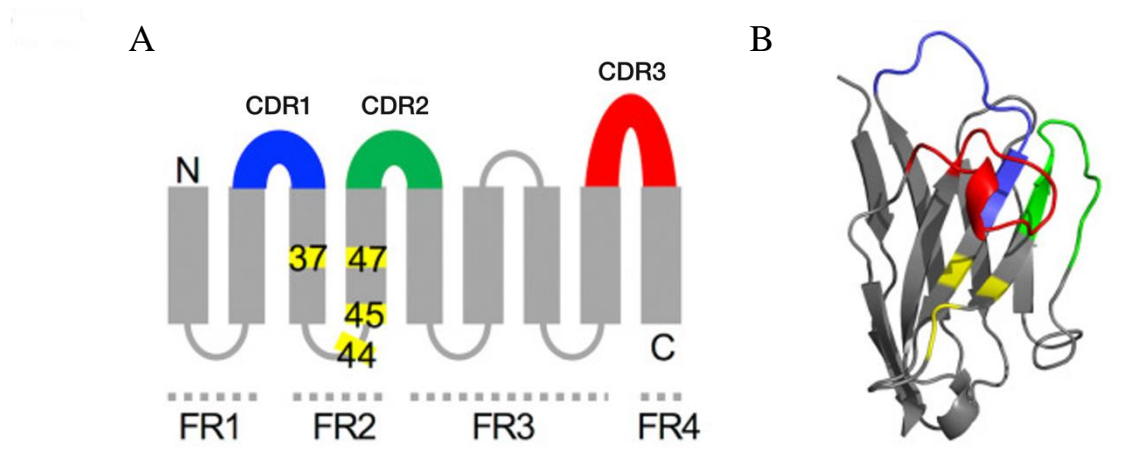


Figure 10. Representation of a nanobody structure. The figure is adapted from the 2018 study by Mitchell & Colwell. (A) The open structure of a nanobody, gray representing framework/skeleton regions, yellow representing VHH-tetrad positions, and blue, yellow, and red representing CDR regions. (B) Cartoon representation of a nanobody, colored according to (A).

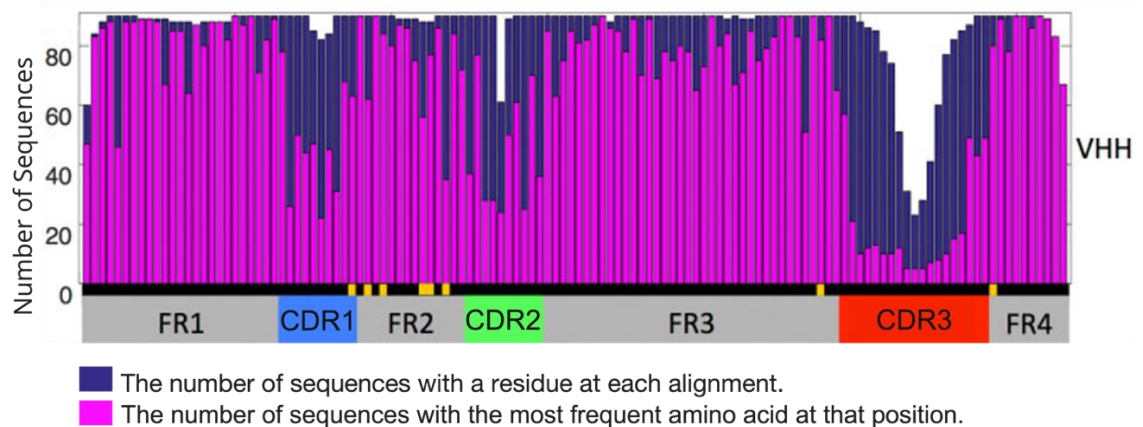


Figure 11. Sequence variability of 90 non-redundant, protein-binding nanobodies. The figure is adapted from the 2018 study by Mitchell & Colwell.

1.2.2. Applications of Nanobodies

Nanobodies achieved recognition in applications as both diagnostic tools and therapeutic agents since their discovery. Nanobodies have applications as therapeutic agents for targeting, inhibiting, tumor imaging, and diagnosis. Their small size, stability, specificity, and solubility provide essential advantages.

Monoclonal antibodies have important therapeutic applications; however, they are held back by their size and tissue penetration capabilities. As an alternative, nanobodies have become more prominent, especially in cancer therapy. Nanobodies have better tissue penetration capabilities than monoclonal antibodies, allowing nanobodies to reach the target tissue.

High specificity and tissue penetration abilities of nanobodies can be used to guide encapsulated cytotoxic drugs to tumor tissue to improve the drug's efficacy. Bivalent or bispecific nanobodies can be developed to improve binding affinity and specificity, leading to improved therapeutic capacity for carried therapeutics.

The small size of the nanobodies can be disadvantageous since they can be quickly secreted from the body via the kidney. The impact of this can allow early imaging of non-kidney lesions since the non-bonded nanobodies are quickly secreted from the body, it can reduce the background signal intensity and the toxicity for tumor imaging applications (Hu et al., 2017; Moradi-Kalbolandi et al., 2020; Sun et al., 2021; C. Wang et al., 2018).

Besides their applications in drug delivery and imaging, nanobodies can also be used for treatment. A bivalent single-domain antibody, caplacizumab, was approved for treating thrombotic thrombocytopenic purpura (TTP) and thrombosis by the FDA in 2019. This drug is the first to be approved for this disease and the first FDA-approved domain antibody (Morrison, 2019). During the COVID-19 pandemic, several nanobodies were proposed against SARS-CoV-2 (Raybould et al., 2021).

In diagnostic applications, nanobodies can be used as detector and capture agents. In the 2016 study of Doerflinger, S. Y and their team, previously characterized nanobodies used for the development of Nanobody-Based Lateral Flow Immunoassay to detect human norovirus. In their studies, Nanobody-based lateral flow immunoassay achieved 80% sensitivity and 86% specificity for norovirus and norovirus-like particles (Doerflinger et al., 2016). In another study by Helma, J. and their team in 2012, a nanobody (CANTDcb1) was developed by immunizing an alpaca with purified HIV-1 capsid protein. In this study, CANTDcb1 and HIV-1 capsid protein were co-expressed in HeLa-Kyoto cells, and co-localization was observed. In their experiments, nanobody achieved a K_D value of 0,16 nM (Helma et al., 2012).

1.3. Computational Methods for Understanding Protein-Protein Interactions

1.3.1. Protein Structure Prediction

Computational prediction of protein structures carries extreme importance when a 3D crystal structure is not available, yet the amino acid sequence of the protein is available. To address the need for a 3D structure in these situations, several homology modeling programs have been developed, tested, and improved. Briefly, these programs can be classified as comparative methods such as SwissModel (Bienert et al., 2017; Guex et al., 2009; Waterhouse et al., 2018), deep learning-based methods such as AlphaFold2 (Jumper et al., 2021; Varadi et al., 2021), and hybrid methods deep-learning or de novo prediction based programs with comparative methods such as trRosetta (Du et al., 2021; Su et al., 2021; W. Wang et al., 2022). Through such programs, predicting and evaluating a protein structure with provided structure assessment tests is possible. Details of these

programs and structure assessment tests are explained in more detail in Section 2.3. Through such programs, predicting and evaluating a protein structure with provided structure assessment tests is possible. Details of these programs and structure assessment tests will be explained in more detail in Section 2.3.

1.3.2. Molecular Docking

Molecular docking is a convenient way to predict possible complex structures using structural information. In molecular docking, it is possible to predict protein-protein, protein-ligand, protein-peptide, and protein-nucleic acid complexes. During docking, provided structures can be rigid or provided with a certain flexibility to improve the accuracy of the resulting complexes. However, as the degrees of flexibility provided to the structures increase, the required computational cost and time also increase. In some docking programs, it is possible to introduce side-chain flexibility to the complexes in post-processing (Lohning et al., 2017). The molecular docking details will be explained in more detail in Section 2.3.

1.3.3. Molecular Dynamics

Molecular dynamics (MD) simulation is a computational method to analyze flexible molecular systems at the atomic scale as a function of time. Systems are placed in an ensemble in these simulations to simulate the desired environment. In MD, the movement of the atoms is computed by a predefined force field (Salmaso & Moro, 2018).

It is the computationally most demanding and realistic method in terms of protein-protein interaction that we will mention in this study. MD simulations are going to be explained in more detail in Section 2.3.

1.4. Aim of the Study

In this study, we aim to understand the molecular details of the interaction of HIV-1 Capsid and the nanobody CANTDcb1 discovered in the 2012 study by Helma J. and colleagues. First, the 3D structure model of CANTDcb1 is generated through comparative and deep learning-based methods. Next, CANTDcb1 models and known HIV-1 capsid structures are used for molecular docking. After evaluating the docking results, the best complex structures are further tested with Molecular Dynamics simulations. Successful identification of the interacting residues would be beneficial to understand the diagnostic capabilities of the CANTDcb1. According to the conservation rates of these residues, we can propose if the CANTDcb1 can be applicable for diagnosing different subtypes of HIV-1.

CHAPTER 2

MATERIALS AND METHODS

2.1. Data Collection

Computational methods discussed in this study are improving with the help of vast amounts of data collected daily. Although these methods, such as homology modeling and molecular docking, are improving in accuracy, they should not be the only source of evaluation. To evaluate the results of these methods, we must have access to high-quality data and evaluate the results accordingly. Therefore, data collection in this study is focused on the sequence and structural data of HIV capsid proteins and nanobodies to evaluate the modeling results of CANTDcb1 and docked complexes of HIV-1 capsid and nanobody proteins.

2.1.1. Protein Data Bank

Protein Data Bank (PDB) is an archive to store, organize, and share 3D structure data of biological macromolecules, primarily protein. The majority of the deposited structure data is created by experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (Berman et al., 2000). However, it is also updated to provide structural data of computationally generated

structures. As of July 2023, >200.000 experimental and >1.000.000 computational structures are freely accessible in PDB. In this study, structures with PDB IDs 4XFX and 3J3Y were sourced from PDB.

2.1.2. The Single Domain Antibody Database

The Single Domain Antibody Database (sdAb-DB) is a source to acquire sequence and CDR data of the nanobodies with the related research papers. Its goal is to enable the use and sharing of existing nanobodies. Data in sdAb-DB are gathered manually through protein databases such as PDB and NCBI, published literature, and user submissions. (Wilton et al., 2018)

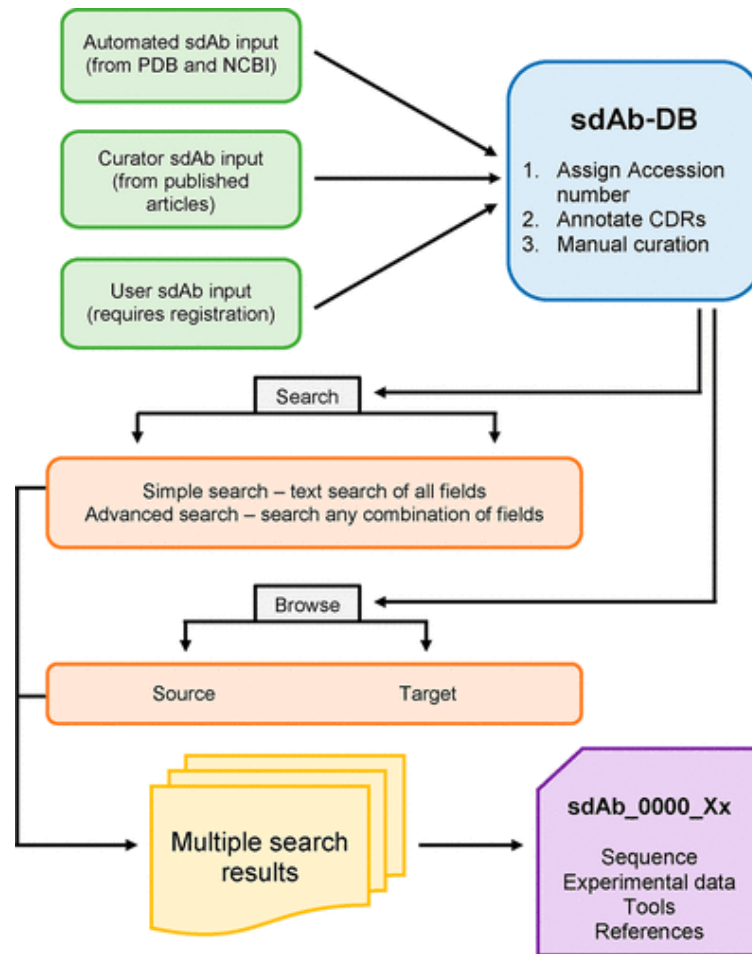


Figure 12. sdAb-DB submission and accession workflow (Wilton et al., 2018).

2.1.3. Los Alamos National Laboratory

Los Alamos National Laboratory hosts an extensive sequence library of HIV. At the moment of the 2021 HIV Sequence Compendium, it hosts more than a million sequences in their database. Their sequence database is accessible through their web page under HIV Sequence Alignments. In the sequence alignment, available data can be

filtered according to alignment type, organism, region, subtype, DNA/Protein, and year. The filtered results are accessible in 24 formats, including FASTA (Apetrei et al., 2021). In this study, 414 HIV-1 capsid and 80 HIV-2 capsid sequences were obtained from Los Alamos National Laboratory HIV Sequence Database.

2.2. Multiple Sequence Alignment of HIV-1 Capsid

HIV capsid sequences obtained from Los Alamos National Laboratory HIV Sequence Database were aligned on the UGene alignment program. The MUSCLE alignment method was used for this alignment to analyze per-residue conservation. The same sequences were again aligned in UCSF Chimera molecular visualization program to visualize them on the 3D structure of HIV-1 capsid with PDBID of 4XFX.

2.3. Modelling the 3D Structure of the Nanobody

The 3D structure of the nanobody developed by Helma J. and their team in 2012 remains unsolved. However, the amino acid sequence of CANTDcb1 is available to us. In our study, we employed several structure prediction programs to model CANTDcb1, including comparative methods to deep learning-based methods.

2.3.1. Homology Modelling

Homology modeling is a computational method used in this study to model the 3D structure of the CANTDcb1 through the amino acid structure. These methods can be further classified as comparative, deep learning-based, and hybrid methods. These methods are evaluated at Critical Assessment of Structure Prediction (CASP) every two years. In this section, we will further discuss homology modeling methods.

2.3.1.1. Comparative Methods

The main idea behind comparative methods is to use available 3D structures with some percent of sequence identity to the desired sequence, also named templates. The reliability of these models is directly proportional to the percent identity and the number of available templates. Also, there can be a problem with different sequences sharing the same 3D structure, but fold-recognition technologies can help overcome this issue. These methods generally follow four steps that can be iteratively repeated if required: template selection, target, and template alignment, building the model, and evaluating the built model. We have selected SwissModel as our comparative homology modeling method in this study. Therefore, the steps will be explained according to how SwissModel works. (Lohning et al., 2017; Waterhouse et al., 2018).

The initial step in these methods is the search for templates. In SwissModel, this can be achieved through three different modes depending on the difficulty of the modeling task: Automated Mode, Alignment Mode, and Project Mode. In Automated Mode, user can simply supply their amino acid sequence and let the system select appropriate templates based on BLAST and HHblits for the modeling task. Users can upload their

target-template alignment in FASTA or Clustal format in Alignment Mode. This method is usually used when the target templates are already known and available. In Project Mode, users have complete control over modeling parameters. This way, users have more control over the modeled structure to improve quality. After the template search, these templates are ranked according to Global Model Quality Estimate (GMQE) and Quaternary Structure Quality Estimate (QSQE). After ranking the templates, they can be selected by the user (Bertoni et al., n.d.; Biasini et al., 2014; Guex et al., 2009; Waterhouse et al., 2018).

After the selection of templates, the modeling of the structure begins by transferring conserved residues in the alignment. Then the backbone of the remaining residues is modeled by loop modeling, followed by constructing non-conserved residues' side chains to build a full-atom model of the target structure. Swiss-Model uses the computational structural biology framework of OpenStructure (Biasini et al., 2013) and ProMod3 in their application to build the model. After the model is built, model quality is analyzed. We will discuss this concept later in Section 2.3.2 (Waterhouse et al., 2018).

2.3.1.2. Deep Learning-Based Methods

Comparative methods fall short when there are no available templates. Therefore, deep learning-based methods are being developed to answer the continuing quest of how proteins fold.

One of the programs that we used in this category is AlphaFold2. AlphaFold2 utilizes neural network architectures with evolutionary data and physical and geometric constraints of protein structures to train these networks. These neural networks are trained to recognize patterns and relationships between amino acids in 3D protein structures. AlphaFold2 network has two main stages. The first stage is named Evoformer block, which processes multiple sequence alignments (MSAs). Followed by the structure

module, it introduces an explicit 3D structure, which is represented by rotations and translations. Details of this architecture can be seen in Figures 13 and 14 (Jumper et al., 2021; Varadi et al., 2021).

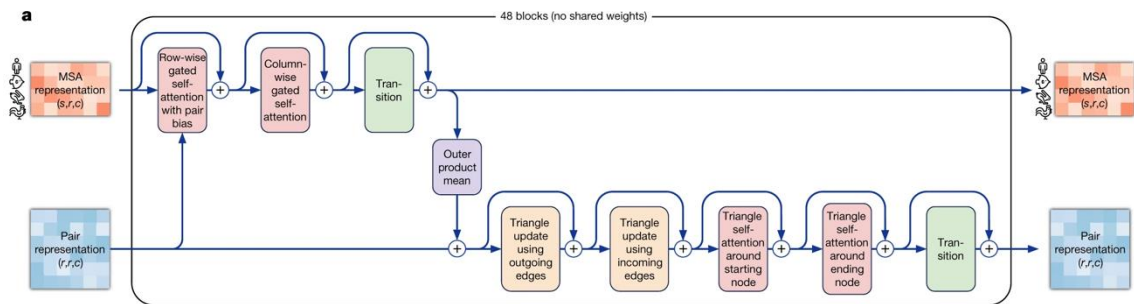


Figure 13. Evaformer block in AlphaFold2 pipeline (Jumper et al., 2021).

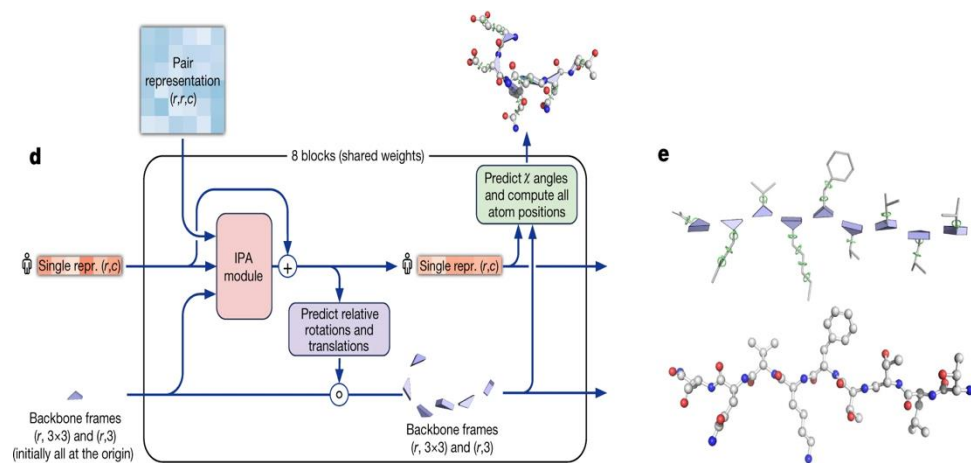


Figure 14. Structure module in AlphaFold2 (Jumper et al., 2021).

Another deep learning-based method utilized in this study is roseTTAFold. This model was served under the Robetta name server during the structure prediction phase of this study). This method is inspired by AlphaFold2, but the module related to the 3D structure building works in parallel to other layers. This approach provides active communication between 1D amino acid sequences, 2D distance map, and 3D coordinate information. Information from these three layers is then combined to produce a 3D structure. (Baek et al., 2021)

2.3.1.3. Hybrid Methods

Hybrid methods used in this study combine knowledge-based approaches with deep learning-based structure prediction algorithms. In our study, we utilized transform-restrained Rosetta (trRosetta). This method comprises two steps to predict a structure. First, the distance and orientations of inter-residue geometries are predicted with a deep neural network. The features in this prediction are derived from a generated multiple sequence alignment, including per-residue and inter-residue properties. If the homologs of the target protein are available, optional parameters can be used as additional inputs. The predicted geometries guide the structure prediction process using direct energy minimization within the Rosetta framework (Du et al., 2021; Su et al., 2021; W. Wang et al., 2022).

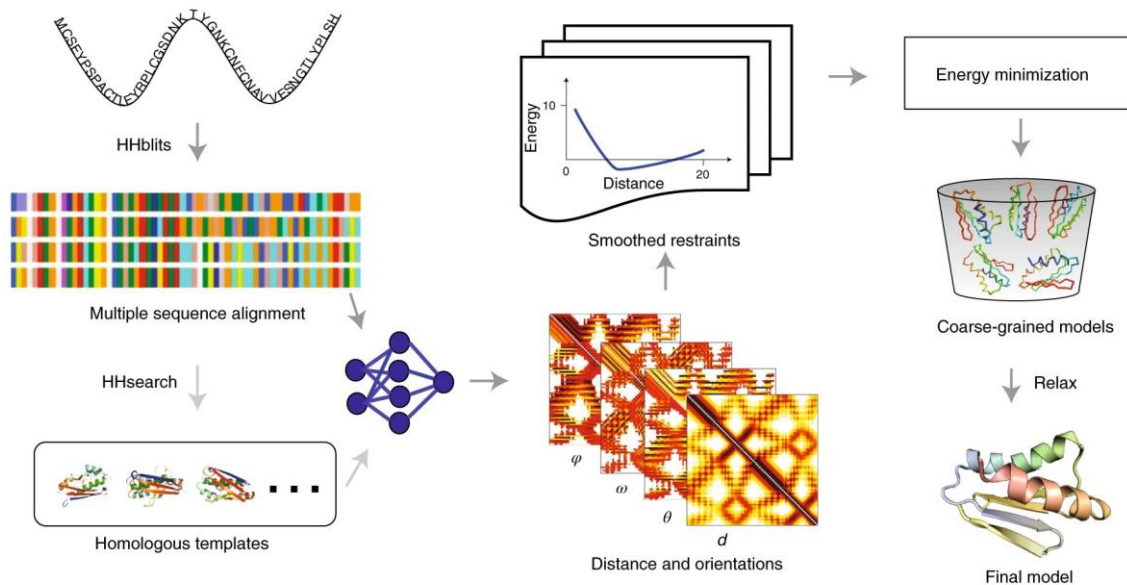


Figure 15. The trRosetta protocol is visualized in a flowchart, outlining the sequential steps involved in the process (Du et al., 2021).

2.3.2. Structure Assessment Test

Structure assessment tests help evaluate the modeled structures' local and global quality. These tests can provide significant insight into how good our modeled structure is and, therefore, explicitly applied to our produced models. The structure assessment tests in this study were done on the SwissModel web service. These tests include the Ramachandran plot, MolProbity evaluation, local distance difference test (IDDT), Qualitative Model Energy Analysis (QMEAN), and QMEAN extended with distance constraints (QMEANDisCo).

2.3.2.1. Ramachandran Plot

All amino acids except N-terminal and C-terminal amino acids in a protein chain have dihedral angles called Φ and Ψ angles. Although these values can be between -180° and 180° , many are impossible due to steric interference (Nelson & Cox, 2017).

A Ramachandran plot can visualize energetically preferred regions for backbone dihedral angles against amino acid residues in a protein structure. An empty Ramachandran plot can be seen in Figure 15. This plot does not include the Proline or Glycine due to their atypical structures. Plots representing only Glycine or only Proline are available if needed. Every amino acid in the protein chain, except N-terminal and C-terminal, will be represented with a local QMEANDisCo score, which we will mention in Chapter 2.3.2.4 on the Ramachandran plot.

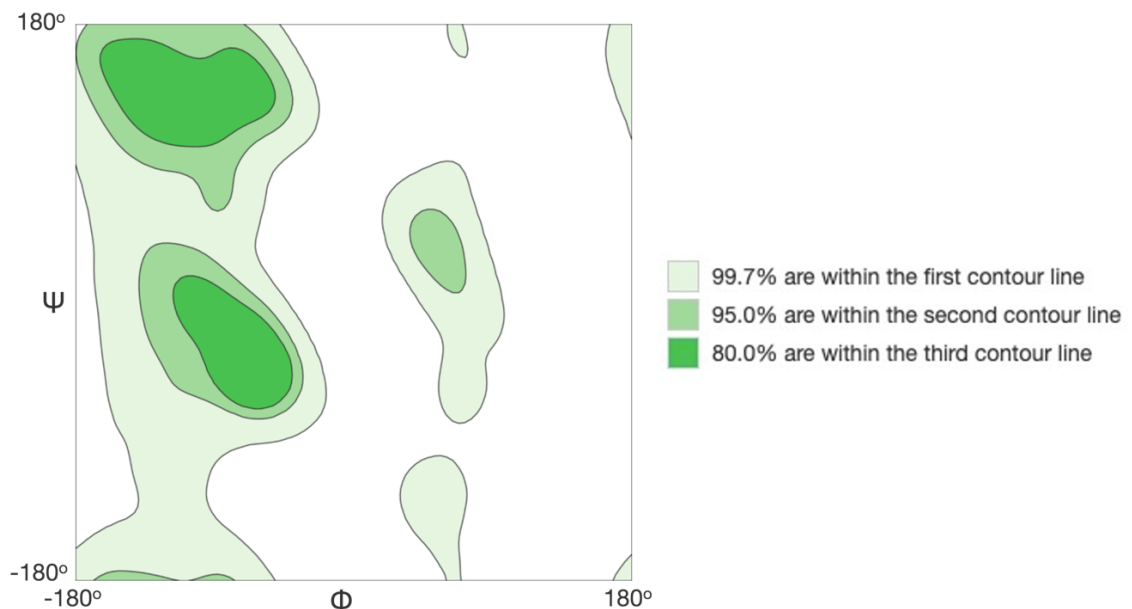
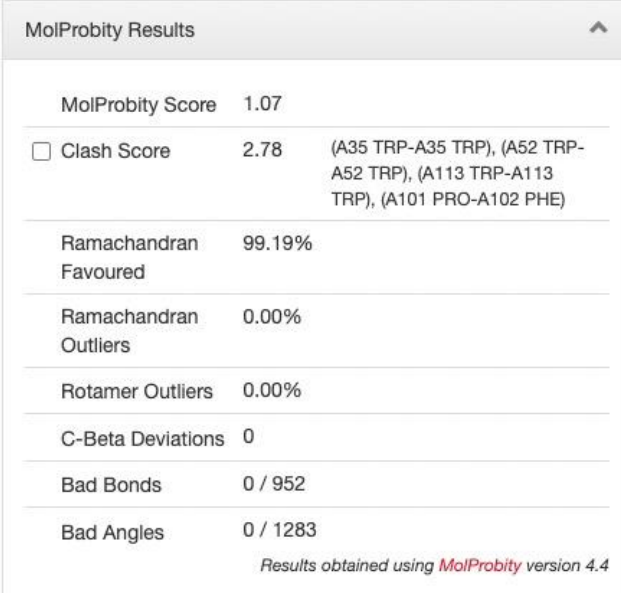


Figure 16. An empty general (No Proline or Glycine) Ramachandran plot from SwissModel with favored regions colored in tones of green.

2.3.2.2. MolProbity Evaluation

MolProbity is another integration in SwissModel's structure assessment tests. It is used to analyze clashes, Ramachandran distribution, and bonds automatically. It first calculates the Clashscore, the number of atom-atom overlapping $\geq 0.4\text{\AA}$ per thousand atoms; in an ideal case, it should be 0. After clashes, MolProbity analyzes the sidechain rotamers; in an ideal case, it should be $<1\%$. Then, MolProbity analyzes the Ramachandran plot for each amino acid for whether it is in a favored position or an outlier. In an ideal case, Ramachandran favored percentage should be $>98\%$, and Ramachandran outliers should be $<0.2\%$. Then, through a weighed calculation on clashes, Ramachandran favored percentage and rotamer outliers, it calculated a MolProbity score. This score helps compare structures relative to each other. It does not indicate an absolute measurement of quality. However, it should be as low as possible (Williams et al., 2018).



MolProbity Results		
MolProbity Score	1.07	
<input type="checkbox"/> Clash Score	2.78	(A35 TRP-A35 TRP), (A52 TRP-A52 TRP), (A113 TRP-A113 TRP), (A101 PRO-A102 PHE)
Ramachandran Favoured	99.19%	
Ramachandran Outliers	0.00%	
Rotamer Outliers	0.00%	
C-Beta Deviations	0	
Bad Bonds	0 / 952	
Bad Angles	0 / 1283	

Results obtained using MolProbity version 4.4

Figure 17. Example output of a MolProbity analysis.

2.3.2.3. Local Distance Difference Test (IDDT)

A model can be evaluated through several global methods, such as root mean square deviation (RMSD) or Global Distance Test (GDT). These global tests have several shortcomings; for example, in RMSD, outliers can dominate the score, be insensitive to missing residues, or be unable to consider flexible domains that can change their orientations naturally, as seen in Figure 17. These shortcomings brought the need for local evaluation techniques that can overcome the flexible domain issues. IDDT score is one way to overcome these issues because it can highlight low-quality regions in the model, independent of domain movements. By comparing the separations between corresponding atoms in the predicted and actual protein structures, the IDDT method evaluates the precision of protein structure predictions. It offers a numerical evaluation between 0-100, 100 meaning identical to the reference structure, of the accuracy of the local prediction (Mariani et al., 2013). In this study, IDDT is not used directly but is a supplement to the QMEANDisCo quality estimate.

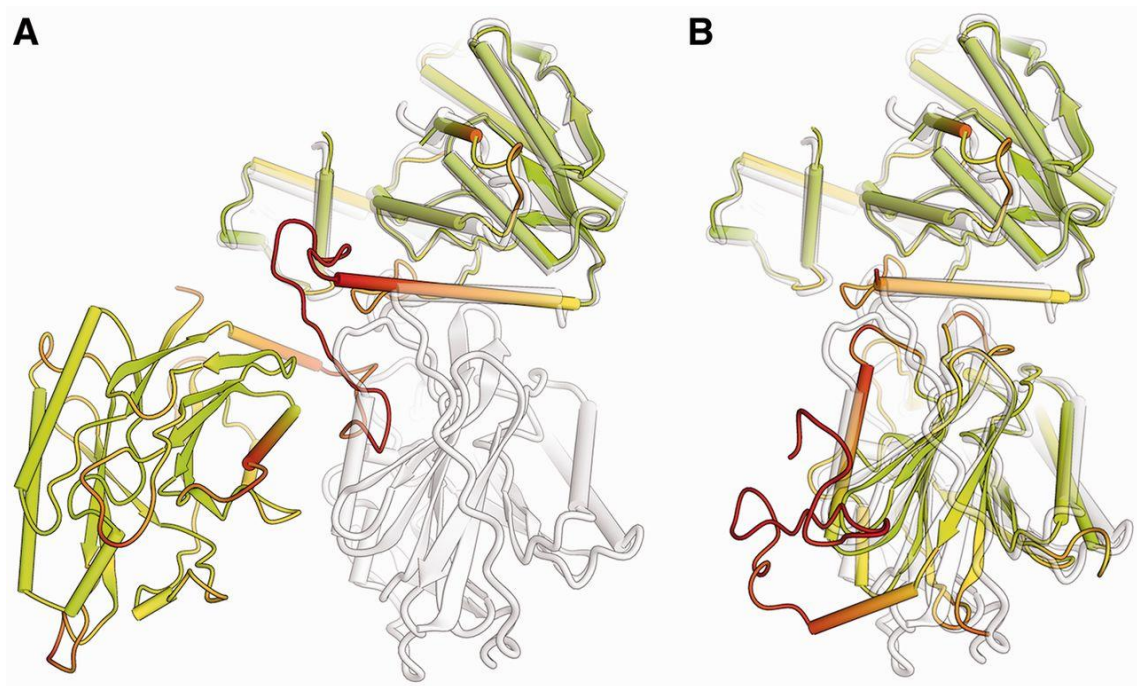


Figure 18. CASP target T0542 reference model and predicted structure model comparison. (A) The model is predicted as full-length, and the first domain is superimposed on the target. (B) Two domains are modeled separately and superposed individually to the target structure. Structures in both panels are colored according to full-length IDDT scores, with green indicating high and red indicating low IDDT scores in the spectrum (Mariani et al., 2013).

2.3.2.4. Quality Estimate: QMEAN and QMEANDisCo

Qualitative Model Energy Analysis (QMEAN) is a composite scoring function for assessing protein structure quality. This scoring function provides us with information about the “degree of nativeness” of the protein. It is calculated through five different structural properties. These properties define the protein and its residues in terms of their

environment, interactions with neighboring residues, and solvent accessibility while considering the secondary structure these residues are in. Depending on the sequence identity, the protein is compared to the reference structures the function identifies. In QMEAN, energies associated with certain interactions in protein structure are calculated from statistical analysis of known protein structures. The function employs a machine-learning model trained from known protein structures for the weighted evaluation of different structural features and energy terms based on their importance in determining protein quality. A combination of these scores and weight is then used to assign a QMEAN score to assign a quality score. The QMEAN score is normalized according to the number of interactions in the structure to make it independent of the structure size. Then the model quality estimates are expressed as z-scores and compared to available crystal structures (Figure 18) (Benkert et al., 2008, 2011).

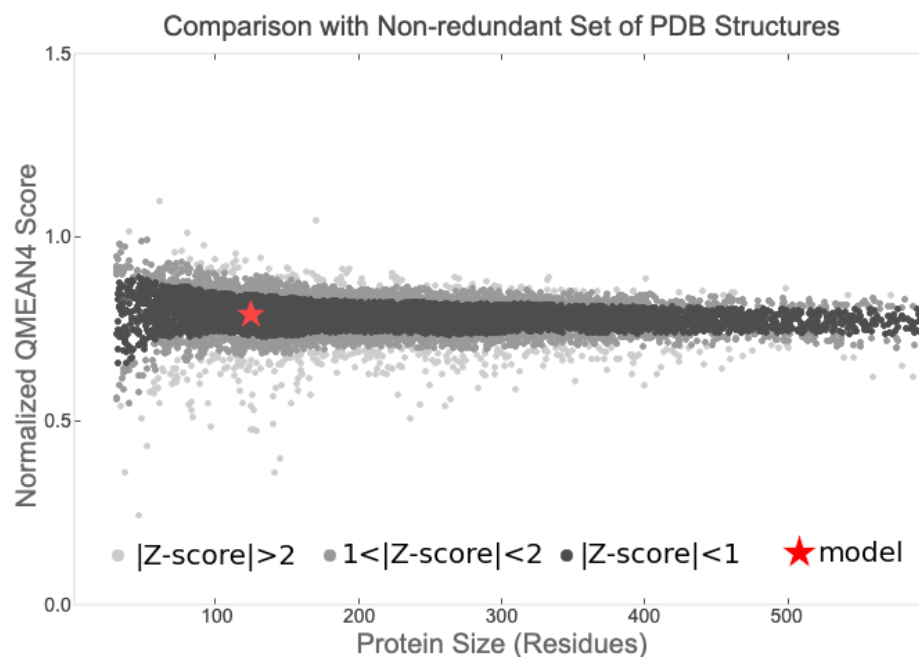


Figure 19. Normalized QMEAN scores are expressed as z-scores in comparison with available crystal structures. The graph shows the size of proteins on the horizontal axis. On the vertical axis, we have the "normalized QMEAN"

score, which tells us how good the protein structure is. Each dot on the graph represents one real protein structure that has been experimentally determined. The black dots represent crystal structures that have a "QMEAN" score within 1 standard deviation of the average score. The grey dots represent structures that have a "QMEAN" score that is between 1 and 2 standard deviations away from the average. The light grey dots represent structures that are even further from the average. The red star represents the model we are interested in. We want to compare how well the model matches the real structures (Benkert et al., 2008).

QMEANDisCo is an extended method of the QMEAN function with consensus-based distance constraints (DisCo) score. It extends QMEAN by introducing pairwise distances from homologous structures to target protein. While QMEAN focuses on the global evaluation of the structure, QMEANDisCo focuses on residue-based evaluation. The average of these local quality estimates represents the QMEANDisCo global score. In this method, an IDDT score with the range [0.0, 1.0] is used, which is predicted from a trained deep learning model. According to the test results, models with a local and global IDDT score >0.6 can be classified as correct (Studer et al., 2020).

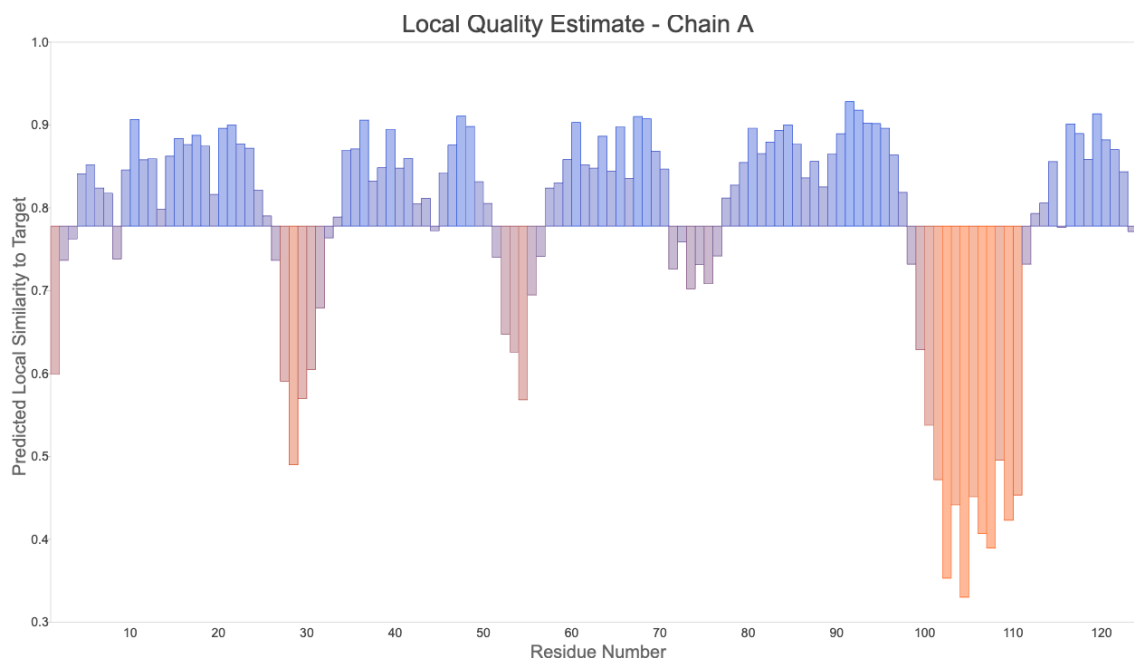


Figure 20. An example of results of QMEANDisCo evaluation. The horizontal axis represents the residue number in the target protein. The vertical axis represents the per residue IDDT score. The baseline for the bars is the global quality estimate score (Studer et al., 2020).

2.4. Determining Protein-Protein Complex Orientation

Knowing the protein-protein complex orientation is crucial for understanding molecular interactions, structural and functional characterization, predicting protein function and behavior, and developing therapeutics. Determining these complexes is possible through both experimental and computational methods. In this section, we will discuss a widely used computational method termed molecular docking (Kozakov et al., 2017).

2.4.1. Molecular Docking

Determining a protein-protein complex structure can be difficult even when we have structures of these proteins are determined experimentally. Accounting for the flexibility and dynamics of complexes can be challenging for experimental methods, and these methods are costly in terms of time and resources. Molecular docking techniques are developed to overcome these challenges and aim for accuracies close to experimental methods. These methods are being benchmarked and improved continuously. Although they are not able to produce the same native complex at once, they can provide us with several possible complexes which have a high probability of containing a native-like complex. Molecular docking methods can be simplified into three steps: (1) the proteins of interest are placed in a grid, and the grid represents a collection of specific points in three-dimensional space that serve as potential locations for placing the ligand during the docking procedure; (2) the ligand is docked into the binding site by exploring various conformations through rotations, translations, and torsional changes within the grid; (3) a scoring function evaluates each conformation, this scoring can be different for each method. Depending on the approach, some docking methods can also include pre- and post-processing steps. (Kozakov et al., 2017; Kurkcuoglu & Bonvin, 2020).

Molecular docking techniques can be discussed under an important property, the flexibility of the structures. Depending on whether there are any flexibilities in structures, we can separate them as rigid-body docking and flexible docking methods. This is not an absolute separation since, in some methods, the structures have no flexibility. In contrast, in others, their structures can have a certain degree of flexibility without being absolutely flexible (Lohning et al., 2017).

In this study, we utilized three docking methods: ZDock, Haddock, and ClusPro. ZDOCK is predominantly classified as a rigid docking method, systematically exploring fixed orientations and translating two protein structures to determine potential binding configurations. It does not explicitly incorporate conformational changes or flexibility; however, it can be combined with supplementary techniques to introduce limited

flexibility or refinement (Pierce et al., 2014). HADDOCK is a docking method that accounts for flexibility and conformational changes in both the receptor and ligand during docking. It utilizes experimental data to guide the docking process and refinement, enabling the exploration of diverse conformations and orientations. (Honorato et al., 2021; Van Zundert et al., 2016). ClusPro is a docking method that employs a combination of rigid-body and flexible docking techniques. It begins with rigid docking, investigating different orientations and translations. Subsequently, a refinement step is applied, allowing for little flexibility in side-chain orientations and minor backbone adjustments. This hybrid methodology improves the precision of the predicted protein-protein complexes (Desta et al., 2020; Kozakov et al., 2013, 2017; Vajda et al., 2017).

2.4.2. Evaluation of the Docked Complexes

Molecular docking methods are powerful tools to determine protein-protein complex orientations. Although these methods are continuously being improved, they produce several possible orientations, which need to be evaluated before selection. In this study, docked complexes were eliminated by docking scores, established non-covalent interactions, and visual inspection of the complexes.

2.4.2.1. Docking Scores and Clusters

In molecular docking methods, scoring functions are crucially essential since these functions are the first step of evaluation in docking. Produced complex orientations are presented to the end user according to their scores. Therefore, the quality of scoring

functions significantly impacts the method's accuracy. In some methods, scoring functions are supplemented with clustering of samples with implications that the number of occurrences of an orientation is related to the "degree of nativeness" of the orientation. In this section, we will discuss the scoring approaches of ZDock, Haddock, and ClusPro.

In ZDock scoring functions: surface complementarity (SC), desolvation free energy (DS), and electrostatics are utilized to calculate a docking score. In shape complementarity, receptor, and ligand are expressed as (l, m, n) with the dimensions of $N \times N \times N$ grid with discrete values from 1 to N . In this expression, N should be large enough to cover coordinate space but not too large to hinder the performance of the calculations. Functions R_{SC} for the receptor and L_{SC} for the ligand used to assign values according to geometric properties shown in equations 2.1 and 2.2. Here ρ is a positive number (Chen & Weng, 2002).

$$R_{SC}(l, m, n) = \begin{cases} 1, & \text{surface of R} \\ \rho i, & \text{core} \\ 0, & \text{empty space} \end{cases} \quad (\text{Equation 2.1})$$

$$L_{SC}(l, m, n) = \begin{cases} 1, & \text{surface of L} \\ \rho i, & \text{core} \\ 0, & \text{empty space} \end{cases} \quad (\text{Equation 2.2})$$

To determine whether an atom in a protein belongs to the surface or core, a computational method was utilized to calculate the solvent-accessible area. In this approach, a water probe with a radius of 1.40\AA is used. If an atom in the protein possesses a solvent-accessible area greater than 1\AA^2 , it is categorized as a surface atom. If the solvent accessible area is equal to or less than 1\AA^2 , the atom is classified as a core atom; after the classification, grid points are assigned accordingly. In this method, ρ values are assigned as 9. The 2002 study of Weng, Z., and their team explained the assignment

procedure of these grid values. Shape complementarity can be calculated with equation 2.3. Correlation between two functions can be computed with Discrete Fourier Transform (DFT) and Inverse Fourier Transform (IFT) as in equation 2.4. In these equations, "o," "p," and "q" values represent the number of grid points that determine the shift of ligand L in relation to receptor R across each dimension and represent the shape complementarity (Chen & Weng, 2002).

$$S_{SC}(o,p,q) = \text{Re} \left[\sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N R_{SC}(l,m,n) \cdot L_{SC}(l+o,m+p,n+q) \right] - \text{Im} \left[\sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N R_{SC}(l,m,n) \cdot L_{SC}(l+o,m+p,n+q) \right] \quad (\text{Equation 2.3})$$

$$S_{SC} = \text{Re} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right] - \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right] \quad (\text{Equation 2.4})$$

In this calculation, core-core contacts contribute to the result by $(\rho i)^2 = -81$, surface-core contacts contribute by $\text{Im}[\rho i] = -9$, surface-surface contacts by 1, and if there is no contact it contributes by 0. After the calculation of shape complementarity, ZDock calculates the DS. Calculations of DS are similar to the SC, however, with different values (equation 2.5), and only one DFT and two IFTs are required to compute this value (equation 2.6). Lastly, ZDock utilizes an approach based on the Coulombic formula to calculate the electrostatic energy in protein-ligand interactions. The method involved correlating the receptor's electric potential with the ligand. This method adopts the approach previously used by Gabb et al. but incorporates partial charges from the CHARMM19 potential. Additionally, to avoid non-physical receptor-core/ligand

contacts, grid points in the core of the receptor were assigned a value of 0 for the electric potential. Final scoring is calculated by the weighted sum of the results of these calculations, shown in equation 2.7. In this method, the default scaling factors were set to 0.01 for α and 0.06 for β . (Chen & Weng, 2002).

$$S_{DS}(o,p,q)=\text{Im} \left[\sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N R_{DS}(l,m,n) \cdot L_{DS}(l+o,m+p,n+q) \right] \quad (\text{Equation 2.5})$$

$$S_{DS}=\frac{1}{2} \times \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{DS}) \cdot \text{DFT}(L_{DS})) \right] \quad (\text{Equation 2.6})$$

$$S=\alpha S_{SC}+S_{DS}+\beta S_{ELEC} \quad (\text{Equation 2.7})$$

A combination of scoring terms that measure the compatibility and quality of the protein-protein complex determines the Haddock docking score. Haddock assesses the shape complementarity of interacting protein surfaces. Van der Waals interactions, which describe attractive and repulsive forces between atoms based on size and configuration, are considered in this evaluation. Electrostatic interactions, as well as charge complementarity of the complex, are considered. Haddock also considers the solvation and desolvation energies to determine the effects of the solvent. As a result, Haddock outputs several different complexes to the user. After generating multiple complex structures, Haddock performs a clustering analysis to identify distinct clusters representing different possible complex conformations. The clustering helps select the

most representative structures and provides insights into the complex's conformational variability. Resulting complexes are ordered depending on their scores as well as cluster sizes. Scripts of the individual calculation are not provided due to their length. The source code of Haddock is available on GitHub, and the scoring functions can be found at <https://github.com/haddock/haddock3>. (Dominguez et al., 2003; *HADDOCK2.4 Manual - Analysis*, 2023; Kurkcuoglu & Bonvin, 2020).

The rigid-body docking step in ClusPro utilizes PIPER, a Fast Fourier Transform (FFT) based docking program. Docking scores in ClusPro are provided directly from the PIPER program. However, it is emphasized that instead of the complexes based on the energy score provided by PIPER, the largest clusters of low-energy structures should be considered. In PIPER, the energy function is expressed by equation 2.8. E_{rep} and E_{attr} represent the repulsive and attractive components of the van der Waals interaction energy, while E_{elec} refers to the electrostatic energy term. The term E_{DARS} corresponds to a pairwise structure-based potential known as the "decoys as the reference state" (DARS), and w_1, w_2, w_3 terms describe the contribution weights. Models built with different weights in energy functions can be selected after the docking process in ClusPro. In PIPER, shape complementarity calculations are done on a 3D grid system; grid points are represented as (l, m, n) . Each grid point is evaluated through R_p and L_p functions shown in equations 2.9 and 2.10. In R_p function: $c_{l,m,n}$ represent the number of atoms that are at the attractive interaction range of the grid point, $r_{l,m,n}$ represents the atoms within the repulsive interaction range of the grid point. The L_p function evaluates whether an atom is present at that grid point. Correlation between these functions results in a shape complementarity term. For determining electrostatic interaction terms, a Generalized Born-type equation is used with constant Born radii. This allows for expressing electrostatic interactions in terms of the receptor's potential field and the ligand's electrostatic charge. The DARS method is used for determining structure-based intermolecular potentials. Details on these functions are available in provided references in detail (Desta et al., 2020; Kozakov et al., 2006, 2013, 2017; Vajda et al., 2017).

$$E = w_1 E_{\text{rep}} + w_2 E_{\text{attr}} + w_3 E_{\text{elec}} + w_4 E_{\text{DARS}} \quad (\text{Equation 2.8})$$

$$R_p(l, m, n) = -c_{l, m, n} + w_1 r_{l, m, n} \quad (\text{Equation 2.9})$$

$$L_p(l, m, n) = \begin{cases} 1 & \text{if } \&(l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases} \quad (\text{Equation 2.10})$$

2.4.2.2. Non-Covalent Interactions and Visual Analysis

In the absence of crystal structures of protein-protein complexes, we can benefit from other experimental data. This data can be used to guide our docking and evaluation of docked poses. In our case, we have several studies to help us with docking. From previous studies, we learned that CANTDcb1 targets the NTD of the HIV-1 capsid protein and does not show an inhibitory effect (Alfadhli et al., 2021; Helma et al., 2012). This information is used to eliminate complexes that target CTD of HIV-1 capsid and interactions that can disrupt the capsid core formation. Since the binding of CANTDcb1 does not disturb capsid core formation, it should not clash with other monomers in the multimeric structures of HIV-1 capsid protein. Important residues that take part in the formation of capsid core are highlighted in Figure 21 (Craveur et al., 2019; Gres et al., 2015):

A
PIVQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVG
GHQAAMQMLKETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTTSTLQE
QIGWMTHNPPIPVGEIYKRWILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRFY
KTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV

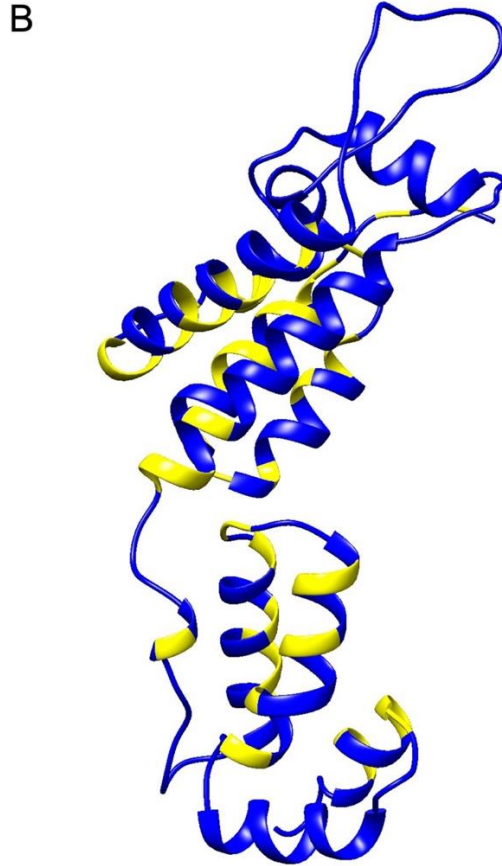


Figure 21. Important residues that take part in the formation of capsid core. (A) Residues highlighted with yellow on HIV-1 capsid sequence. (B) Residues highlighted on the 3D structure of HIV-1 capsid structure. Yellow highlighted residues are important in the formation of capsid core. The figure is prepared in UCSF Chimera.

In this study, non-covalent interactions are analyzed through PDBsum Generate, and visual analysis of docked poses is evaluated with UCSF Chimera.

2.5. Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations are computational methods that provide us with atomic-level details of the dynamics of systems. These methods are based on Newton's equations of motion. During MD simulations, coordinates, velocities, forces, and potential energies are determined at each iterative step, which is denoted by the term "time step". Lowering the time step value can increase simulation accuracy but at the cost of computational resources. Properties of the system can be divided in two: macroscopic system properties that represent volume (V), pressure (P), temperature (T), and number of atoms (N); microscopic system properties that represent velocities (v_i) and positions (r_i). MD simulations provide us with trajectories, which are time-dependent changes in the system (Zheng et al., 2018). In this section, we will briefly explain the simulation ensembles and force fields in MD.

2.5.1. Simulation Ensembles

Simulation ensembles are mainly concerned with macroscopic system properties, volume (V), pressure (P), temperature (T), and number of atoms (N). These ensembles are; canonical ensemble (NVT), microcanonical ensemble (NVE), and isothermal-isobaric ensemble (NPT). Names of the ensembles are assigned according to what is being held constant in the system. In NVT, the number of atoms, volume, and temperature of the system are constant. Energy exchange is possible for the systems in this ensemble. In NVE, the number of atoms, volume, and energy of the system are constant, representing an isolated system. In NPT, the temperature and pressure of the system are constant, representing an isothermal and isobaric system (Zheng et al., 2018). In this study,

equilibration steps are completed in an NVT ensemble, and production steps are completed in an NPT ensemble at 303.15K constant temperature.

2.5.2. Periodic Boundary Condition (PBC)

PBC is a concept to keep in mind when analyzing MD simulations. In large-scale models, it is impractical to have an infinite or too large of a system. Therefore, the simulation environment in MD simulations is expressed with one cell, replicated to surround itself infinitely, and these replicas are called periodic images. Without considering PBC, atoms at the edge of the simulation box would experience different forces than other atoms. In PBC, if an atom leaves the system from one edge of the box, it will reappear on the opposite side of the system (Yu & Dalby, 2020). MD trajectories are visualized with VMD, and VMD at its base state does not consider the PBC. To overcome this, we will utilize an MDTraj script before visualizing it in VMD.

2.5.3. Force Fields

Force fields in MD simulations are a potential energy field representing the topology and motion of atoms in the system. The molecular properties of simple molecules like water are described with a spectrum constant force field. To describe the properties of more complex molecules, empirical potential function force fields are usually utilized. For this potential energy calculation, non-binding potential, bonding stretching term potential, angle bending term potential, torsion (dihedral) angle term potential, out-of-plane bending term potential, and coulombic interaction term potential

can be utilized (Zheng et al., 2018). For the MD simulations in this study, the Charmm36m force field is used.

2.5.3.1. CharmmGUI and Charmm36m Force Field

Charmm is a commonly used classic force field. Potential energy is expressed in terms of $U(\vec{R})$, internal terms include bond (b), valence angle (θ), Urey–Bradley (UB, S), dihedral angle (φ), improper angle (x), and backbone torsional correction ($CMAF, \varphi, \psi$) contributions (Equation 2.11). The Verlet-type integrator in Charmm can be used for velocity reassignment and velocity scaling (Brooks et al., 2009; Huang et al., 2016).

$$\begin{aligned}
U(\vec{R}) = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \\
& + \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 \\
+ & \sum_{dihedrals} K_\varphi(1 + \cos(n\varphi - \delta)) + \sum_{impropers} K_\omega(\omega - \omega_0^2) \\
& + \sum_{non-bonded\ pairs} \left\{ \varepsilon_{ij}^{min} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^6 \right] \right. \\
& \left. + \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r_{ij}} \right\} + \sum_{residues} U_{CMAF}(\varphi, \psi)
\end{aligned} \tag{2.11}$$

In this study, we are utilizing the CharmmGUI input generator for OpenMM simulation using the Charmm36m force field. For the simulation ensembles, as we have

mentioned earlier, equilibration steps are completed in an NVT ensemble, and production steps are completed in the NPT ensemble at 303.15K. Molecules are placed in a rectangular water box. In the water box, K^+ and Cl^- ions are placed with the Monte Carlo method, and ion concentration is set to 0.15. (Jo et al., 2008; Lee et al., 2015).

CHAPTER 3

RESULTS AND DISCUSSION

3.1. HIV-1 Capsid Protein Multiple Sequence Alignment

Amino acid sequence conservation of HIV-1 capsid protein is important for his study. In our hypothesis, if CANTDcb1 interacts with a highly conserved region of the capsid protein, we can further utilize CANTDcb1 for diagnosing multiple HIV-1 subtypes.

A subset of sequences representing major subtypes of HIV-1 is acquired from the Los Alamos National Laboratory HIV Sequence database. A total of 414 HIV-1 capsid sequences were aligned, and upon visual inspection, the HIV-1 capsid protein showed significant conservation, clearly visible in Figure 22. After assessing sequence conservation, we applied our alignment result in Chimera. We then highlighted the regions with above 90% sequence conservation on the 3D crystal structure of HIV-1 capsid in Figure 23.

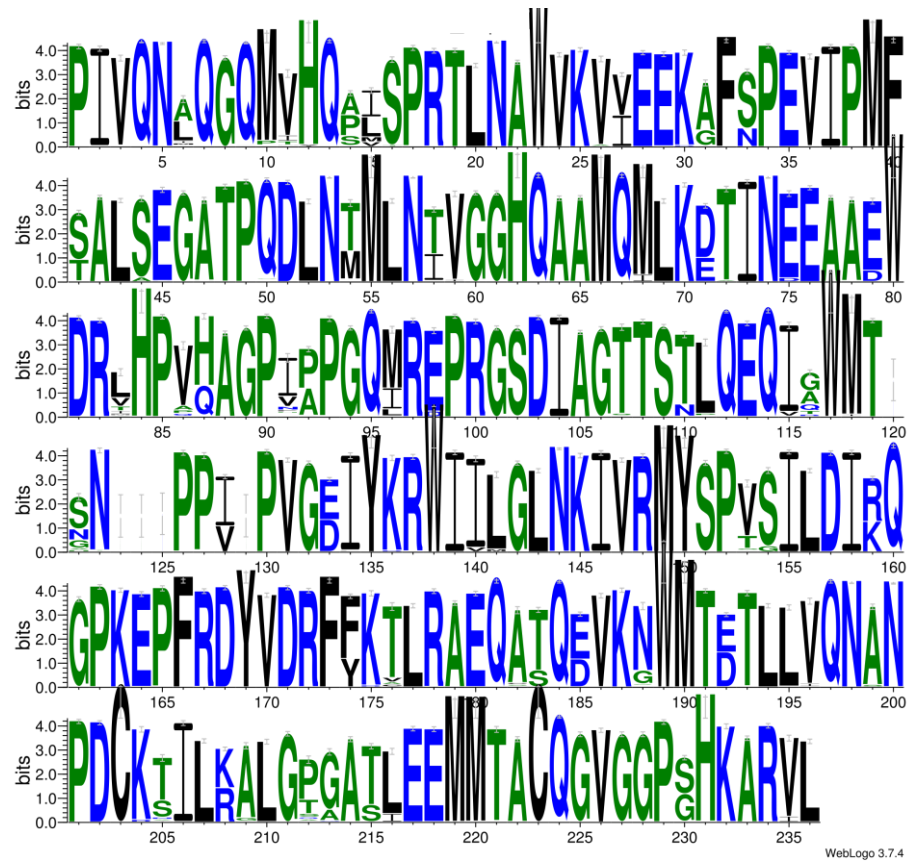


Figure 22. Sequence logo representation of 414 aligned HIV-1 capsid sequences. The horizontal axis represents the residue number, and single-letter representation is used for the residues. The size of the letter indicates how dominant that residue is across multiple sequences. The sequence logo is prepared in WebLogo 3.7.4.

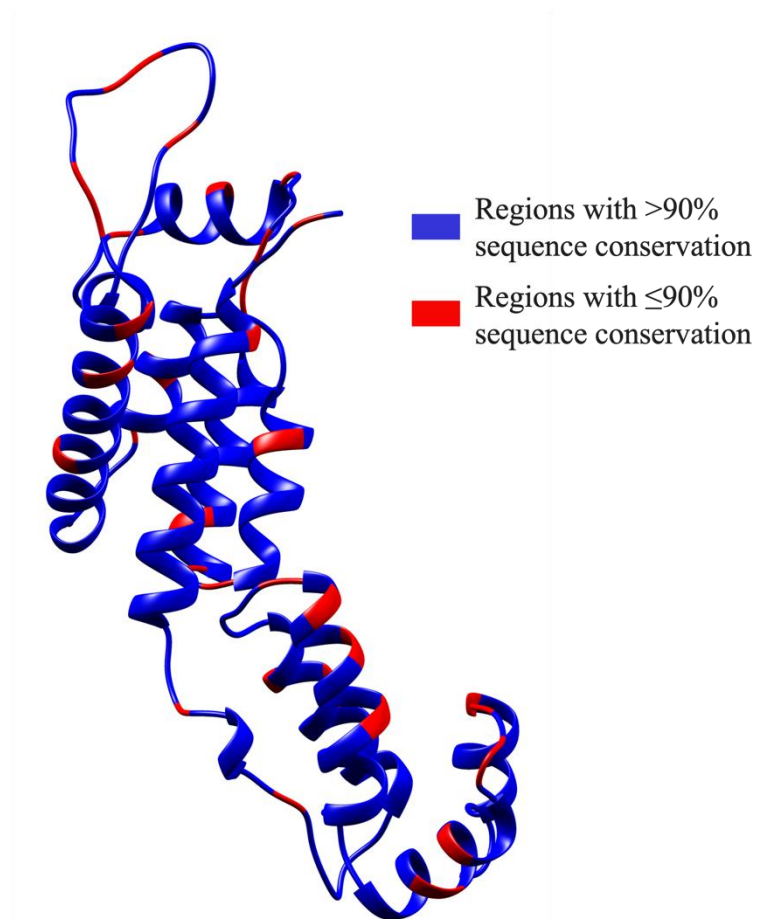


Figure 23. HIV-1 Capsid protein (PDBID: 4XFX) highlighted according to sequence conservation. Residues with above 90% conservation are highlighted with blue, and residues with less or equal to 90% conservation are highlighted with red. The figure is prepared on UCSF Chimera.

3.2. Modeling Nanobody 3D Structure

Modeling of the CANTDcb1 was done on four different methods. The main reasons behind building multiple models are (1) to assess variances in conformation

between models built by different methods and (2) if there are considerable variances, we can use multiple models in our further analysis. For modeling, we did not supply any template structures or sequence alignment. All necessary information was provided by the methods themselves. At the time of modeling, among these methods, only AlphaFold2 did not have any user interface. Therefore, we used a Python script to model the structure. The script is provided in Appendix A. Modeled structures are superimposed with CANTDcb1 amino acid sequence and CDR regions highlighted in Figure 24.

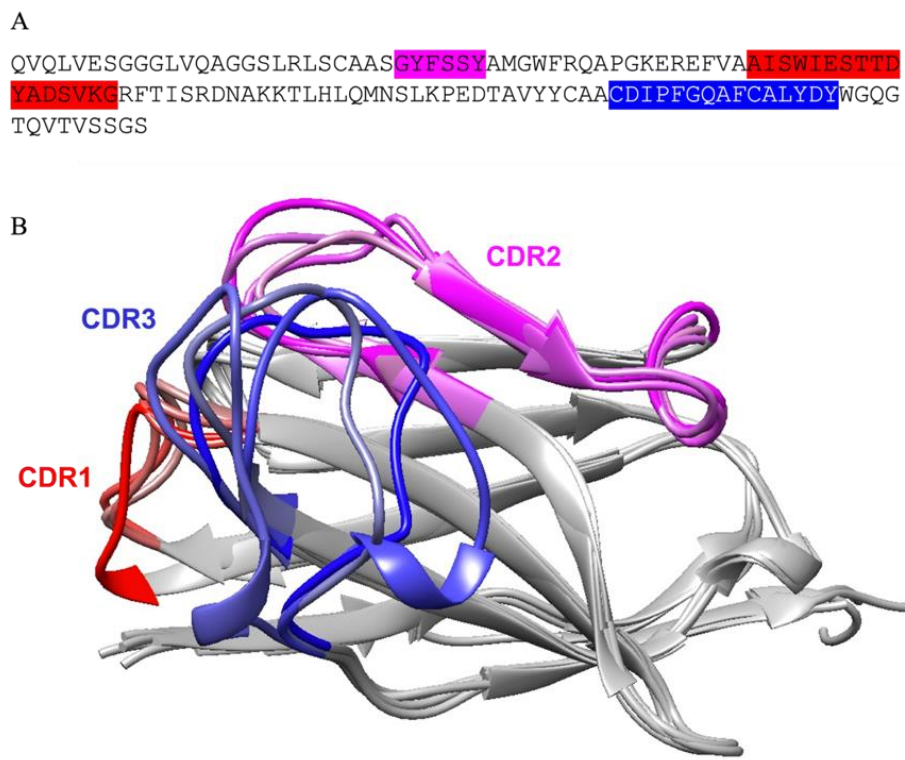


Figure 24. Sequence 3D computational models of CANTDcb1. (A) The amino acid sequence of CANTDcb1 with CDR regions are highlighted. Magenta, red, and blue represent CDR1, CDR2, and CDR3, respectively. (B) 3D computational models of CANTDcb1 are superimposed. Models in this image are built by SwissModel, trRosetta, Robetta, and AlphaFold2. Models are visualized in UCSF Chimera.

In visual inspection, in all four models, skeleton/framework regions were very similar and consistent. However, we saw some variation in CDR3. The reason for this can be the flexibility and length of the CDR3 of this nanobody. In visual inspection, the trRosetta model's CDR3 formed a more 'open' conformation in comparison to other models.

3.3. Structure Assessment of the Nanobody

Upon visual inspection, modeled structures were satisfactory to continue with structure assessment tests. All structure assessment tests were conducted in the SwissModel service.

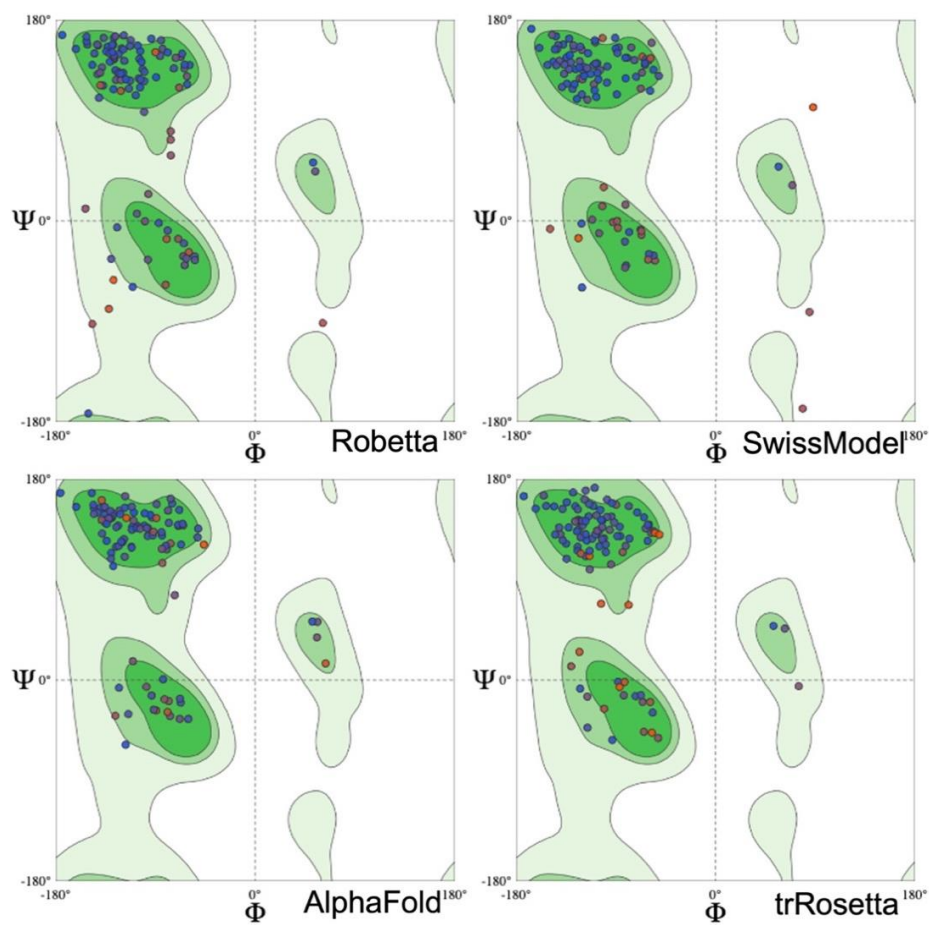


Figure 25. Evaluation of CANTDcb1 on Ramachandran plot. All residues are colored according to their local QMEANDisco scores. Red represents a low score, and blue represents a high score.

Table 2. MolProbity results of the CANTDcb1 models.

	Robetta	SwissModel	AlphaFold2	trRosetta
MolProbity Score	3.04	1.51	2.23	1.07
Ramachandran Favored	94.31%	95.87%	98.37%	99.19%
Ramachandran Outliers	0.81%	2.48%	0.81%	0%
Bad Bonds	2/976	0/965	32/975	0/952
Bad Angles	1/1323	2/1308	11/1321	0/1283

Upon receiving the Ramachandran plots and MolProbity results, we wanted to establish a baseline in comparison to the native structure. For the baseline, four structures from PDB were selected: 4GFT, 4P2C, 5OCL, and 6SSP. All these structures were crystallized in a protein-nanobody complex form. We extracted the nanobodies of these complexes for structure assessment tests. Baseline structures showed Ramachandran favoredness between 93.7%-99.15%. There were no bad bonds or angles in three of the results, 4GFT had two bad bonds and one bad angle. As a result, of this analysis, we eliminated the AlphaFold2 structure due to having a high number of bad bonds and angles.

Global and local QMEANDisCo scores of the modeled structures were evaluated. In Figure 25, we can see the skeleton/framework regions have the highest local scores. This is expected since these regions are mostly conserved in nanobodies. More flexible loop regions and CDRs showed the lowest local scores, especially in CDRs. It is again an expected result, CDRs of nanobodies can have high variability in length, and their sequence is less conserved. trRosetta model received a global QMEANDisCo score of 0.78, and the rest of the models got a score of 0.8. For global score evaluation, scores

above 0.6 are counted as satisfactory. In this analysis, all models behaved as expected. Normalized QMEAN4 scores were also analyzed for these structures (Figure 26). All the modeled structures stayed within $|z - score| \leq 1$ zone.

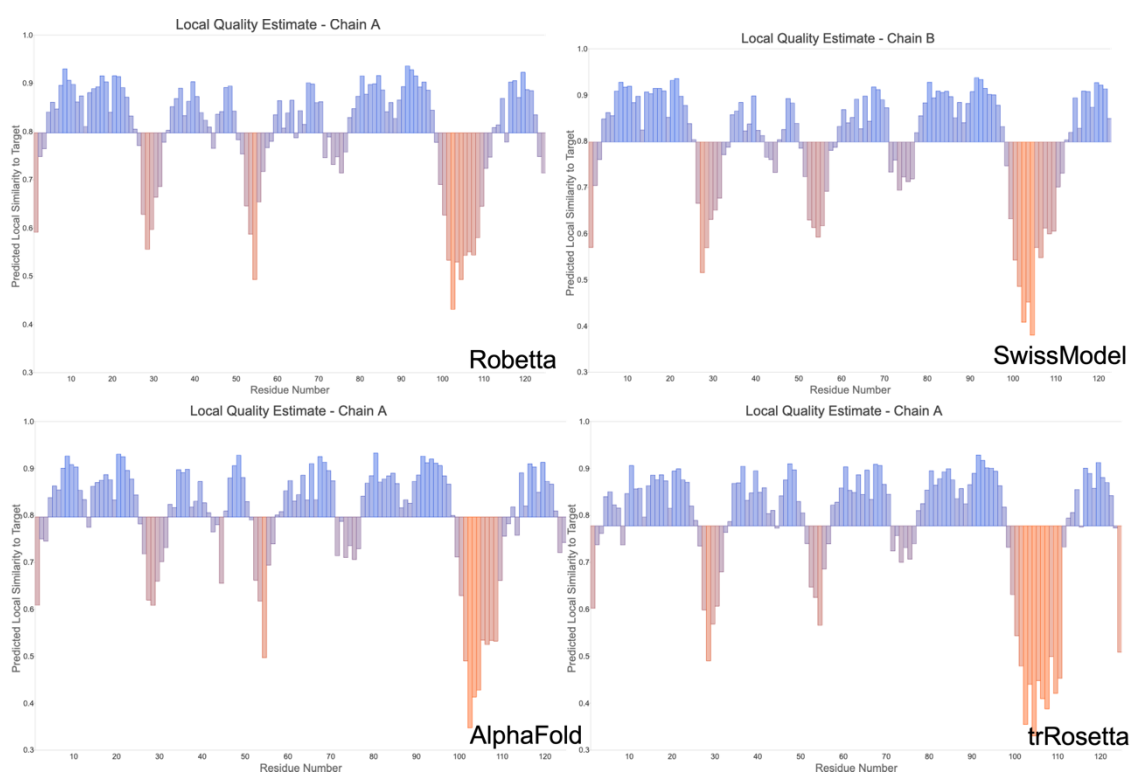


Figure 26. Local QMEANDisCo charts of the CANTDcb1 models. Horizontal axis represents residue number. Vertical axis represents QMEANDisCo evaluation results. Baseline for the bars is set to the model's global QMEANDisCo score.

In conclusion of this analysis, we decided to continue with Robetta and trRosetta models. The Robetta model achieved the highest MolProbity score and showed acceptable performance with high Ramachandran favoredness, low Ramachandran

outliers, and low bad angle and bad bond counts. trRosetta showed acceptable performance also, but it was chosen instead of SwissModel, due to having a relatively different CDR3 conformation than other models.

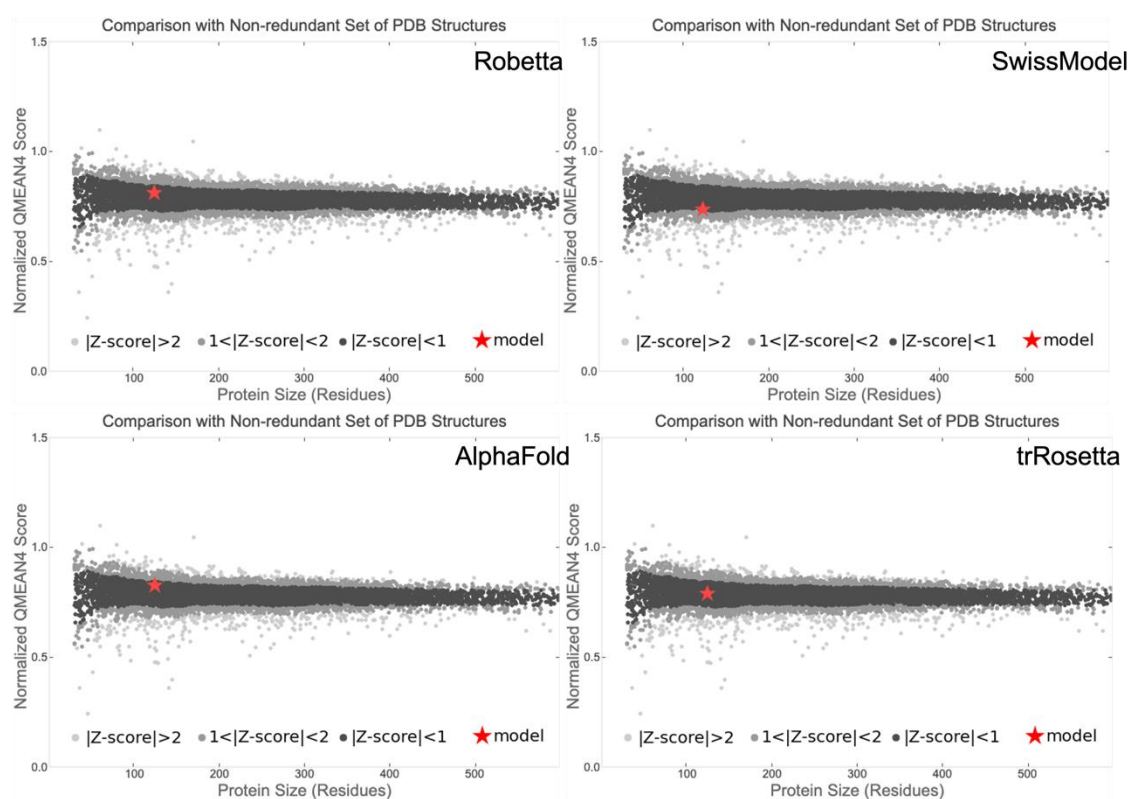


Figure 27. Normalized QMEAN scores of CANTDcb1models expressed as z-scores in comparison with available crystal structures. The graph shows the size of proteins on the horizontal axis. On the vertical axis, we have the "normalized QMEAN" score, which tells us how good the protein structure is. Each dot on the graph represents one real protein structure that has been experimentally determined. The black dots represent crystal structures that have a "QMEAN" score within 1 standard deviation of the average score. The grey dots represent structures that have a "QMEAN" score that is between 1 and 2 standard deviations away from the average. The light grey dots represent structures that

are even further from the average. The red star represents the model we are interested in.

3.4. HIV-1 Capsid Protein and Nanobody Interaction Analysis

Initial docking trials followed a blind docking procedure. In blind docking, we do not introduce any per-residue attraction or repulsion to the docking method. Later, we introduce repulsions and/or attractions to the docking methods to guide the docking process.

The position of CANTDcb1 and possible clashes with capsid monomers in the formation of three hexamers are the first evaluation criteria. From experimental data, we know CANTDcb1 interacts with the NTD of capsid and can bind to the multimeric capsid. If CANTDcb1 in the complexes interacts mostly with the CTD of the capsid or causes clashes in the capsid formation of three hexamers, it is eliminated. Capsid formation of three hexamers is used because it allows us to visualize the position of CANTDcb1 in a multimeric capsid.

3.4.1. ZDock Blind Docking

Initial docking attempts were performed on ZDock 3.0.2 and used the Robetta model. 10 blind docking and 3 guided docking runs were performed, resulting in 130 complexes. The majority of the blind-docked complexes showed a tendency to interact with CTD. Therefore, we introduced restrictions to the CTD of the capsid to guide CANTDcb1 to interact with the NTD of the capsid. This resulted in CANTDcb1 only

targeting the NTD of the capsid, but the majority of the interactions were established through skeleton/framework regions. Due to the performance, it was decided that ZDock is not the appropriate method for this study. Docking studies were continued with ClusPro and Haddock 2.4.

3.4.2. ClusPro Blind Docking

In ClusPro, as mentioned before, it is possible to evaluate complexes by four different scoring functions, and there is also an antibody-antigen specific docking mode. Here we have performed 9 blind docking, including capsid monomer, pentamer (PDBID: 3P05), and the formation of three hexamers (derived from PDBID: 3J3Y) with the Robetta model. In addition, we performed 6 more blind docking with antibody mode in the same conditions. In total, 330 blind-docked complexes were analyzed. Docking results from ClusPro showed a better tendency to target NTD. We decided that results could be improved by introducing attraction and repulsion terms or active residues. However, results from antibody mode caused similar issues to ZDock. These blind docking results were later used to define active residues. We created a subset from ClusPro docking results, where complexes were able to establish salt-bridge(s). Residues involved in these salt bridges were identified in Figures 27-30. From these residues, Arg132 in HIV-1 capsid, and Asp99 and Asp111 from CANTDcb1 CDR3 were selected for attraction active residues.

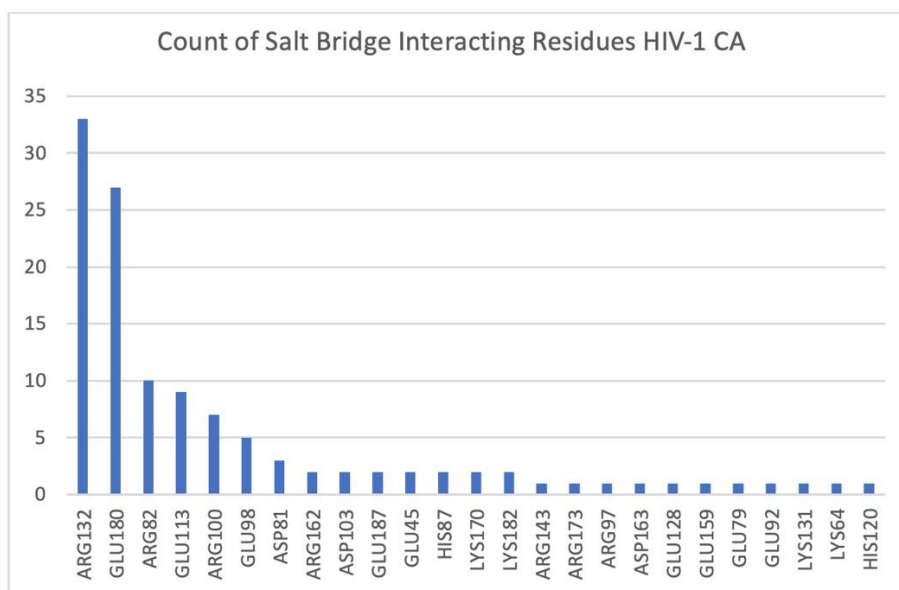


Figure 28. Residues commonly involved in salt bridges in HIV-1 capsid. The horizontal axis represents residues in HIV-1 CA, and the vertical axis represents the number of times these residues appear in complex interfaces.

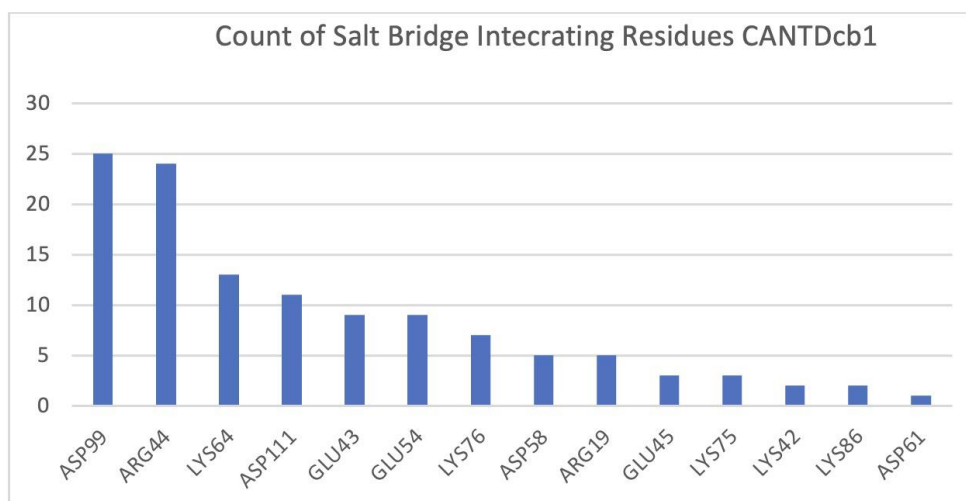


Figure 29. Residues commonly involved in salt bridges in CANTDcb1. The horizontal axis represents residues in HIV-1 CA, and the vertical axis represents the number of times these residues appear in complex interfaces.

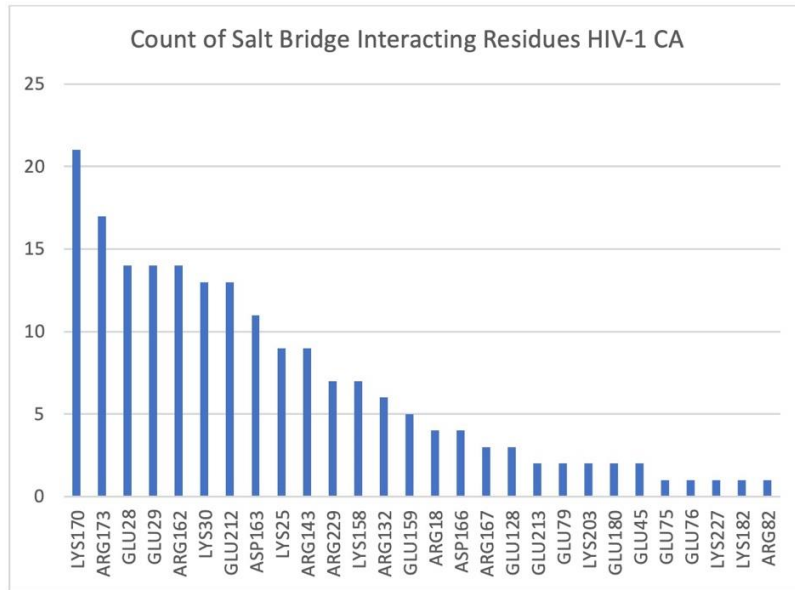


Figure 30. Residues commonly involved in salt bridges in HIV-1 capsid in antibody mode. The horizontal axis represents residues, and the vertical axis represents the number of occurrences in of these residues.

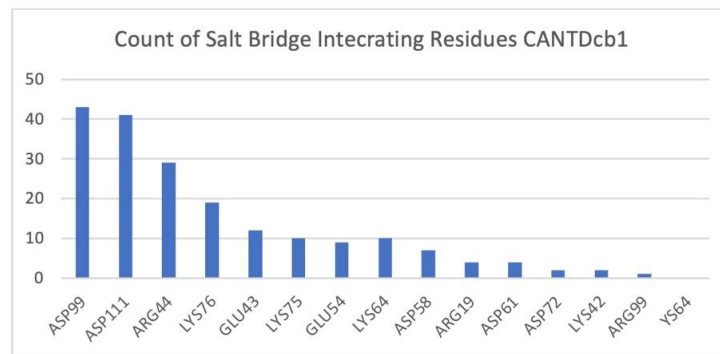


Figure 31. Residues commonly involved in salt bridges in CANTDcb1 in antibody mode. The horizontal axis represents residues, and the vertical axis represents the number of occurrences in of these residues.

3.4.3. Haddock Guided Docking and MD Simulations

For the remainder of this study, we continued only with the guided docking process. For guiding the docking process, several attraction and repulsion terms or active residues were considered. For clarification, attraction and repulsion terms were used in ClusPro, and active residues were used in Haddock. Active residues include residues in HIV-1 capsid NTD and the most common residues that are involved in salt bridges. These active residues were used in combination as listed below:

- (1) HIV-1 CA:NTD - CANTDcb1:Asp99
- (2) HIV-1 CA:NTD - CANTDcb1:Asp111
- (3) HIV-1 CA:NTD - CANTDcb1:Asp99+Asp111
- (4) HIV-1 CA:Arg132 - CANTDcb1:Asp99
- (5) HIV-1 CA:Arg132 - CANTDcb1:Asp111
- (6) HIV-1 CA:Arg132 - CANTDcb1:Asp99+Asp111

For Haddock runs, we used the Robetta model. Complexes were evaluated with docking scores and cluster sizes in addition to the criteria mentioned earlier in this section. Complex 4 in group (1), complex 1 in group (2), complex 10 in group (3), complex 1 in group (4), complex 13 in group (5), and complex 1 in group (6) were selected for further inspection. These complexes are evaluated in PDBSum to determine interactions between capsid and CANTDcb1, and they are visually inspected. Complex 4 in group (1) was not able to establish a salt bridge. Complex 1 in the group (2), complex 10 in group (3), and complex 1 in group (4) showed major clashes when superimposed to tri-hexameric capsid structure. Therefore, these structures were eliminated. Complex 13 in group (5) and complex 1 in group (6) was able to form a salt bridge, although it was not the desired one, and they did not display any major clashes when superimposed to tri-hexameric capsid structures. The orientation of CANTDcb1 in these complexes are acceptable, and it is decided to further evaluate these structures with MD simulations.

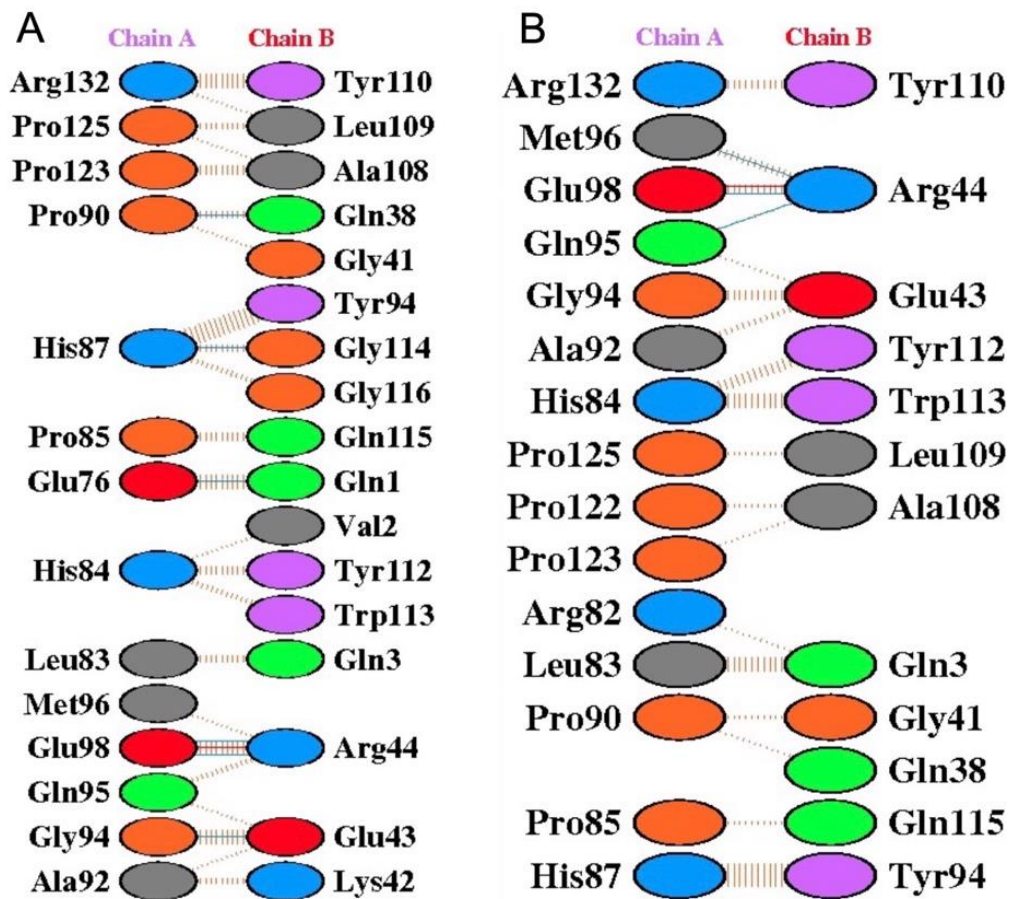


Figure 32. PDBSum results of restricted Haddock run. In both figures, Chain A represents HIV-1 capsid, and Chain B represents CANTDcb1. (A) Complex 13 in group (5), (B) complex 1 from group (6).

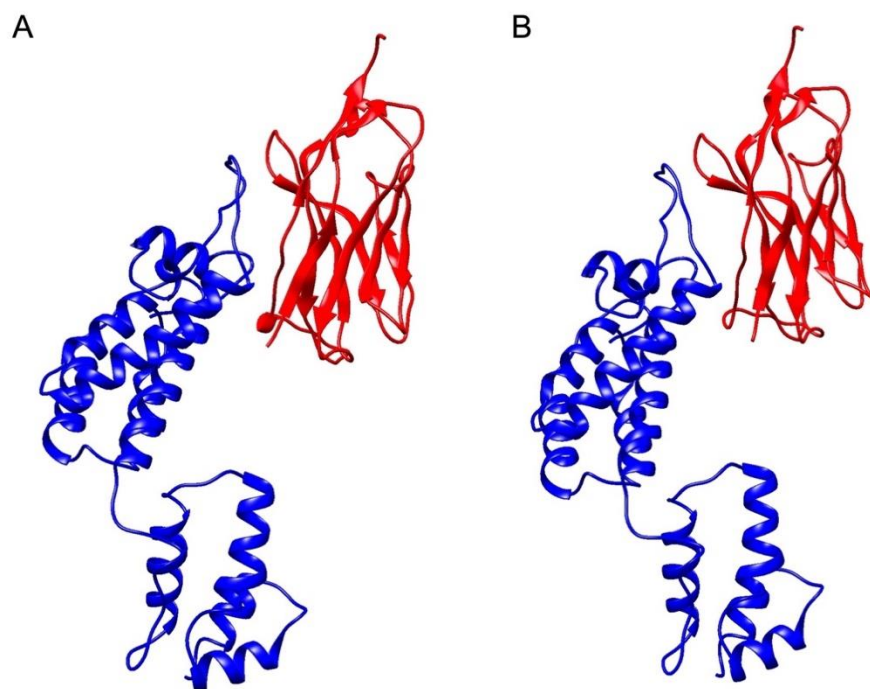


Figure 33. Docked complexes of restricted Haddock run. (A) Complex 13 in group (5), (B) complex 1 from group (6). Figure is prepared on UCSF Chimera.

We started our MD runs with complex 13 of the group (5). While evaluating this run, all complex was aligned. Then we calculated the RMSD of HIV-1 capsid and CANTDcb1 (Figure 34). During the run, there were no major changes in the structures of the proteins. As seen in Figure 33, this complex has one possible salt bridge between HIV-1 CA:Glu98 and CANTDcb1:Arg44. During the MD run, this electrostatic interaction was not conserved, and structures drifted away from their positions (Figure 35). We saw a similar result in complex 1 of the group (6), as seen in Figures 36 and 37. In conclusion, both models failed in our MD runs.



Figure 34. RMSD calculations of the structures through the simulation of complex 13 of group (5) compared to the starting structures.

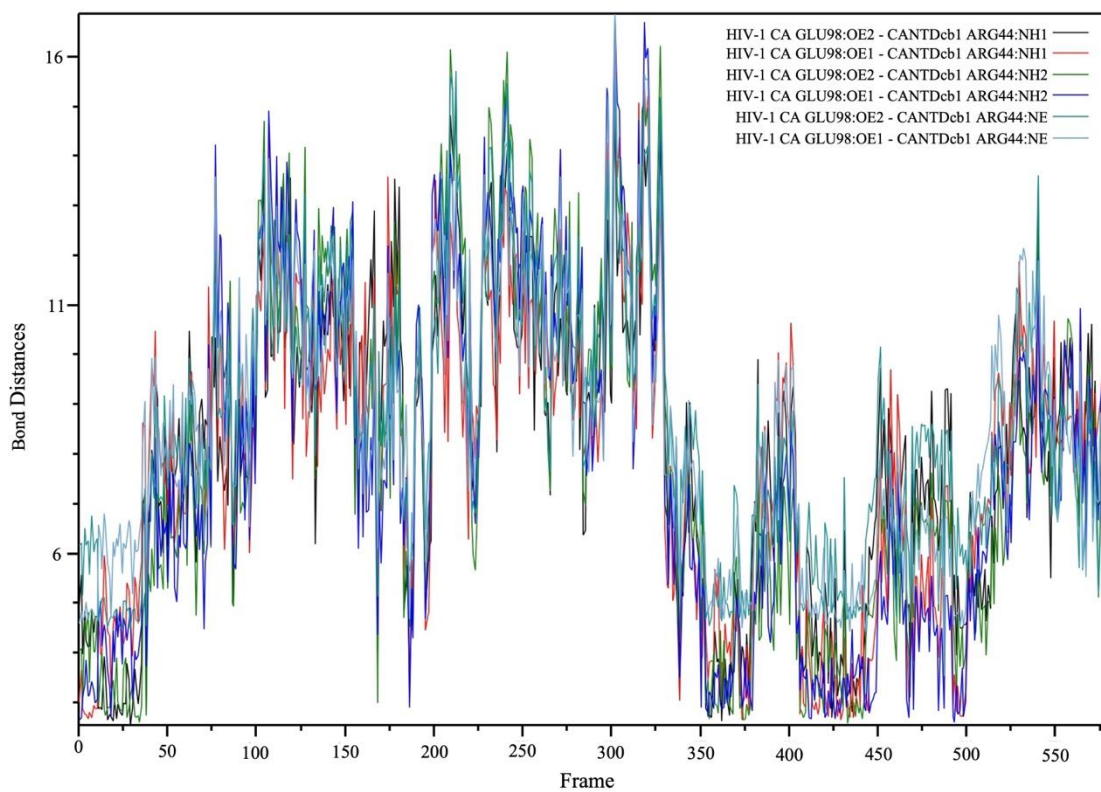


Figure 35. HIV-1 CA:Glu98 and CANTDcb1:Arg44 distances per atom throughout the simulation of complex 13 of group (5).

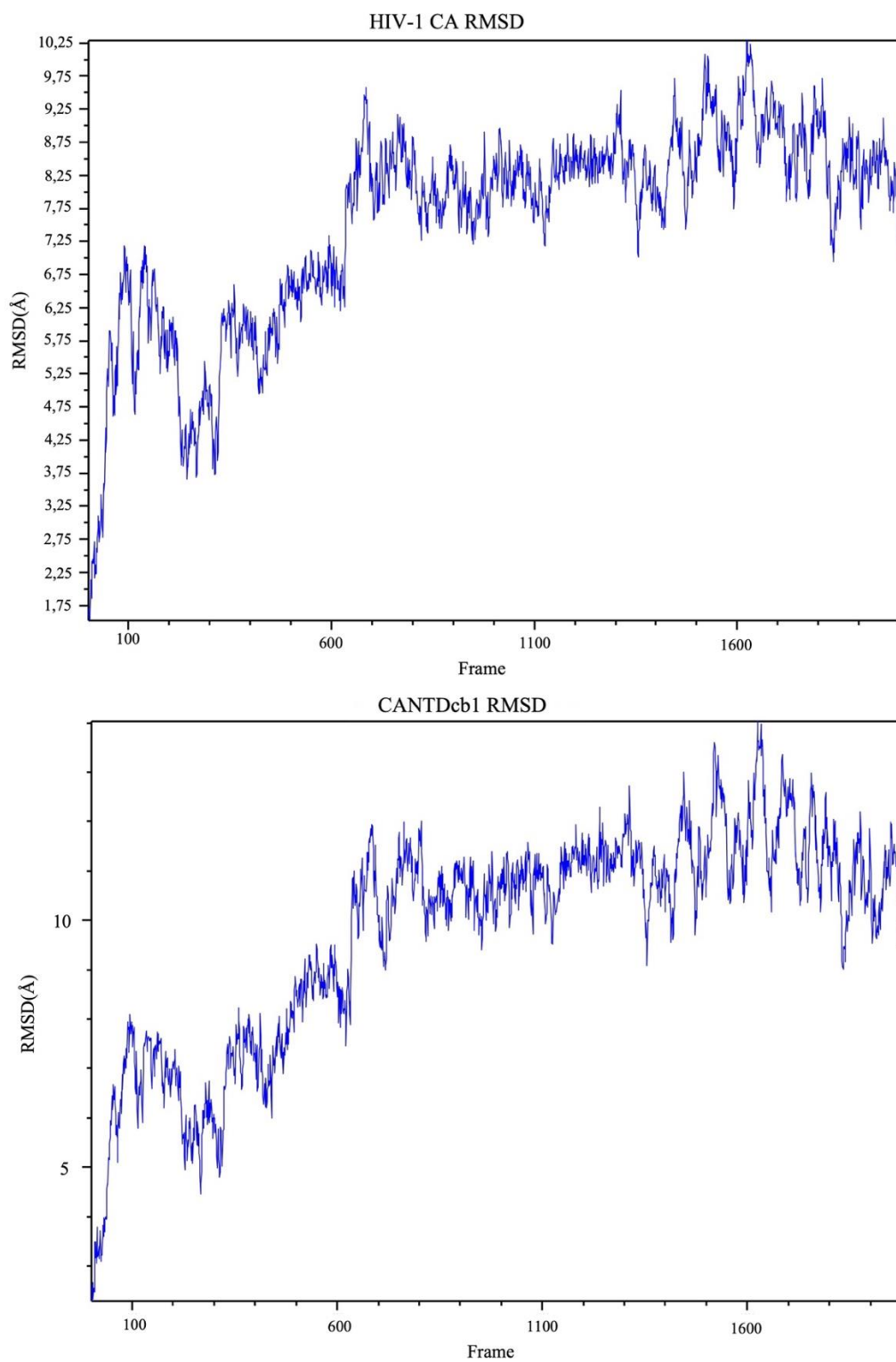


Figure 36. RMSD calculations of the structures through the simulation of complex 1 of group (6).

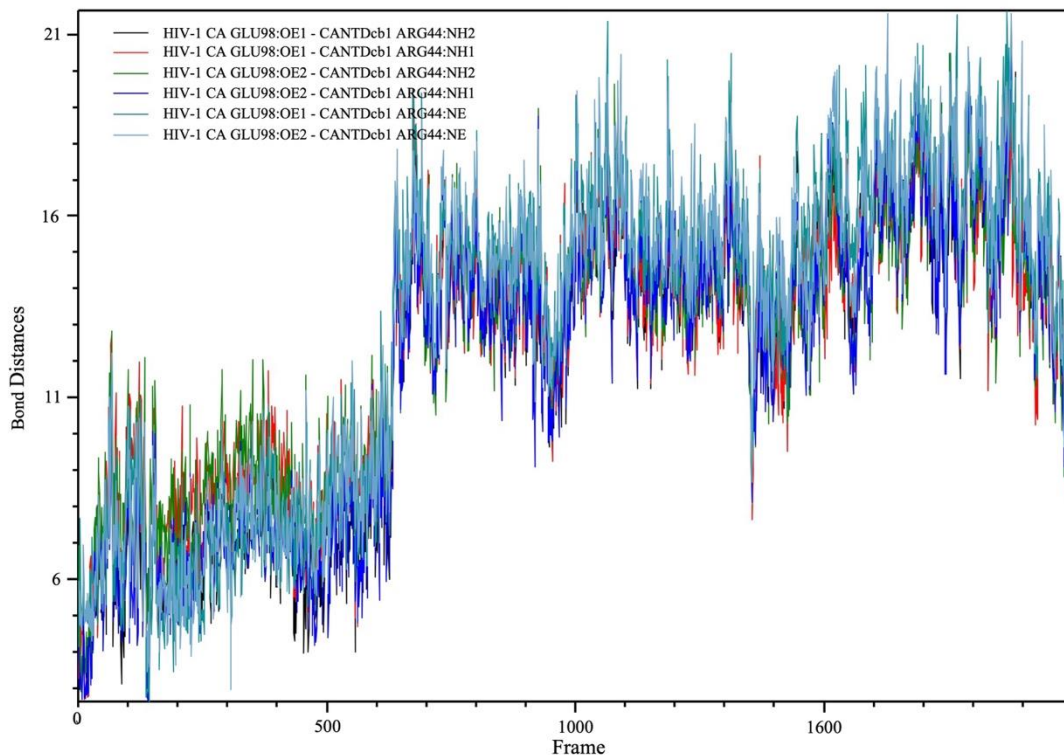


Figure 37. HIV-1 CA:Glu98 and CANTDcb1:Arg44 distances through the simulation of complex 1 of group (6).

3.4.4. ClusPro Guided Docking and MD Simulations

Docking studies were conducted on ClusPro to introduce repulsion terms. It was realized that, unlike Haddock, ClusPro was very strict with repulsion terms, as was observed in capsid CTD repulsion terms. Since strict terms are needed to guide the docking, we preferred to continue with ClusPro. As we have mentioned earlier, there are several residues that take part in establishing the necessary interactions to form the capsid core (Figure 21). We introduce these residues as repulsion terms during ClusPro docking.

Up until this point, only the Robetta model of the CANTDcb1 was used. Here we introduced the trRosetta model as well, doing two docking runs.

We started this docking process with the Robetta model. Of the docked complexes, the first complex showed the best performance. This complex had three possible salt bridges (Figure 38) and was in a favorable position to not cause any major clashes with the trihexameric capsid structure.

After the evaluation, we decided to perform an MD run on this complex. In this run, the complex was stable throughout the simulation. Only one binding residue pair was unstable during the simulation (Figures 38-42). Since the results of this MD run were satisfactory, we decided to perform two more MD runs to determine if the results were replicable.

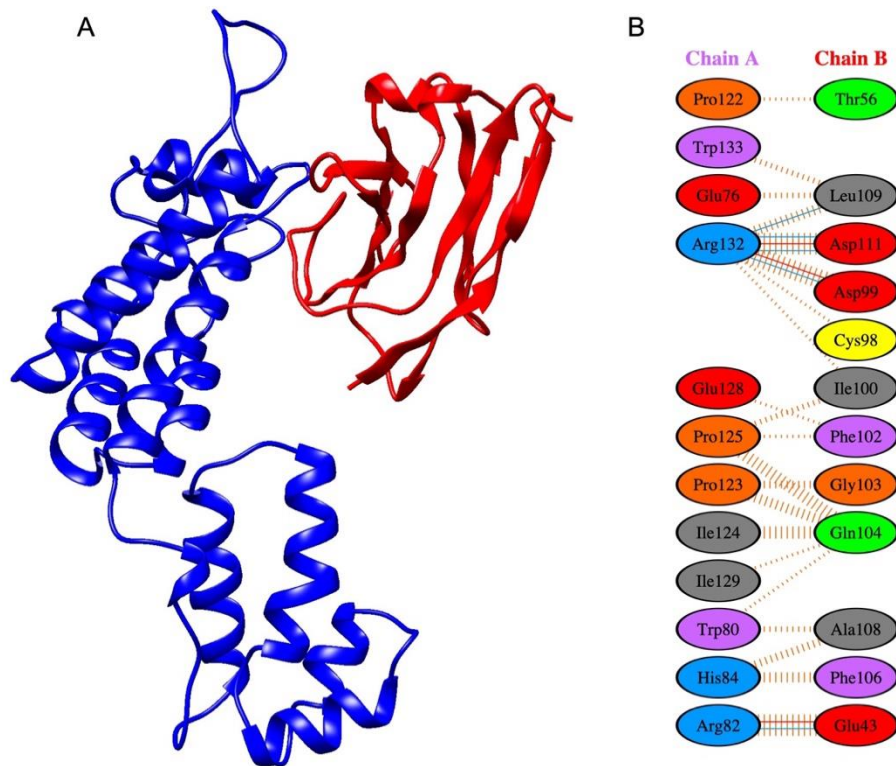


Figure 38. First of the docking results from the Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interaction to form the capsid core. (A) 3D visualization of the docked pose. The figure is prepared on UCSF Chimera. (B) PDBSum result of the docked pose.

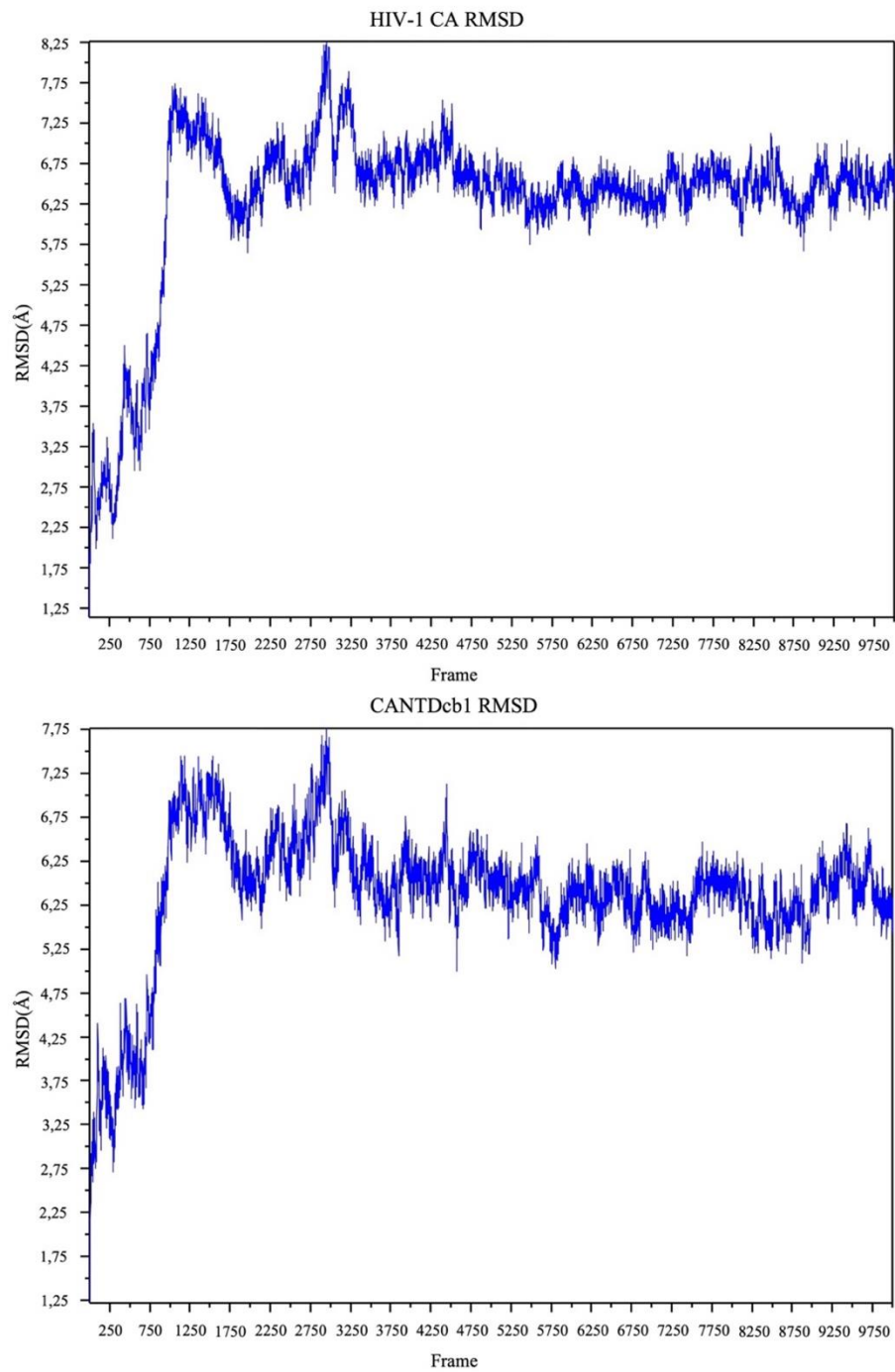


Figure 39. RMSD calculations of the first complex of the docking results from the Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interactions to form the capsid core.

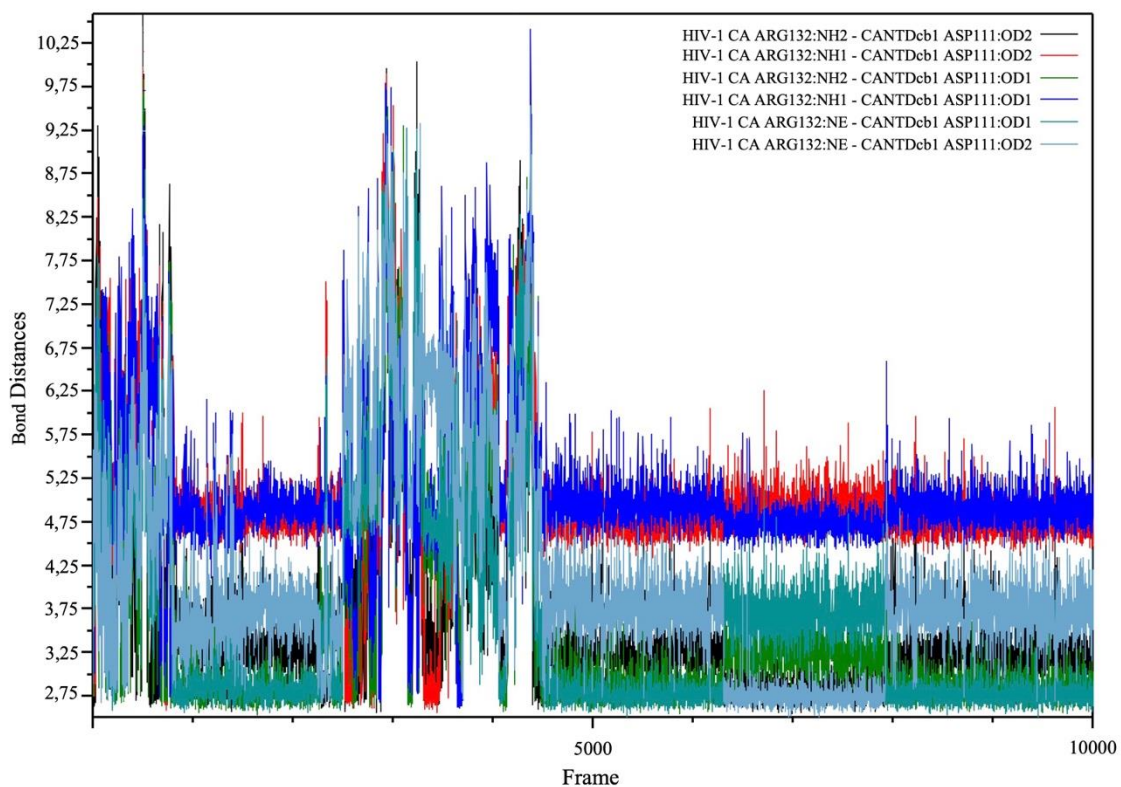


Figure 40. HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as shown with atomic distances. This is the first complex of the docking results from Robetta model docking with repulsion terms applied for residues that takes part in establishing necessary interaction to form capsid core.

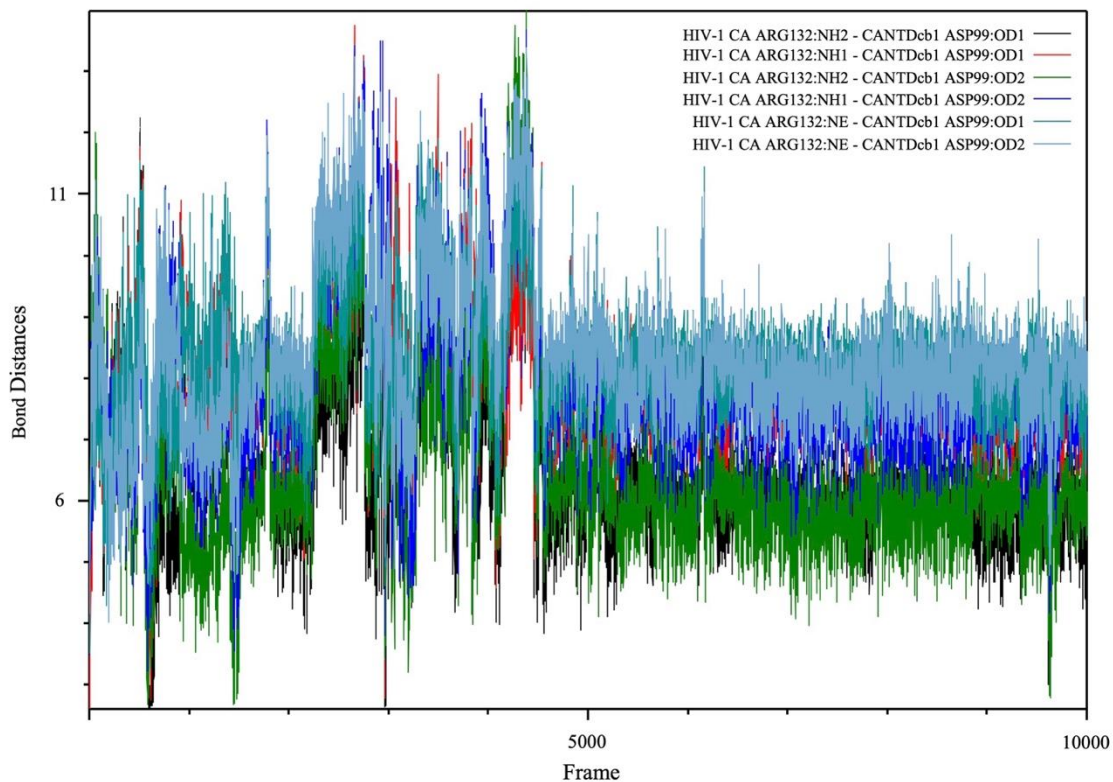


Figure 41. HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as shown with atomic distances. This is the first complex of the docking results from Robetta model with repulsion terms applied for residues that takes part in establishing necessary interaction to form capsid core.

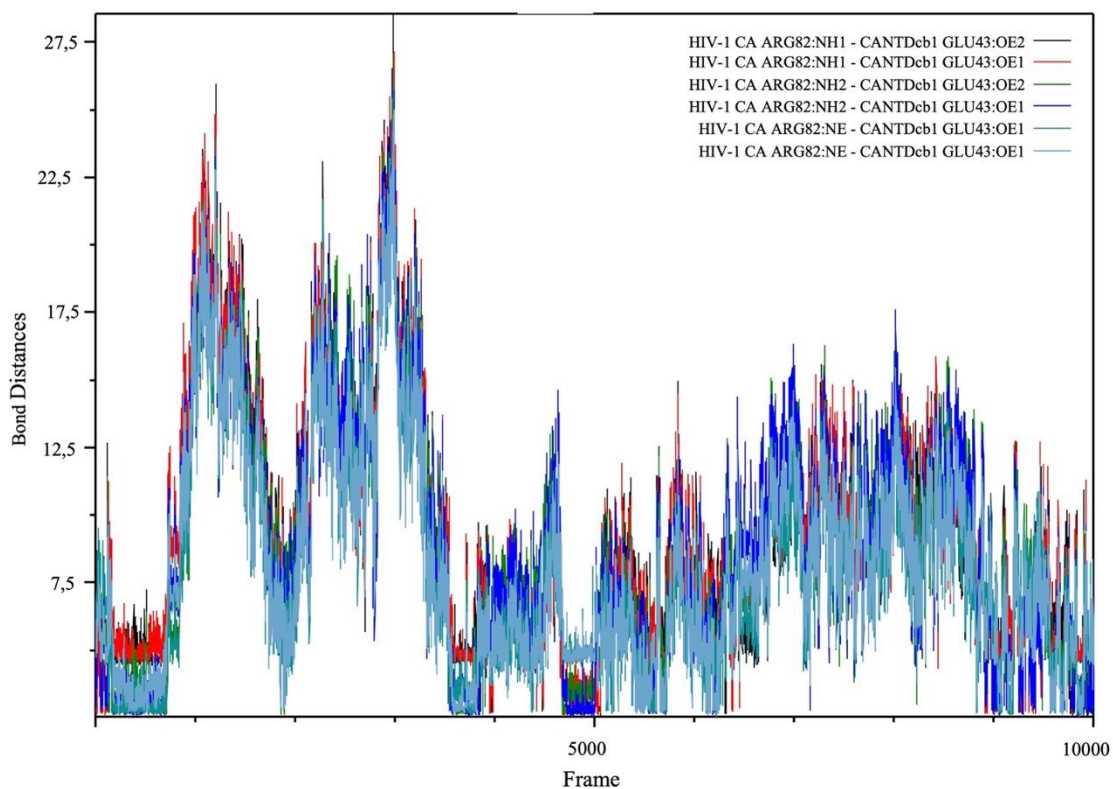


Figure 42. HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as shown with atomic distances. This is the first complex of the docking results from Robetta model docking with repulsion terms applied for residues that takes part in establishing necessary interaction to form capsid core.

In the repeated runs, the results were not replicated. The outcome of the MD simulations were different in all runs. In the repeated runs, structures were not stabilized during simulation or separated completely, as seen in Figures 43-50.

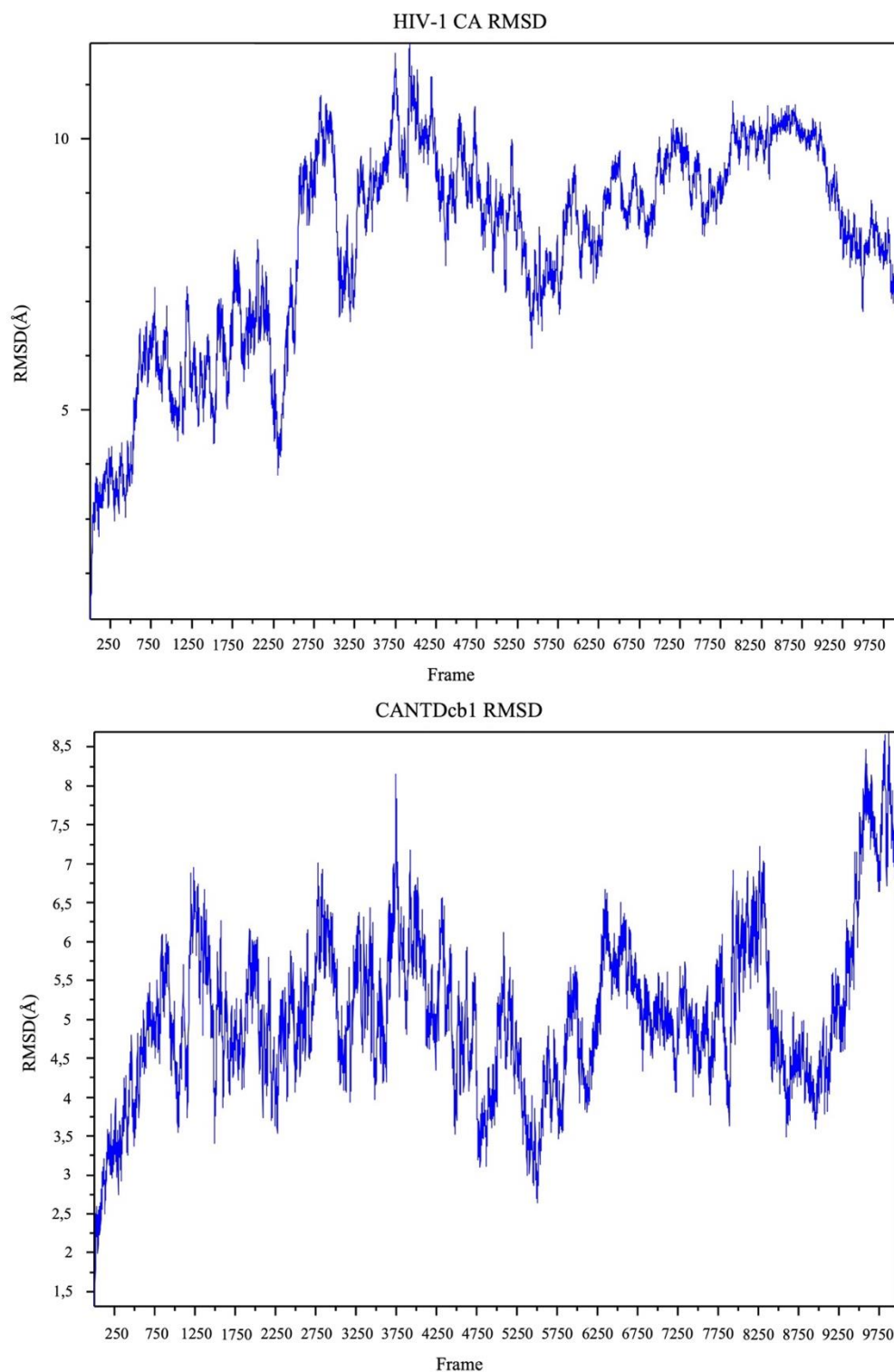


Figure 43. Second run RMSD calculations of the first complex of the docking results from Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interaction to form the capsid core.

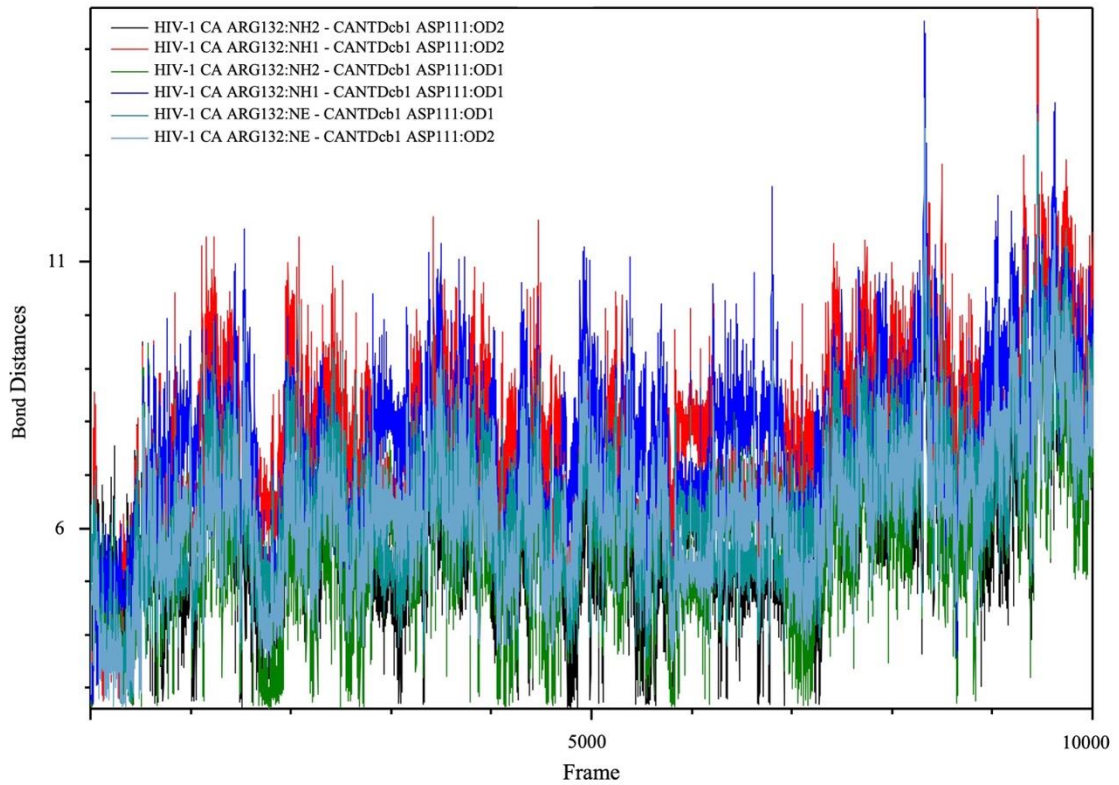


Figure 44. Second run HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for residues that takes part in establishing the necessary interaction to form the capsid core.

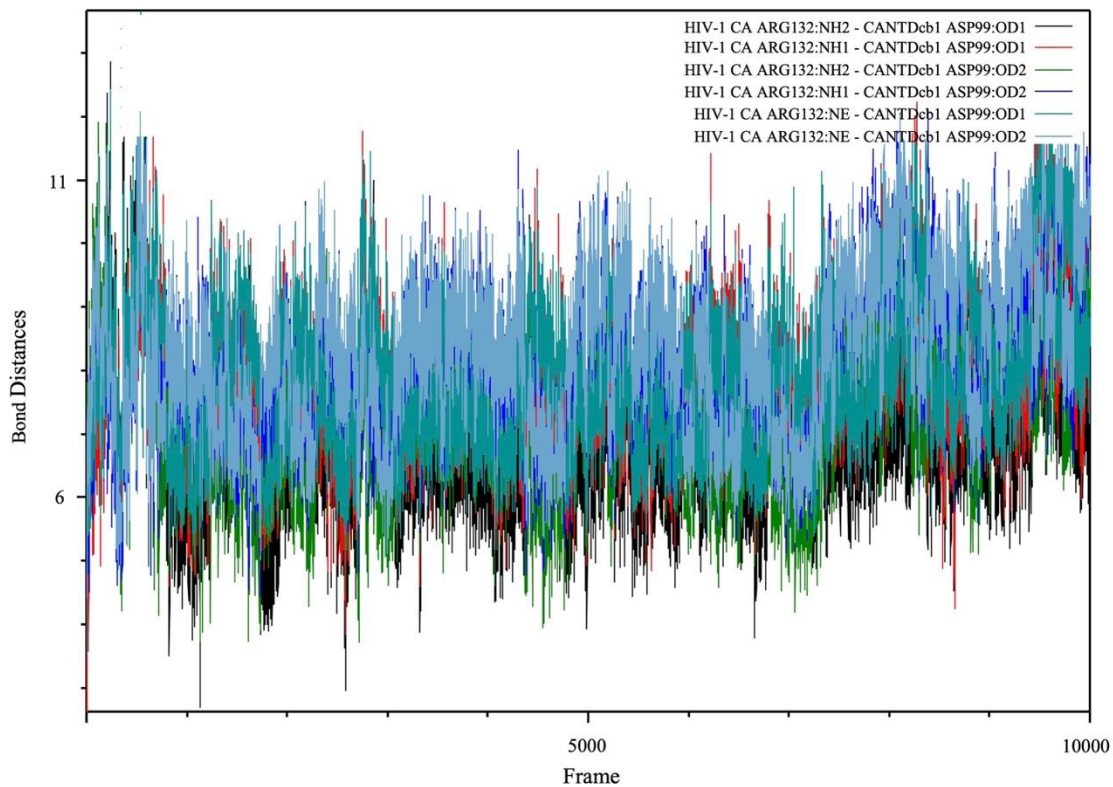


Figure 45. Second run HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interaction to form the capsid core.

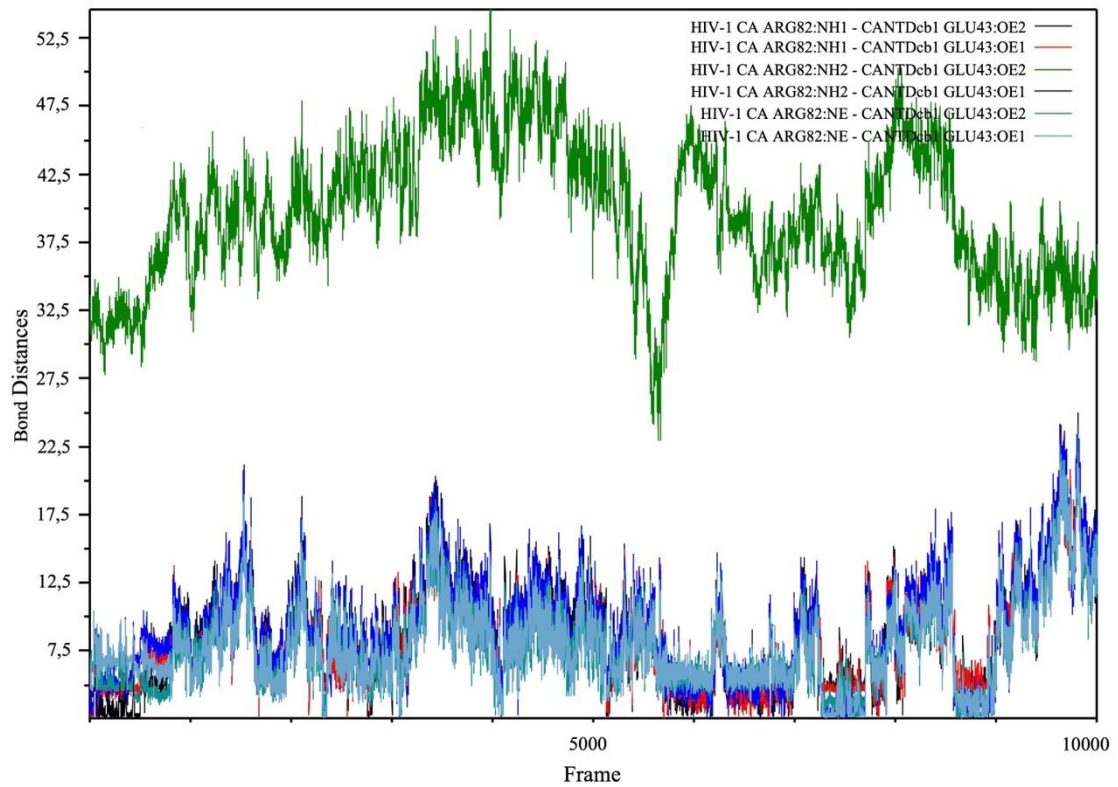


Figure 46. Second run HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that takes part in establishing the necessary interaction to form capsid core.

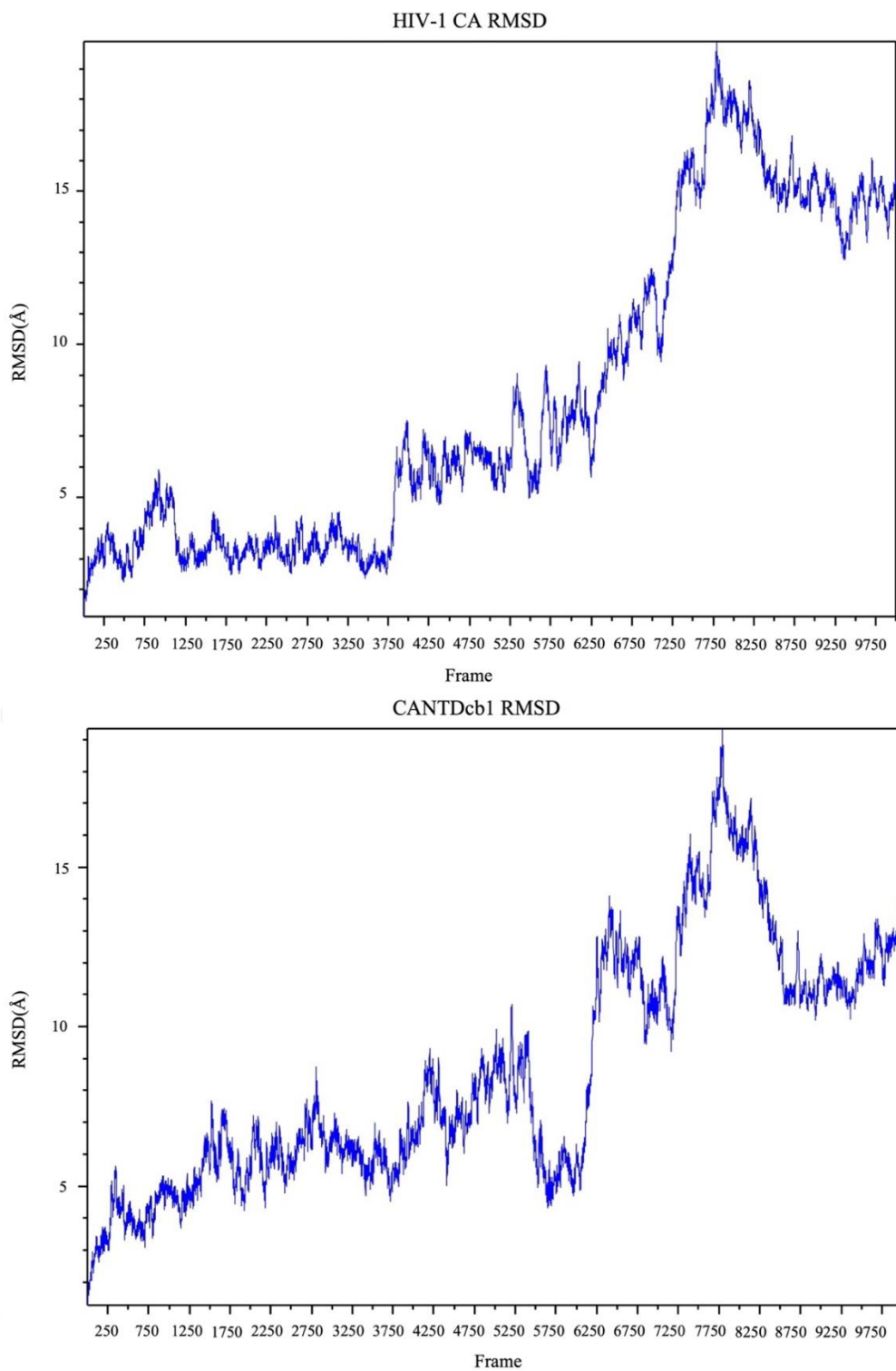


Figure 47. Third run RMSD calculations of the first complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

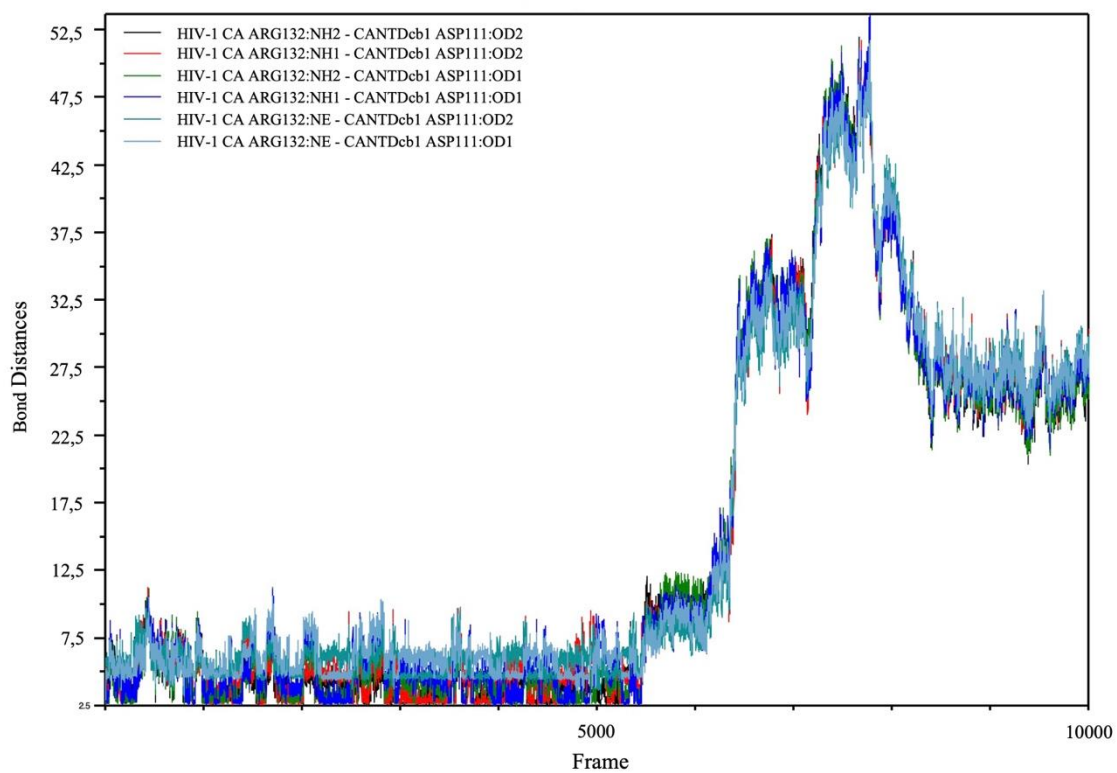


Figure 48. Third run HIV-1 CA:Arg132 and CANTDcb1:Asp111 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

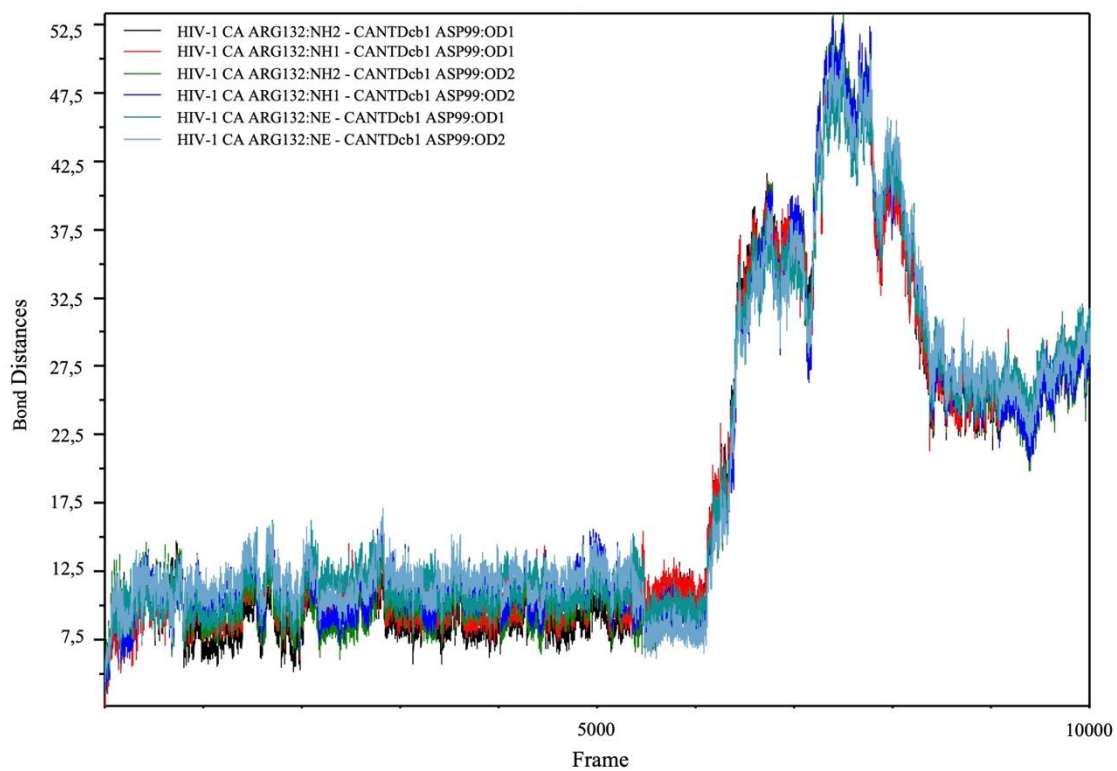


Figure 49. Third run HIV-1 CA:Arg132 and CANTDcb1:Asp99 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

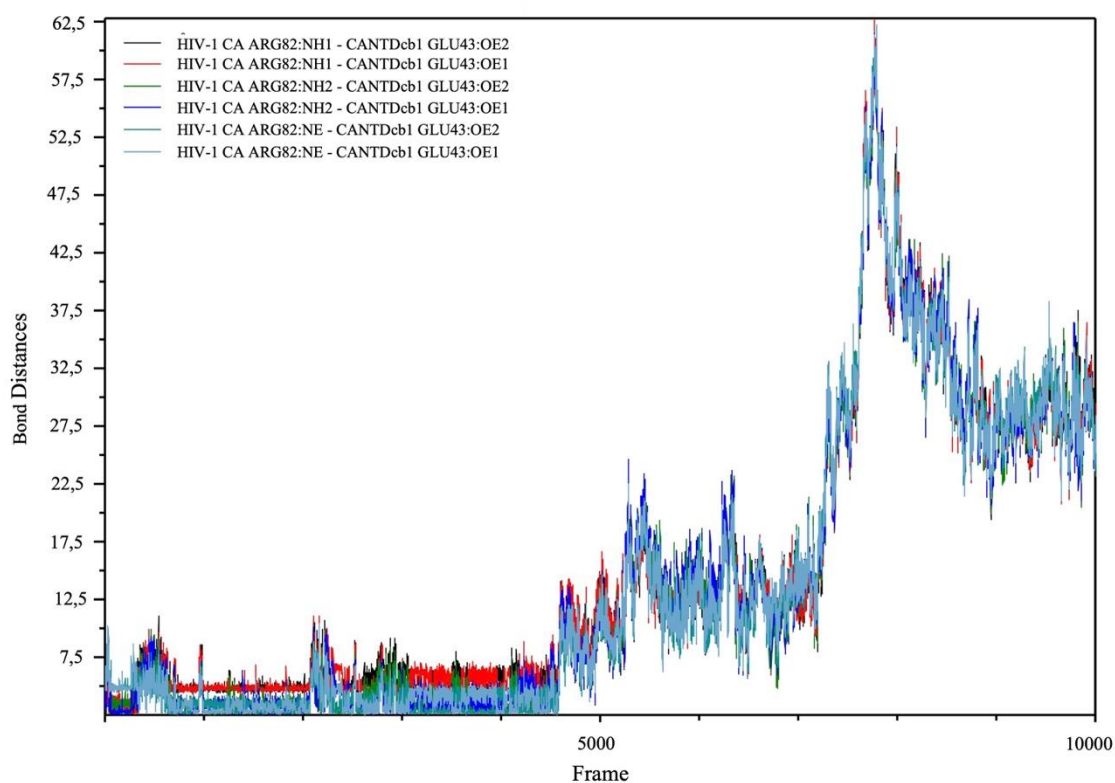


Figure 50. Third run HIV-1 CA:Arg82 and CANTDcb1:Glu43 interaction as shown with atomic distances. This is the first complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

During the visual inspection of docked complexes, model 8 showed a different conformation than the rest. Therefore, we evaluated the interactions involved in this complex on PDBSum. The majority of these interactions were hydrophobic interactions (Figure 51). Due to our curiosity, we performed three MD simulations with this model.

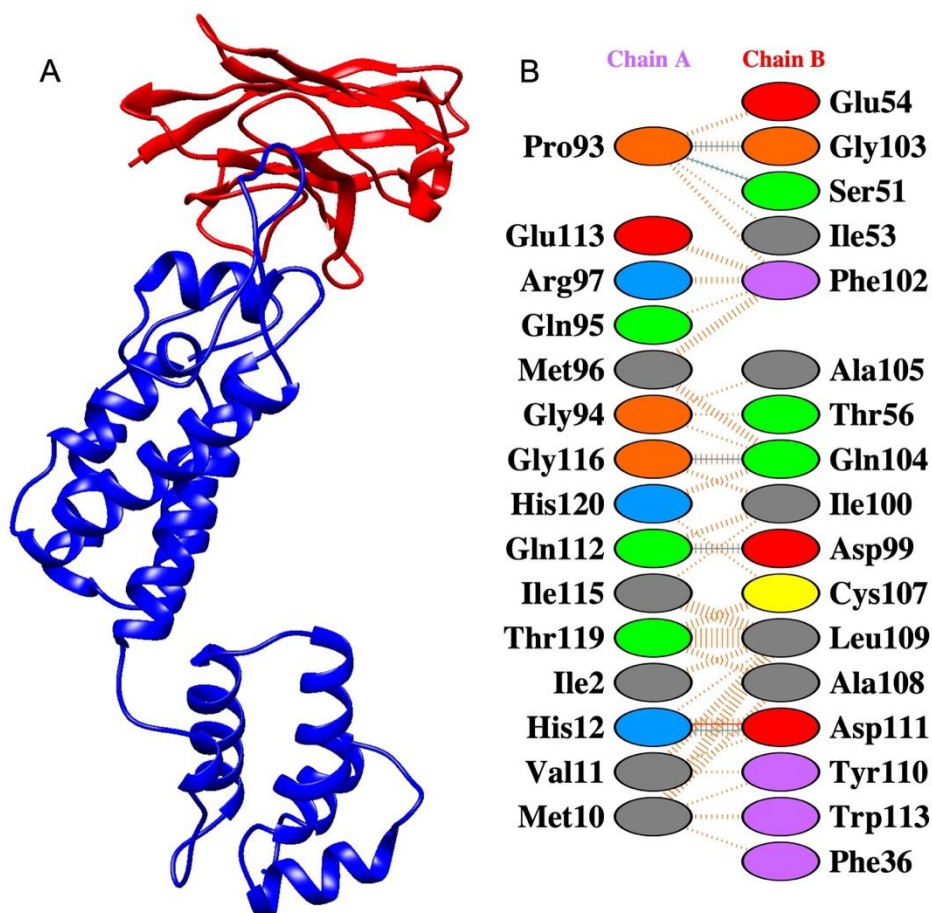


Figure 51. Eighth Complex of the docking results from Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core. (A) The 3D visualization of the docked pose. The figure is prepared on UCSF Chimera. (B) PDBSum result of the docked pose.

This model did not have any salt bridges, so we followed a different way to analyze binding this MD run. We prepared three runs for this complex; however, only the first run is completed for 100 ns. The second run is at around 80 ns, and the third run is yet to start at the time of writing. First, we performed a visual assessment for the two MD runs, then calculated their RMSD throughout the simulation (Figures 52 and 53).

Throughout both simulations, capsid and CANTDcb1 stayed bound. To better understand the binding dynamics, contacts between the proteins were identified and presented in Table 3. Previous research on nanobody-antigen interactions has revealed distinct characteristics compared to traditional antibodies. Nanobodies exhibit a higher prevalence of hydrophobic interactions, particularly involving amino acids Ile, Val, and Leu. Additionally, Glu, Asn, Asp, Ser, Thr, and Tyr residues have been identified as frequent contributors to antigen binding, with Tyr playing a prominent role among them. Furthermore, the CDR3 region of the nanobody has been identified as the primary region in establishing contact with the antigen. The interactions observed in Table 3 align with the findings from these earlier studies, further supporting the significance of these specific features in nanobody-antigen recognition (Liu et al., 2022; Mitchell & Colwell, 2018a). These interactions involve hydrogen bonds and hydrophobic interactions. All highlighted contacts are hydrogen bonds, and common contacts between the two simulation runs were highlighted in magenta (Table 3). Almost all these interactions are through the CDRs of CANTDcb1, especially the CDR3. According to the calculations of PDBSum, 738\AA^2 and 773\AA^2 solvent accessible surface area between capsid and CANTDcb1 was buried, respectively. These contacts were first plotted in Figure 54, then analyzed individually in Figure 55. In conclusion, the capsid-CANTDcb1 complex was stable and stayed bound during the two MD runs. This model will be analyzed further with the completion of the second and third MD runs.

Table 3. Interactions in two MD runs of model 8 complex of the docking results from Robetta model docking with repulsion terms applied for residues that take part in establishing necessary interaction to form capsid core binding analysis. Pairs highlighted in yellow and magenta represent potential hydrogen bonds and common interactions between different runs, respectively.

RUN 1 (HIV-1 CA – CANTDcb1)	RUN 2 (HIV-1 CA – CANTDcb1)
THR119:HB – ILE100:CD1	MET10:O - LEU109:CA
GLY94:HA2 – PHE102:HA	MET10:C - LEU109:HA
HIS120:HA – PHE102:CD1	MET10:O - LEU109:HA
HIS120:HA – PHE102:HD1	MET10:O - LEU109:HB3
HIS120:HA – PHE102:CE1	VAL11:HA - LEU109:HB3
GLY116:HA2 – PHE102:HE1	HIS12:H - LEU109:HB3
GLY116:O - PHE102:HE1	MET10:HA - LEU109:CD2
THR119:OG1 - PHE102:HE1	MET10:HA - LEU109:HD21
HIS120:N - PHE102:HE1	MET10:HA - LEU109:HD22
HIS120:HA - PHE102:HE1	MET10:HA - LEU109:HD23
GLY116:HA3 - PHE102:HZ	MET10:O - LEU109:C
GLY116:HA2 - PHE102:HZ	MET10:O - TYR110:N
GLY116:O - PHE102:HZ	MET10:C - TYR110:H
GLY94:C - PHE102:HD2	MET10:O - TYR110:H
GLY94:O - PHE102:HD2	VAL11:HA - TYR110:H
THR119:HB - CYS107:O	VAL11:HA - TYR110:O
THR119:CG2 - CYS107:O	MET10:O - TRP113:HZ2
MET10:HA - ALA108:O	
MET10:O - LEU109:CA	
MET10:CA - LEU109:HA	
MET10:HA - LEU109:HA	

(cont. on the next page)

Cont. of Table 3

MET10:C - LEU109:HA	
MET10:O - LEU109:HA	
THR119:CG2 - LEU109:HB2	
THR119:HG21 - LEU109:HB2	
THR119:HG22 - LEU109:HB2	
THR119:HG23 - LEU109:HB2	
MET10:O - LEU109:C	
MET10:O - TYR110:N	
TYR110-H MET10-C	
TYR110-H MET10-O	
TYR110-H VAL11-HA	
TYR110-O MET10-O	
TYR110-O VAL11-HA	
TRP113-HZ2 MET10-C	
TRP113-HZ2 MET10-O	

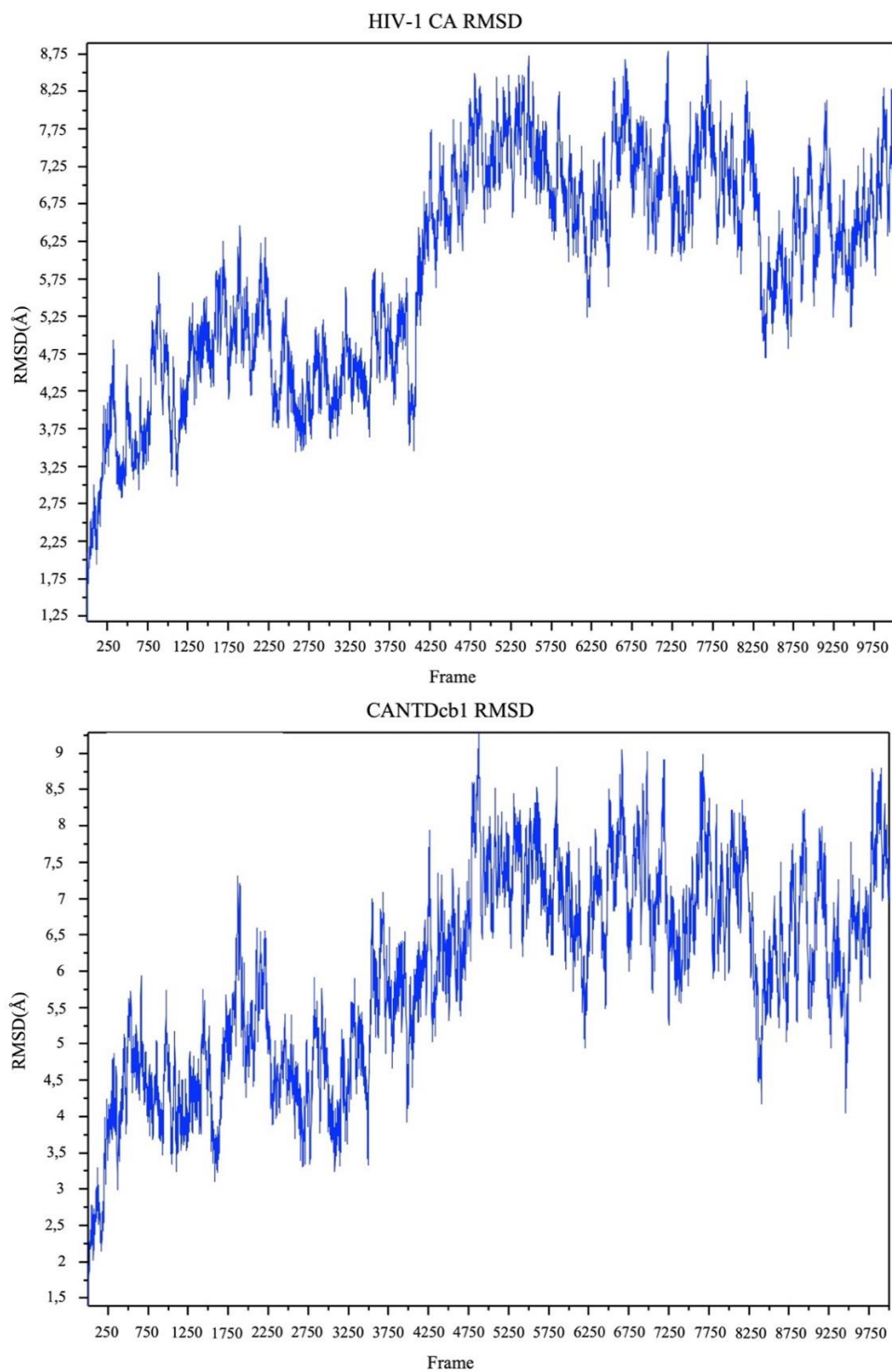


Figure 52. First run RMSD calculations of the eighth complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

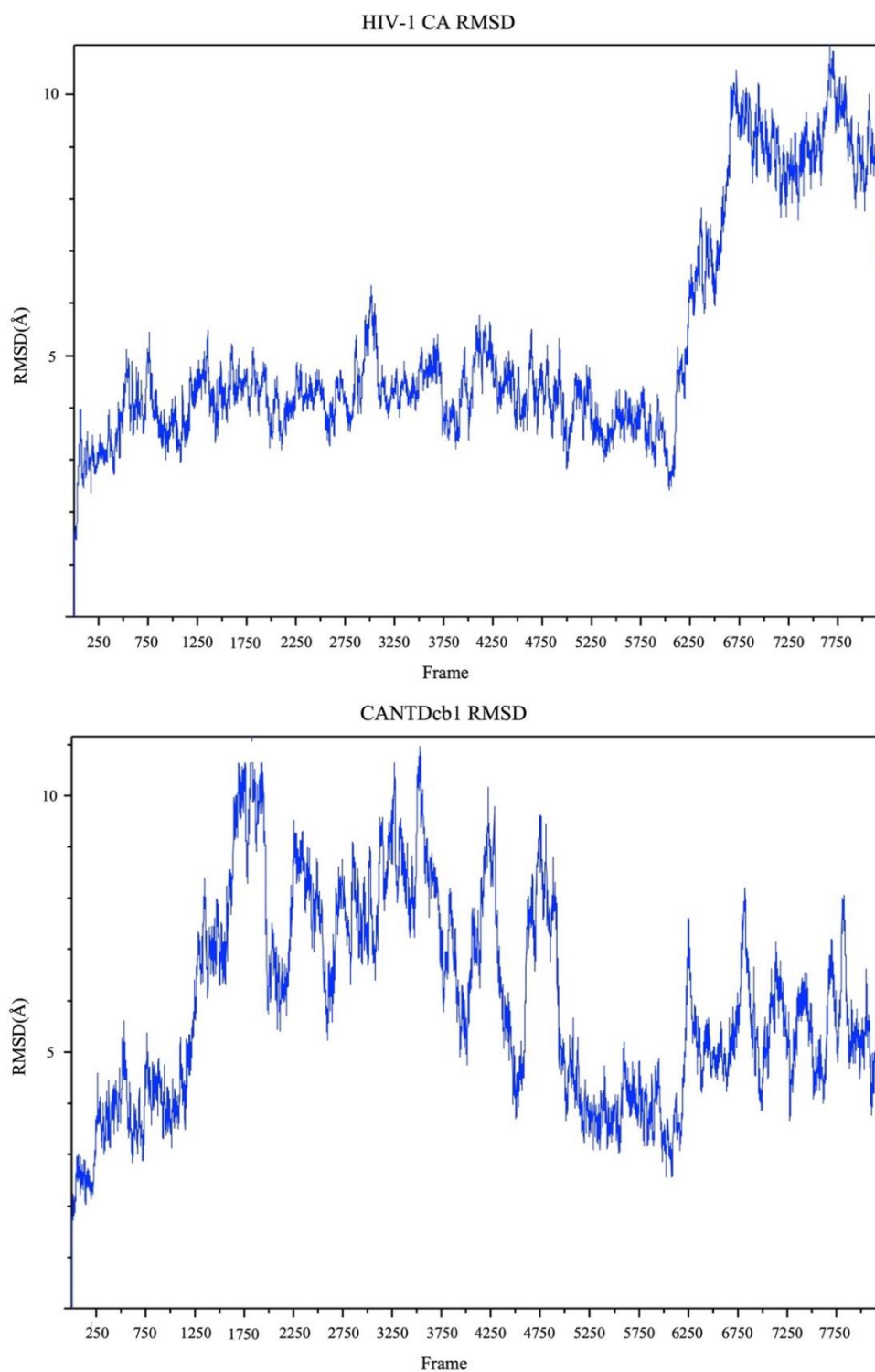


Figure 53. Second run RMSD calculations of the eighth complex of the docking results from the Robetta model docking with repulsion terms applied for the residues that take part in establishing the necessary interaction to form the capsid core.

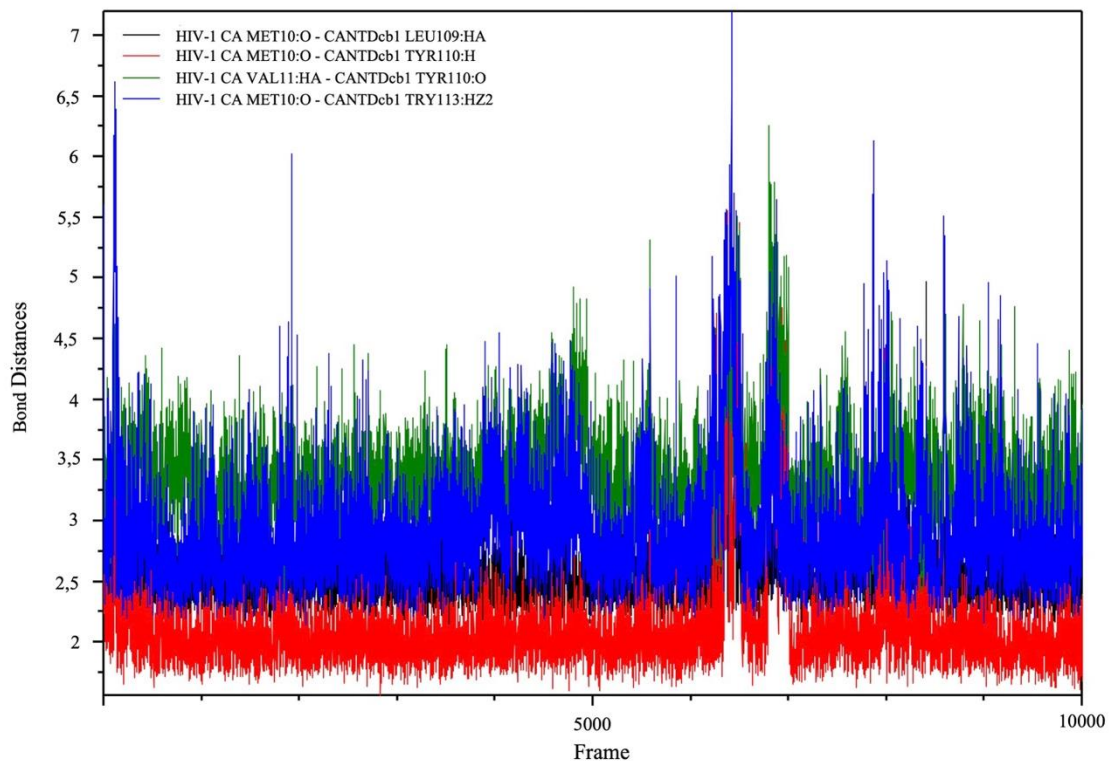


Figure 54. First run point of view for the eighth complex of the docking results from the Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interaction to form capsid core binding analysis.

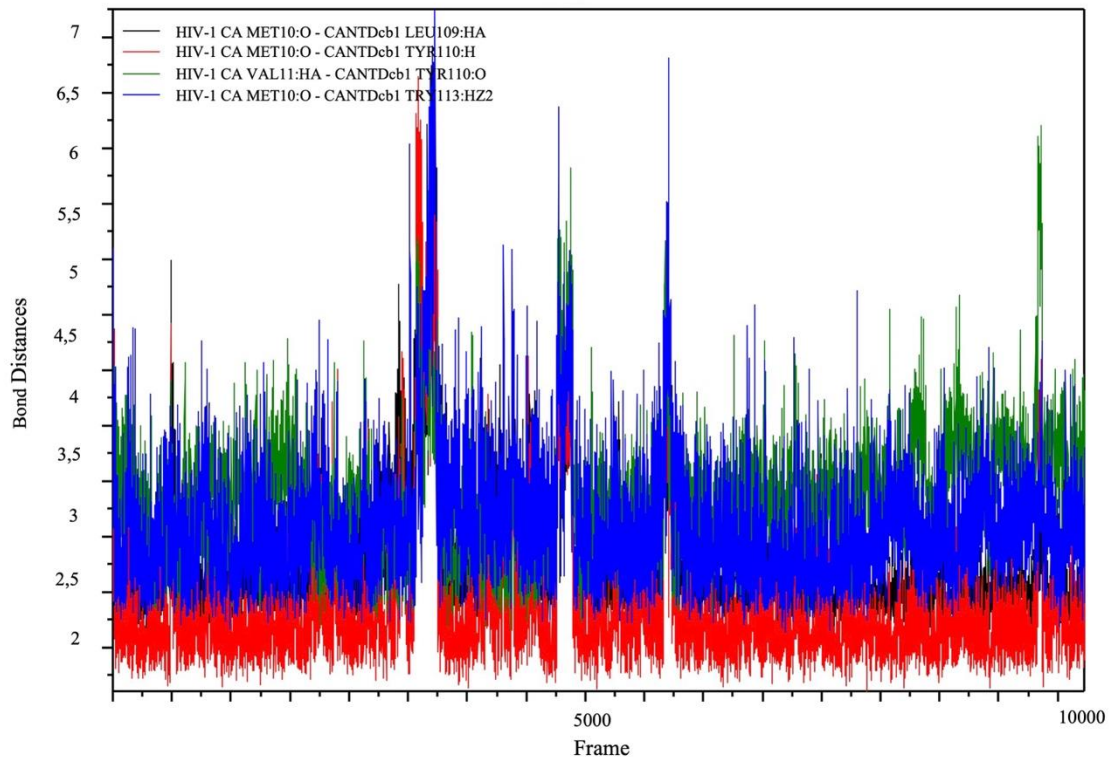


Figure 55. Second run point of view for the eighth complex of the docking results from the Robetta model docking with repulsion terms applied for residues that take part in establishing the necessary interaction to form capsid core binding analysis.

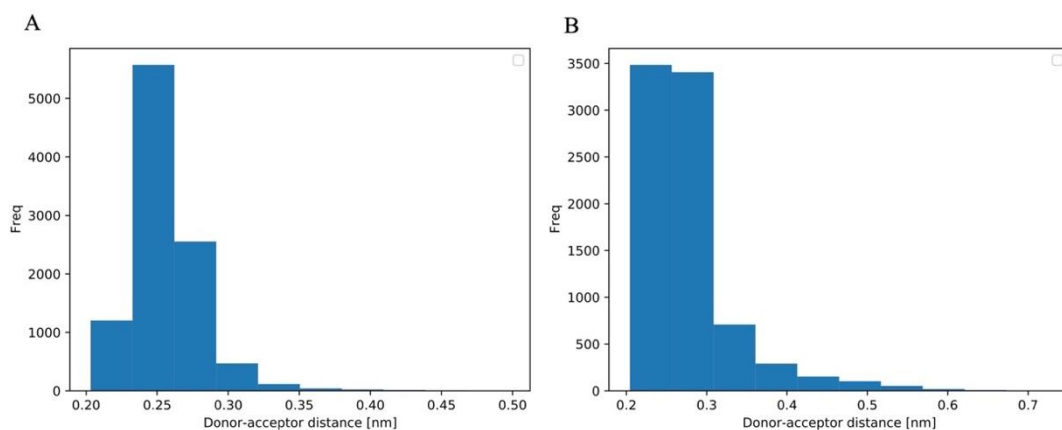


Figure 56. 80th percent analysis of HIV-1 CA Met10:O-CANTDcb1 LEU109:HA. (A) First run evaluation of the bond. (B) Second run evaluation of the bond.

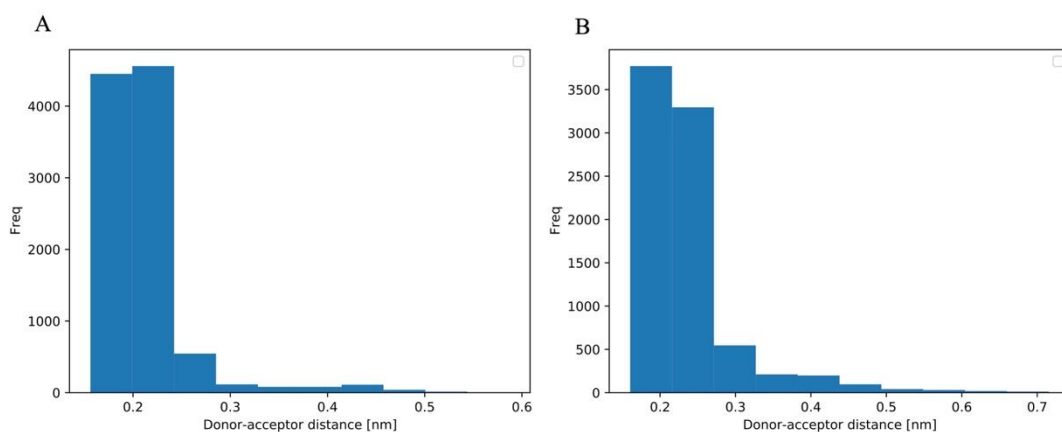


Figure 57. 80th percent analysis of HIV-1 CA Met10:O-CANTDcb1 Tyr110:H. (A) First run evaluation of the bond. (B) Second run evaluation of the bond.

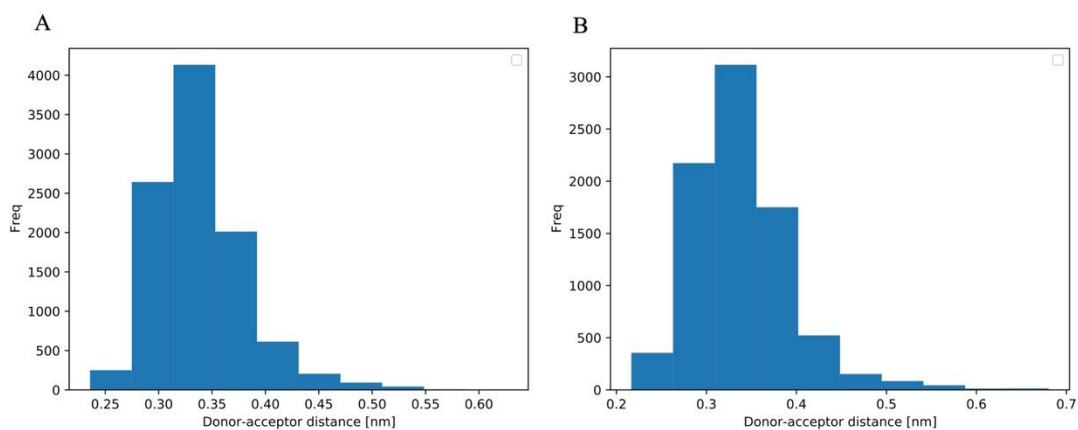


Figure 58. 80th percent analysis of HIV-1 CA Val11:HA-CANTDcb1 Tyr110:O. (A) First run evaluation of the bond. (B) Second run evaluation of the bond.

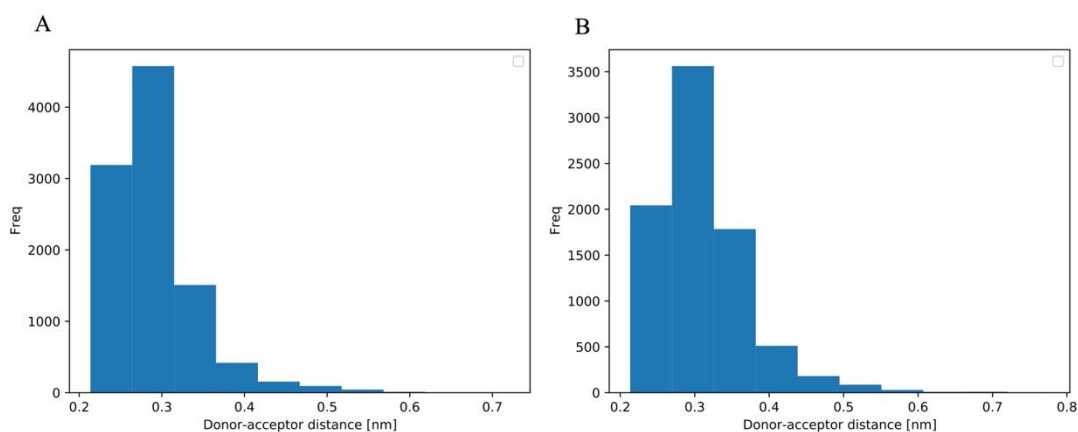


Figure 59. 80th percent analysis of HIV-1 CA Val10:O-CANTDcb1 Tyr113:HZ2. (A) First run evaluation of the bond. (B) Second run evaluation of the bond.

In our next docking run, we used the trRobetta model. Compared to the Robetta, this model had a more open CDR conformation. Among the docking results, we selected the first model. Although the first model does not have any salt bridge, it has a more favorable position (Figure 60). After the evaluation, we decided to perform an MD run on this complex. However, MD simulations for this complex are not yet concluded.

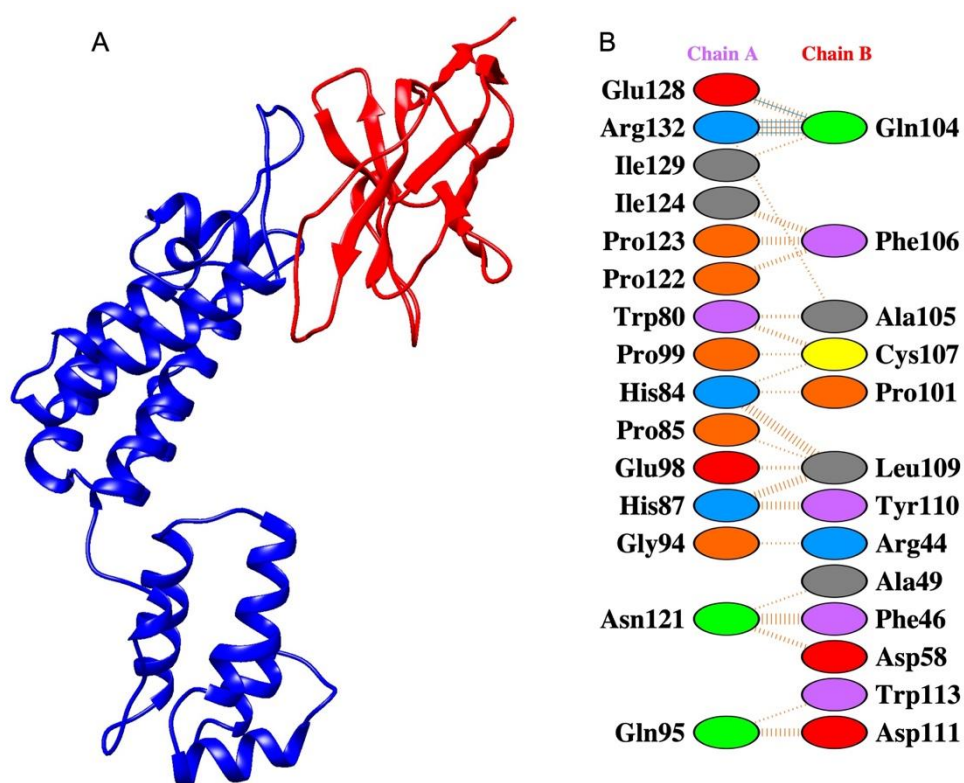


Figure 60. First complex of the docking results from the trRosetta model docking with repulsion terms applied for residues that takes part in establishing necessary interaction to form capsid core. (A) 3D visualization of the docked pose. Figure is prepared on UCSF Chimera. (B) PDBSum result of the docked pose.

As a result of the molecular docking attempts and interaction analysis, we decided that guided docking is the appropriate method to follow. From these docking runs, the eighth complex of the docking results from Robetta model docking on ClusPro with repulsion terms applied for residues that takes part in establishing necessary interaction to form capsid core performed the best. trRosetta was docked with the same parameters, however, MD simulation results are not available yet.

CHAPTER 4

CONCLUSION

Diagnosis of HIV is a crucial step in preventing the spread of the virus. The aim of this study is to understand the interactions between HIV-1 capsid and CANTDcb1, a previously developed nanobody. This information will prove beneficial for developing a diagnostic assay.

We modeled CANTDcb1 3D structure through SwissModel, AlphaFold2, trRosetta, and Robetta. These models were evaluated with structure assessment tests, and trRosetta and Robetta models were selected. Both models proceeded to be docked with HIV-1 capsid protein. Robetta model was docked at ZDock, ClusPro, and Haddock, and trRosetta was docked at ClusPro. Binding data from blind docking studies were extracted to be used as parameters in guided docking. In guided docking, attraction and repulsion terms and active residues were applied. Among the guided docking results, complexes were selected for MD simulations according to visual evaluation and binding analysis. These complexes are:

- Complex 13 of HIV-1 CA:Arg132 – CANTDcb1:Asp111 group,
- Complex 1 of HIV-1 CA:Arg132 - CANTDcb1:Asp99+111 group,
- Complex 1 of the Robetta model docking with repulsion terms applied for residues that take part in establishing necessary interactions to for capsid core
- Complex 8 of the Robetta model docking with repulsion terms applied for residues that take part in establishing necessary interactions to for capsid core,
- Complex 1 of the trRosetta model docking with the same repulsion terms.

A total number of 4 MD simulations are discussed in this study. Models were evaluated according to stability of the starting individual protein structures and protein-

protein complexes throughout the simulation. Number and type of non-covalent interactions (salt bridge, hydrogen bonds and hydrophobic interactions) and buried solvent accessible surface area are important parameters investigated.

In MD simulations of complex 13 of HIV-1 CA:Arg132 – CANTDcb1:Asp111 group and complex 1 of HIV-1 CA:Arg132 - CANTDcb1:Asp99+111 group, complexes were not stable and proteins drifted away from each other. In MD simulations of complex 1 of the Robetta model was stable in the first MD run, however, the other two MD runs of this complex were not stable and positions of the complexes after 100ns were completely different from each other. Complex 8 of the Robetta model was stable in the first and second runs, the third run is not complete yet . In the first two MD runs of complex 8, there are consistently present H-bonds and many hydrophobic interactions are observed in around 750Å² buried surface area between the two proteins. Despite the absence of any electrostatic interaction, HIV-1CA – CANTDcb1 complex stayed close together. Complex 8 of Robetta model has promising results but further docking and MD runs need to be conducted to be able to elucidate the molecular details of HIV-1 Capsid and CANTDcb1 interaction.

REFERENCES

- Alfadhli, A., Romanaggi, C. A., Barklis, R. L., Merutka, I., Bates, T. A., Tafesse, F. G., & Barklis, E. (2021). Capsid-specific nanobody effects on HIV-1 assembly and infectivity. *Virology*, *562*, 19–28. <https://doi.org/10.1016/j.virol.2021.07.001>
- Apetrei, C., Hahn, B., Rambaut, A., Wolinsky, S., Brister, J. R., Keele, B. F., Fraser, C., Singh, A., Abfalterer, W., Fischer, W., Foley, B., Korber, B., Macke, J., Szinger, J. J., Wagh, K., & Yoon, H. (2021). *HIV Sequence Compendium 2021 Los Alamos HIV Sequence Database and Analysis Staff*. <https://www.hiv.lanl.gov/>
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Bao, G., Tang, M., Zhao, J., & Zhu, X. (2021). Nanobody: a promising toolkit for molecular imaging and disease therapy. *EJNMMI Research*, *11*(1), 1–13. <https://doi.org/10.1186/S13550-021-00750-5/TABLES/2>
- Bbosa, N., Kaleebu, P., & Ssemwanga, D. (2019). HIV subtype diversity worldwide. *Current Opinion in HIV and AIDS*, *14*(3), 153–160. <https://doi.org/10.1097/COH.0000000000000534>
- Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, *27*(3), 343–350. <https://doi.org/10.1093/bioinformatics/btq662>
- Benkert, P., Tosatto, S. C. E., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function and Genetics*, *71*(1), 261–277. <https://doi.org/10.1002/prot.21715>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. <https://doi.org/10.1093/NAR/28.1.235>
- Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (n.d.). *Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology OPEN*. <https://doi.org/10.1038/s41598-017-09654-8>
- Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary

- information. *Web Server Issue Published Online*, 42.
<https://doi.org/10.1093/nar/gku340>
- Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., & Schwede, T. (2013). OpenStructure: An integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5), 701–709.
<https://doi.org/10.1107/S0907444913007051>
- Bienert, S., Waterhouse, A., De Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, 45(D1), D313–D319.
<https://doi.org/10.1093/NAR/GKW1132>
- Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoseck, M., Im, W., ... Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), 1545–1614. <https://doi.org/10.1002/jcc.21287>
- Chen, R., & Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Structure, Function and Genetics*, 47(3), 281–294. <https://doi.org/10.1002/prot.10092>
- Cornett, J. K., & Kirn, T. J. (2013). Laboratory diagnosis of HIV in adults: A review of current methods. In *Clinical Infectious Diseases* (Vol. 57, Issue 5, pp. 712–718). <https://doi.org/10.1093/cid/cit281>
- Craveur, P., Gres, A. T., Kirby, K. A., Liu, D., Hammond, J. A., Deng, Y., Forli, S., Goodsell, D. S., Williamson, J. R., Sarafianos, S. G., & Olson, A. J. (2019). Novel intersubunit interaction critical for hiv-1 core assembly defines a potentially targetable inhibitor binding pocket. *MBio*, 10(2).
<https://doi.org/10.1128/mBio.02858-18>
- Deshmukh, L., Schwieters, C. D., Grishaev, A., Ghirlando, R., Baber, J. L., & Clore, G. M. (2013). Structure and Dynamics of Full Length HIV-1 Capsid Protein in Solution NIH Public Access. *J Am Chem Soc*, 135(43), 16133–16147.
<https://doi.org/10.1021/ja406246z>
- Desta, I. T., Porter, K. A., Xia, B., Kozakov, D., & Correspondence, S. V. (2020). Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure*, 28, 1071–1081. <https://doi.org/10.1016/j.str.2020.06.006>
- Doerflinger, S. Y., Tabatabai, J., Schnitzler, P., Farah, C., Rameil, S., Sander, P., Koromyslova, A., & Hansman, G. S. (2016). Development of a Nanobody-Based Lateral Flow Immunoassay for Detection of Human Norovirus. *MSphere*, 1(5).
<https://doi.org/10.1128/msphere.00219-16>

- Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731–1737. https://doi.org/10.1021/JA026939X/SUPPL_FILE/JA026939XS120021128_085857.TXT
- Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D., & Yang, J. (2021). The trRosetta server for fast and accurate protein structure prediction. In *Nature Protocols* (Vol. 16, Issue 12, pp. 5634–5651). Nature Research. <https://doi.org/10.1038/s41596-021-00628-9>
- Engelman, A., & Cherepanov, P. (2012). The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology* *2012 10:4*, *10*(4), 279–290. <https://doi.org/10.1038/nrmicro2747>
- Ganser-Pornillos, B. K., Yeager, M., & Sundquist, W. I. (2008). The structural biology of HIV assembly. *Current Opinion in Structural Biology*, *18*(2), 203–217. <https://doi.org/10.1016/j.sbi.2008.02.001>
- Gray, E. R., Bain, R., Varsaneux, O., Peeling, R. W., Stevens, M. M., & McKendry, R. A. (2018). P24 revisited: A landscape review of antigen detection for early HIV diagnosis. In *AIDS* (Vol. 32, Issue 15, pp. 2089–2102). Lippincott Williams and Wilkins. <https://doi.org/10.1097/QAD.0000000000001982>
- Gres, A. T., Kirby, K. A., Kewalramani, V. N., Tanner, J. J., Pornillos, O., & Sarafianos, S. G. (2015). *X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability*. <https://www.science.org>
- Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *ELECTROPHORESIS*, *30*(S1), S162–S173. <https://doi.org/10.1002/elps.200900140>
- HADDOCK2.4 manual - Analysis*. (2023, July 5).
- Helma, J., Schmidthals, K., Lux, V., Nüske, S., Scholz, A. M., Kräusslich, H. G., Rothbauer, U., & Leonhardt, H. (2012). Direct and Dynamic Detection of HIV-1 in Living Cells. *PLoS ONE*, *7*(11). <https://doi.org/10.1371/journal.pone.0050026>
- HIV Infection | BioNinja*. (n.d.). Retrieved June 20, 2023, from <https://ib.bioninja.com.au/standard-level/topic-6-human-physiology/63-defence-against-infectio/hiv-infection.html>
- Honorato, R. V., Koukos, P. I., Jiménez-García, B., Tsaregorodtsev, A., Verlato, M., Giachetti, A., Rosato, A., & Bonvin, A. M. J. J. (2021). Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Frontiers in Molecular Biosciences*, *8*, 729513. <https://doi.org/10.3389/FMOLB.2021.729513/BIBTEX>

- Hu, Y., Liu, C., & Muyldermans, S. (2017). Nanobody-Based Delivery Systems for Diagnosis and Targeted Tumor Therapy. *Frontiers in Immunology*, 8. <https://doi.org/10.3389/fimmu.2017.01442>
- Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., Grubmüller, H., & Mackerell, A. D. (2016). *charmm36m: an improved force field for folded and intrinsically disordered proteins*. 14(1). <https://doi.org/10.1038/nMeth.4067>
- Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11), 1859–1865. <https://doi.org/10.1002/JCC.20945>
- Johnston, M. I., & Hoth, D. F. (1993). Present Status and Future Prospects for HIV Therapies. *Science*, 260, 1286–1293. <https://doi.org/10.1126/science.7684163>
- Julien, J. P., Cupo, A., Sok, D., Stanfield, R. L., Lyumkis, D., Deller, M. C., Klasse, P. J., Burton, D. R., Sanders, R. W., Moore, J. P., Ward, A. B., & Wilson, I. A. (2013). Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science*, 342(6165), 1477–1483. https://doi.org/10.1126/SCIENCE.1245625/SUPPL_FILE/JULIEN.SM.PDF
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold2. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R., & Vajda, S. (2013). How good is automated protein docking? *Proteins: Structure, Function and Bioinformatics*, 81(12), 2159–2166. <https://doi.org/10.1002/prot.24403>
- Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65(2), 392–406. <https://doi.org/10.1002/PROT.21117>
- Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein-protein docking. *Nature Protocols*, 12(2), 255–278. <https://doi.org/10.1038/nprot.2016.169>
- Kurkcuoglu, Z., & Bonvin, A. M. J. J. (2020). Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein-protein and protein-DNA systems. *Proteins: Structure, Function and Bioinformatics*, 88(2), 292–306. <https://doi.org/10.1002/prot.25802>
- Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, C. L., Mackerell, A. D., Klauda, J. B., & Im, W. (2015). *CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations*

Using the CHARMM36 Additive Force Field.
<https://doi.org/10.1021/acs.jctc.5b00935>

- Liu, C., Lin, H., Cao, L., Wang, K., & Sui, J. (2022). Research progress on unique paratope structure, antigen binding modes, and systematic mutagenesis strategies of single-domain antibodies. In *Frontiers in Immunology* (Vol. 13). Frontiers Media S.A. <https://doi.org/10.3389/fimmu.2022.1059771>
- Lohning, A. E., Levonis, S. M., Williams-Noonan, B., & Schweiker, S. S. (2017). A Practical Guide to Molecular Docking and Homology Modelling for Medicinal Chemists. *Current Topics in Medicinal Chemistry*, 17(18). <https://doi.org/10.2174/1568026617666170130110827>
- Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21), 2722–2728. <https://doi.org/10.1093/bioinformatics/btt473>
- McFadden, W. M., Snyder, A. A., Kirby, K. A., Tedbury, P. R., Raj, M., Wang, Z., & Sarafianos, S. G. (2021). Rotten to the core: antivirals targeting the HIV-1 capsid core. *Retrovirology*, 18(1), 1–24. <https://doi.org/10.1186/s12977-021-00583-z>
- Mishra, P. M., Verma, N. C., Rao, C., Uversky, V. N., & Nandi, C. K. (2020). Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. In *Progress in Molecular Biology and Translational Science* (Vol. 174, pp. 1–78). Elsevier B.V. <https://doi.org/10.1016/bs.pmbts.2020.03.001>
- Mitchell, L. S., & Colwell, L. J. (2018a). Analysis of nanobody paratopes reveals greater diversity than classical antibodies. *Protein Engineering, Design and Selection*, 31(7–8), 267–275. <https://doi.org/10.1093/protein/gzy017>
- Mitchell, L. S., & Colwell, L. J. (2018b). Comparative analysis of nanobody sequence and structure data. *Proteins: Structure, Function and Bioinformatics*, 86(7), 697–706. <https://doi.org/10.1002/prot.25497>
- Miyazaki, Y., Miyake, A., Doi, N., Koma, T., Uchiyama, T., Adachi, A., & Nomaguchi, M. (2017). Comparison of biochemical properties of HIV-1 and HIV-2 capsid proteins. *Frontiers in Microbiology*, 8(JUN). <https://doi.org/10.3389/fmicb.2017.01082>
- Moradi-Kalbolandi, S., Sharifi-K, A., Darvishi, B., Majidzadeh-A, K., Jalili, N., Sadeghi, S., Mosayebzadeh, M., Sanati, H., Salehi, M., & Farahmand, L. (2020). Evaluation the potential of recombinant anti-CD3 nanobody on immunomodulatory function. *Molecular Immunology*, 118, 174–181. <https://doi.org/10.1016/j.molimm.2019.12.017>
- Morrison, C. (2019). Nanobody approval gives domain antibodies a boost. *Nature Reviews. Drug Discovery*, 18(7), 485–487. <https://doi.org/10.1038/D41573-019-00104-W>

- Nelson, D. L., & Cox, M. M. (2017). *Lehninger Principles of Biochemistry* (7th Ed.). W. H. Freeman.
- Perilla, J. R., Hadden-Perilla, J. A., Gronenborn, A. M., & Polenova, T. (2021). Integrative structural biology of HIV-1 capsid protein assemblies: combining experiment and computation. *Current Opinion in Virology*, 48(June 2020), 57–64. <https://doi.org/10.1016/j.coviro.2021.03.005>
- Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T., & Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics (Oxford, England)*, 30(12), 1771–1773. <https://doi.org/10.1093/BIOINFORMATICS/BTU097>
- Raybould, M. I. J., Kovaltsuk, A., Marks, C., & Deane, C. M. (2021). CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, 37(5), 734–735. <https://doi.org/10.1093/bioinformatics/btaa739>
- Rihn, S. J., Wilson, S. J., Loman, N. J., Alim, M., Bakker, S. E., Bhella, D., Gifford, R. J., Rixon, F. J., & Bieniasz, P. D. (2013a). Extreme Genetic Fragility of the HIV-1 Capsid. *PLoS Pathogens*, 9(6). <https://doi.org/10.1371/journal.ppat.1003461>
- Rihn, S. J., Wilson, S. J., Loman, N. J., Alim, M., Bakker, S. E., Bhella, D., Gifford, R. J., Rixon, F. J., & Bieniasz, P. D. (2013b). Extreme Genetic Fragility of the HIV-1 Capsid. *PLoS Pathogens*, 9(6). <https://doi.org/10.1371/journal.ppat.1003461>
- Salmaso, V., & Moro, S. (2018). Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. In *Frontiers in Pharmacology* (Vol. 9, Issue AUG). Frontiers Media S.A. <https://doi.org/10.3389/fphar.2018.00923>
- Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics*, 36, 1765–1771. <https://doi.org/10.1093/bioinformatics/btz828>
- Su, H., Wang, W., Du, Z., Peng, Z., Gao, S.-H., Cheng, M.-M., Yang, J., Su, H., Wang, W., Du, Z., Peng, Z., Yang, J., Gao, S.-H., & Cheng, M.-M. (2021). *Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates*. <https://doi.org/10.1002/adv.202102592>
- Sun, S., Ding, Z., Yang, X., Zhao, X., Zhao, M., Gao, L., Chen, Q., Xie, S., Liu, A., Yin, S., Xu, Z., & Lu, X. (2021). Nanobody: A Small Antibody with Big Implications for Tumor Therapeutic Strategy. *International Journal of Nanomedicine, Volume 16*, 2337–2356. <https://doi.org/10.2147/IJN.S297631>
- The Structural Biology of HIV*. (n.d.). Retrieved June 20, 2023, from <https://cdn.rcsb.org/pdb101/learn/resources/structural-biology-of-hiv/index.html#>

- UNAIDS 2021 Adults and children living with HIV. (2021). <https://www.unaids.org/en/resources/documents/2022/core-epidemiology-slides>.
- Vajda, S., Yueh, C., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R., & Kozakov, D. (2017). New additions to the ClusPro server motivated by CAPRI. *Proteins: Structure, Function and Bioinformatics*, 85(3), 435–444. <https://doi.org/10.1002/prot.25219>
- Valdés-Tresanco, M. S., Molina-Zapata, A., Pose, A. G., & Moreno, E. (2022). Structural Insights into the Design of Synthetic Nanobody Libraries. *Molecules*, 27(7), 2198. <https://doi.org/10.3390/molecules27072198>
- Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastiris, P. L., Karaca, E., Melquiond, A. S. J., Van Dijk, M., De Vries, S. J., & Bonvin, A. M. J. J. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4), 720–725. <https://doi.org/10.1016/J.JMB.2015.09.014>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2021). AlphaFold2 Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50. <https://doi.org/10.1093/nar/gkab1061>
- Visseaux, B., Damond, F., Matheron, S., Descamps, D., & Charpentier, C. (2016). Hiv-2 molecular epidemiology. *Infection, Genetics and Evolution*, 46, 233–240. <https://doi.org/10.1016/J.MEEGID.2016.08.010>
- Wang, C., Wang, Q., Ji, B., Pan, Y., Xu, C., Cheng, B., Bai, B., & Chen, J. (2018). The Orexin/Receptor System: Molecular Mechanism and Therapeutic Potential for Neurological Diseases. In *Frontiers in Molecular Neuroscience* (Vol. 11). Frontiers Media S.A. <https://doi.org/10.3389/fnmol.2018.00220>
- Wang, W., Peng, Z., & Yang, J. (2022). Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science*, 2, 804–814. <https://doi.org/10.1038/s43588-022-00373-3>
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1), W296–W303. <https://doi.org/10.1093/nar/gky427>
- Wilk, T., Gross, I., Gowen, B. E., Rutten, T., De Haas, F., Welker, R., Krausslich, H.-G., Krausslich, K., Boulanger, P., & Fuller, S. D. (2001). Organization of Immature Human Immunodeficiency Virus Type 1. *JOURNAL OF VIROLOGY*, 75(2), 759–771. <https://doi.org/10.1128/JVI.75.2.759-771.2001>

- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science: A Publication of the Protein Society*, 27(1), 293. <https://doi.org/10.1002/PRO.3330>
- Wilton, E. E., Opyr, M. P., Kailasam, S., Kothe, R. F., & Wieden, H. J. (2018). SdAb-DB: The Single Domain Antibody Database. *ACS Synthetic Biology*, 7(11), 2480–2484. https://doi.org/10.1021/ACSSYNBIO.8B00407/ASSET/IMAGES/LARGE/SB-2018-00407W_0004.JPEG
- Yu, H., & Dalby, P. A. (2020). A beginner's guide to molecular dynamics simulations and the identification of cross-correlation networks for enzyme engineering. *Methods in Enzymology*, 643, 15–49. <https://doi.org/10.1016/BS.MIE.2020.04.020>
- Zhang, Y., Qian, H., Love, Z., & Barklis, E. (1998). Analysis of the Assembly Function of the Human Immunodeficiency Virus Type 1 Gag Protein Nucleocapsid Domain. In *JOURNAL OF VIROLOGY* (Vol. 72, Issue 3). <https://journals.asm.org/journal/jvi>
- Zheng, L., Alhossary, A. A., Kwoh, C. K., & Mu, Y. (2018). Molecular dynamics and simulation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3, pp. 550–566). Elsevier. <https://doi.org/10.1016/B978-0-12-809633-8.20284-7>

APPENDIX A

Python script of AlphaFold2:

```
import os

import mock

import numpy as np

import pickle

import py3Dmol

from typing import Dict

from AlphaFold2.common import protein

from AlphaFold2.data import pipeline

from AlphaFold2.data import templates

from AlphaFold2.model import data

from AlphaFold2.model import config

from AlphaFold2.model import model

# setup which models to use

# note for demo, we are only using model_1

model_runners = {}

models = ["model_1"] #,"model_2","model_3","model_4","model_5"]

for model_name in models:
```

```

model_config = config.model_config(model_name)

model_config.data.eval.num_ensemble = 1

model_params = data.get_model_haiku_params(model_name=model_name,
data_dir=".")

model_runner = model.RunModel(model_config, model_params)

model_runners[model_name] = model_runner

def mk_mock_template(query_sequence):

    # mock template features

    output_templates_sequence = []

    output_confidence_scores = []

    templates_all_atom_positions = []

    templates_all_atom_masks = []

    for _ in query_sequence:

        templates_all_atom_positions.append(np.zeros((templates.residue_constants.atom_type
_num, 3)))

        templates_all_atom_masks.append(np.zeros(templates.residue_constants.atom_type_nu
m))

        output_templates_sequence.append('-')

        output_confidence_scores.append(-1)

    output_templates_sequence = ".join(output_templates_sequence)

```

```

templates_aatype =
templates.residue_constants.sequence_to_onehot(output_templates_sequence,

templates.residue_constants.HHBLITS_AA_TO_ID)

template_features = {'template_all_atom_positions':
np.array(templates_all_atom_positions)[None],

'template_all_atom_masks': np.array(templates_all_atom_masks)[None],

'template_sequence': [f'none'.encode()],

'template_aatype': np.array(templates_aatype)[None],

'template_confidence_scores': np.array(output_confidence_scores)[None],

'template_domain_names': [f'none'.encode()],

'template_release_date': [f'none'.encode()]

return template_features

def predict_structure(
    prefix: str,
    data_pipeline: pipeline.DataPipeline,
    model_runners: Dict[str, model.RunModel],
    random_seed: int):

    """Predicts structure using AlphaFold2 for the given sequence."""

```

```

# Get features.

feature_dict = data_pipeline.process()

# Run the models.

plddts = {}

for model_name, model_runner in model_runners.items():

    processed_feature_dict = model_runner.process_features(feature_dict,
random_seed=random_seed)

    prediction_result = model_runner.predict(processed_feature_dict)

    unrelaxed_protein =
protein.from_prediction(processed_feature_dict,prediction_result)

    unrelaxed_pdb_path = f'{prefix}_unrelaxed_{model_name}.pdb'

    plddts[model_name] = prediction_result['plddt']

with open(unrelaxed_pdb_path, 'w') as f:

    f.write(protein.to_pdb(unrelaxed_protein))

return plddts

query_sequence =
"QVQLVESGGGLVQAGGSLRLSCAASGYFSSYAMGWFRQAPGKEREFVAAISW
IESTTDYADSVKGRFTISRDNAKKTLHLQMNSLKPEDTAVYYCAACDIPFGQAF
CALYDYWGQGTQVTVSSKLAAALE"

# mock pipeline for testing

data_pipeline_mock = mock.Mock()

```

```
data_pipeline_mock.process.return_value = {  
    **pipeline.make_sequence_features(sequence=query_sequence,  
                                     description="none",  
                                     num_res=len(query_sequence)),  
    **pipeline.make_msa_features(msas=[[query_sequence]],  
                                 deletion_matrices=[[0]*len(query_sequence)]),  
    **mk_mock_template(query_sequence)  
}
```

```
plddts = predict_structure(  
    prefix="test",  
    data_pipeline=data_pipeline_mock,  
    model_runners=model_runners,  
    random_seed=0)
```