METHODS OF ELECTROSTATIC ANALYSIS FOR BIOMOLECULAR STRUCTURES

by

Scott P. Morton

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in

Computational Science

Middle Tennessee State University

December, 2020

Dissertation Committee:

Dr. Joshua L. Phillips, *Chair*

Dr. Jing Kong

Dr. Rachel Leander

Dr. Stephen Wright

For my mother, father and sister, to whom I promised to finish this effort before they passed, a promise I am truly proud to keep.

ACKNOWLEDGMENTS

My deepest thanks and appreciation to my advisor Dr. Joshua L. Phillips, for the opportunity to explore Molecular Dynamics and bio-molecular electrostatic analysis. His depth of knowledge and patience with my quirky points of view are admirable and I am honored to have been taught by him. I would also like to thank Dr. Ralph M. Butler for all the time spent discussing points of education and his bottomless depth of knowledge in parallel processing. To the entire College of Basic and Applied Sciences and the Computer Sciences department at MTSU in particular, I am so humbled to be educated by some of the finest scientists on Earth. Thank you.

I thank my committee members for taking the time to read this dissertation, running me through the breadth of content and for providing guidance, feedback and support.

Additionally, I want to thank the entire Computational Sciences program for the guidance provided by Dr. John Wallin, the programs director, and the remaining program staff for all of your support.

Most importantly, I want to thank my wife and partner in life for putting up with this for the past eleven years. This was difficult to accomplish and I simply could not have done so without your support.

ABSTRACT

Decades of research have yet to provide a vaccine for the human immunodeficiency virus which causes acquired immune deficiency syndrome. The virus sequence varies at high rates once infection occurs, but changes in the RNA sequence that defines the virus are further convoluted by the limited number of variations that can infecting another host during heterosexual intercourse. Current theoretical research has turned attention to genital mucosa pH levels over systemic pH levels in the quest to determine the transmission bottleneck observed. Previous research in this field developed a computational approach for determining pH sensitivity that indicated higher potential for transmission at mucosa pH levels present during intercourse. The process was extended to incorporate multiple program / multiple data operations, advanced compression for accumulated data and a principal component analysis (PCA)-based machine learning technique for classification of gp120 proteins against a known transmitted variant; This method is called Biomolecular Electro-Static Indexing (BESI). The process was further extended to the residue level by a method termed Electrostatic Variance Masking (EVM) and used in conjunction with BESI to determine structural differences present among various subspecies across HIV Clade. Results indicate that variable loop composition outside of the core selected by EVM may be responsible for binding affinity observed in many other studies and that pH modulation of residues selected by EVM may influence specific regions of the viral envelope protein involved in protein-protein interactions. Further research has shown that pH affects binding free energy, a measure of contribution solvation has for interactions between two molecules. These data indicate that a functional range of pH exists and is different for gp120/CD4 interactions compared to gp120/broadly neutralizing antibody interactions. The methods presented in this dissertation have been applied extensively to HIV gp120 proteins individually and in protein-protein interaction simulations. Protein interaction simulations for gp120 with human CD4 protein (found on the surface of T lymphocytes, mono-

cytes, dendritic cells and brain microglia), and gp120 with broadly neutralizing antibodies provide unique insight not easily or economically achievable with traditional laboratory methods. The pipeline and methods are easily adapted to other protein structures, such as SARS-CoV-2 spike protein with human angiotensin-converting enzyme 2, and should provide valuable and unique insights into interactions where environmental factors, like pH, may modulate how two proteins interact.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I : PROLOGUE

Life as we know it functions through proteins: linear assemblies of amino acids; amino acids are monomer molecules comprised of a central carbon atom bonded to a hydrogen atom (CH), referred to as the alpha carbon, in a covalent bond with an amino group (NH_2), a carboxyl group ($COOH$), and an R group as shown in Figure 1.1. The characteristic that distinguishes the approximately twenty unique amino acids is the composition of the side chain bonded to the alpha carbon referred to as the R group. The amino group of one amino acid may bind to the carboxyl group of another in what is referred to as a peptide bond. Polypeptides are series of amino acids, linked together by peptide bonds, with an amino terminator at one end and a carboxyl terminator on the opposite end, as shown in Figure 1.2. Proteins are specific sequences of amino acids that constitute the building blocks of biological systems that form organisms from simple single cell life forms to highly complex systems, such as human beings.

This dissertation explores protein structure electrostatic charges and the effects environmental pH has on structure charge. These methods are useful for studying individual proteins or interactions between two proteins, provided the assemblies have experimentally determined three dimensional representations available. Methods of determination may be X-Ray Crystallography (Drenth and Mesters, 2007), Nuclear Magnetic Resonance Spectroscopy (Cavanagh, 2007), Three Dimensional Electron Microscopy (Frank, 2006) or any other means of producing 3D coordinate representations of molecules. For protein interaction analysis, these methods require three dimensional representations of the individual protein structures in isolation as well as the bound structures.

Such experimentally determined structures are available for human immunodeficiency virus (HIV) gp120 glycoprotein, human immune system T-cell CD4 substructures, and broadly neutralizing antibodies. These proteins are the subject of many studies in the field of HIV research, particularly in terms of transmission and immune system response. For

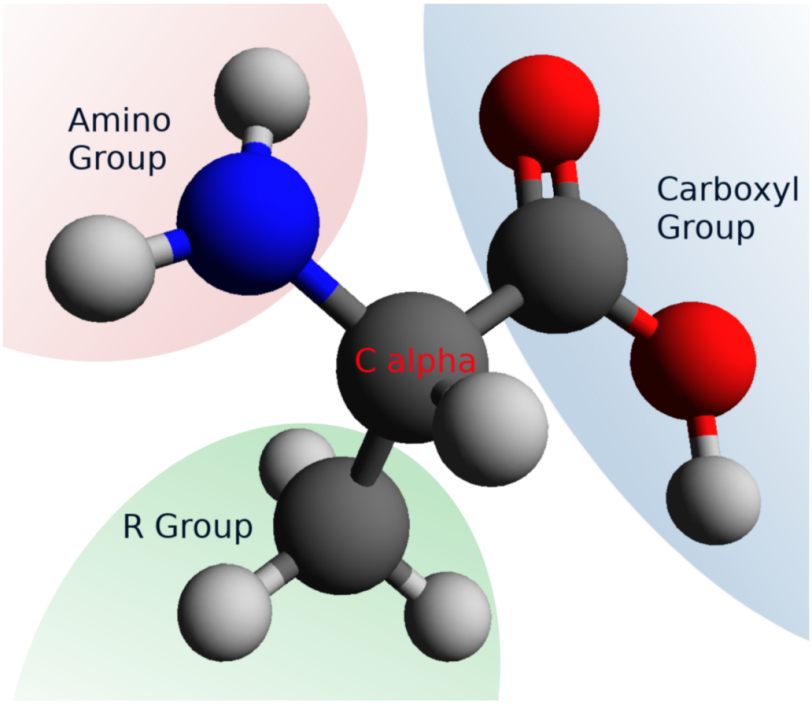

Figure 1.1: The amino acid Alanine showing the amino group (red shading) bound to the alpha carbon *CH* (center) that is bound to the carboxyl group (blue shading). The R group (green shading) for Alanine is CH_3. Carbon is depicted as dark gray, hydrogen is light gray, oxygen is red and nitrogen is blue. Amino acid produced by Avogadro (Hanwell et al., 2012).

more than thirty years, scientists have been exploring potential avenues towards a vaccine for HIV and while progress been made, no vaccine has been produced as of this writing.

Researchers in the field of HIV have concluded that an effective vaccine requires a greater understanding of the mechanics involved with the infection process (Haynes and Mascola, 2017; Fauci, 2016; Haynes et al., 2016; Mascola and Haynes, 2013). To fill this need requires a structural analysis of the molecules involved and how knowledge of the environment impacts molecular structure and molecular interactions. Previous research developed a pipeline for generating electrostatic surface charge data of molecular structures, to provide a novel means of investigating how molecular interactions are modulated by pH.

The next chapter provides background information on HIV infection and a survey of relevant research from the literature. Data acquisition is explained in Chapter III, includ-

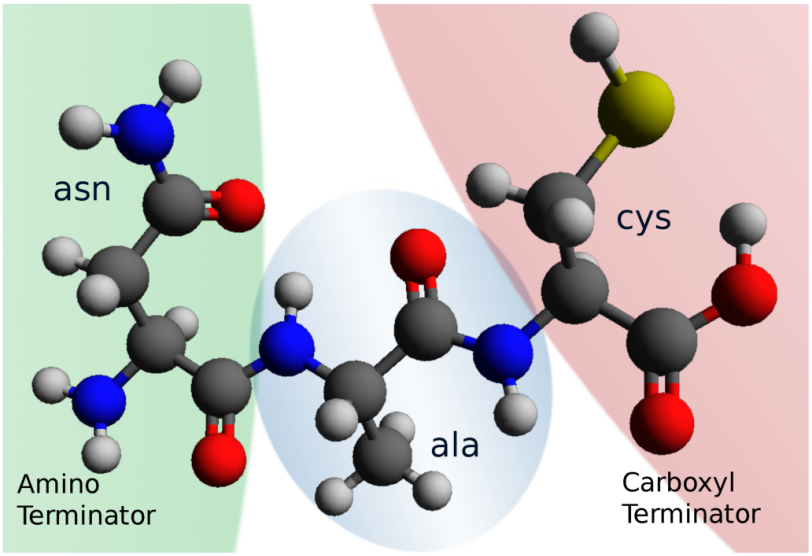

Figure 1.2: A simple protein chain with shading to distinguish each amino acid: Asparagine (green), Alanine (blue), and Cysteine (red). The amino terminal is to the left and the carboxyl terminal is to the right. Carbon is depicted as dark gray, hydrogen is light gray, oxygen is red, sulfur is yellow and nitrogen is blue. Protein chain produced by Avagadro (Hanwell et al., 2012).

ing details around the use and execution of third party applications. Chapter IV provides methods of analysis, including detailed results for each approach.

The following are related publications:

> "pH Dependent Binding Energies of Broadly Neutralizing Antibodies" (In preparation)
>
> Morton, Scott P., and Joshua L. Phillips. "Computational Electrostatics Predict Variations in SARS-CoV-2 Spike and Human ACE2 Interactions." BioRxiv, Cold Spring Harbor Laboratory, June 2020, p. 2020.04.30.071175, doi:10.1101/2020.04.30.071175. (Submitted for publication)
>
> Morton, Scott P., et al. "High-Throughput Structural Modeling of the HIV Transmission Bottleneck." Proceedings of the 2017 IEEE International Conference on Bioin-

formatics and Biomedicine - BIBM-HPCB '17, vol. 2017-Janua, IEEE Press, 2017, doi:10.1109/BIBM.2017.8217952.

Morton, Scott P., et al. "Sub-Class Differences of PH-Dependent HIV GP120-CD4 Interactions." Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18, ACM Press, 2018, pp. 663–68, doi:10.1145/3233547.3233711.

Morton, Scott P., et al. "The Molecular Basis of PH-Modulated HIV Gp120 Binding Revealed." Evolutionary Bioinformatics, vol. 15, SAGE Publications Sage UK: London, England, Jan. 2019, p. 117693431983130, doi:10.1177/1176934319831308.

Morton, Scott P., et al. "A JSON-Based Markup Language for Deploying Virtual Clusters via Docker." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications PDPTA'16, CSREA Press ©, 2016, pp. 251–57, http://worldcomp-proceedings.com/proc/p2016/PDP3139.pdf.

CHAPTER II : HUMAN IMMUNODEFICIENCY VIRUS

Transmission

The Human Immunodeficiency Virus (HIV) was first identified in 1983; vaccine research ensued immediately and is still underway at the time of this writing. HIV is typically transmitted in a mucosa pool during sexual intercourse; other means of transmission may include, but are not limited to: blood transfusions, sharing syringes, and organ transplants. While these additional methods are potential means of infection, they obfuscate the problem of transmission by introducing numerous contextual variables. Many proteins on the external surface of the virion assist with attachment to the target cell, however, the initial interaction is between the viral envelope (Env) glycoprotein gp120 and human glycoprotein CD4 (Klatzmann et al., 1984). CD4 is found on the surface of T lymphocytes, monocytes, dendritic cells, and brain microglia (Wyatt, 1998). Wyatt et al. also explain in detail that the major function of CD4 binding is to induce exposure of chemo-kine receptors through conformational changes in gp120 which facilitate the process of membrane fusion and leads to the eventual introduction of reproductive material into the host cell required for proliferation to take place.

Figure 2.1 presents a simple diagram of the infection process: An HIV virion attaches to the host cell membrane at CD4 [1] and penetrates the cell [2]. Once entry into the host cell is completed, viral RNA is injected into the cell and undergoes reverse transcription to produce viral DNA [3]. The viral DNA is integrated with cell DNA [4] where transcription produces messenger RNA (mRNA) [5] which is translated into protein structures [6] and assembled into a viral core [7]. The completed virus is ejected from the cell through a process called budding [8] and released [9]. Glycoprotein gp120 (blue) and host cell protein CD4 (red) are expanded for clarity to the left of the diagram, the CD4 binding site (CD4bs) is indicated by the green circle. Because the interaction between gp120 and CD4

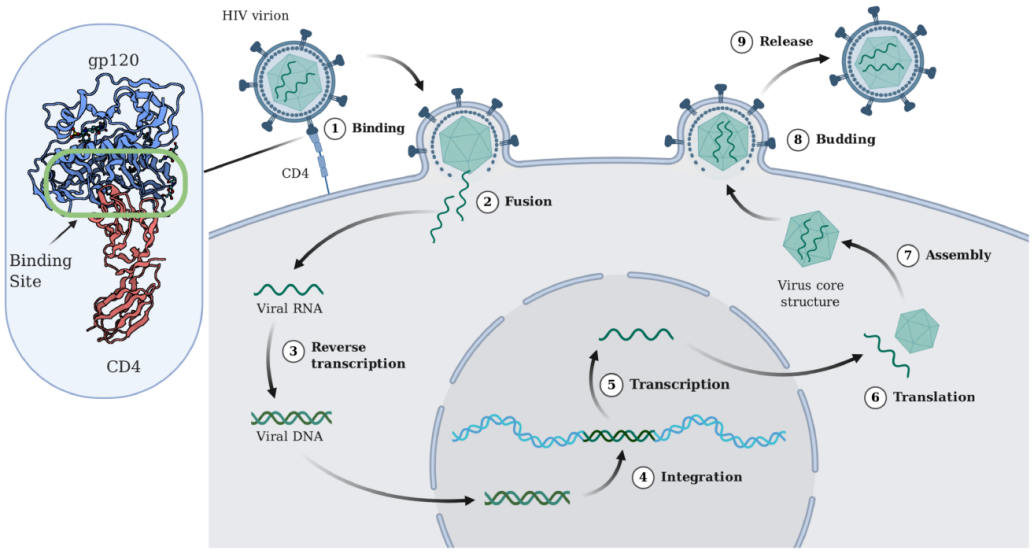

Figure 2.1: Viral envelope glycoprotein gp120 (blue) to host T-cell periphery protein CD4 (red) binding must take place in order for HIV to infect another cell. The binding site is indicated by the green circle. An HIV virion attaches to the host cell membrane at CD4 and begins to penetrate the cell. Once entry into the host cell is completed, the retro virus forces the cell to replicate the viral genetic code repeatedly to proliferate the virus.

initiates the infection process, variables that effect this interaction; e.g. variations in gp120, broadly neutralizing antibodies (bnAbs, see below) and pH, are of interest. CD4 is a functional protein specific to immune cell types previously mentioned, thus it is not subject to distinguishing sequence changes over time *in vivo*. The same cannot be said of bnAbs, as these proteins constitute a dynamic immune response by the body to deal with infections by nature.

In contrast to CD4 and bnAbs, the virus is subject to a high rate of evolutionary change after introduction into the body. HIV uses reverse transcription, the process of converting RNA into DNA, for viral replication upon entry into the target cell. Reverse transcription ultimately has viral DNA integrated into the nucleus of the host cell and is transcribed into mRNA as a template for protein assembly. It is during this process that variations of the HIV genome occurs and most interesting in this case is gp120 that must perform the

binding action. Thus, every infected cell has the potential to generate a new variation of HIV based on random errors in the reverse transcription process.

Clinical studies have shown that the majority of HIV variants present in the genital tract is not responsible for transmission from one host to another (Boeras et al., 2011); this observation suggest that some mechanism, circumstance or combination of the two stifles or enhances transmission. These data indicate a cyclical process in which the high rate of viral evolution determines the potential for transmission between hosts after the initial infection takes place. The following studies support this conjecture:

Researchers investigating variable loop lengths in gp120 glycoprotein also examined evolutionary relationships through the use of maximum likelihood (ML) trees and discovered what was termed as an "extreme bottleneck" related to transmission (Derdeyn et al., 2004), in other words, the majority of HIV virions are incapable of transmitting between hosts. Although the study concludes that assertions of a bottleneck do not constitute conclusive proof of the 'bottleneck' hypothesis, it does provide an avenue of investigation that had not been pursued previously.

Another study used a mathematical model of random viral evolution along with phylogenetic tree constructions to study the transmission bottleneck. The results suggest that "78 out of 102 subjects were infected by a single variant of the virus" (Keele et al., 2008). Keele et al. also revealed that the 24 remaining subjects were infected by fewer than 6 variants of HIV. These data support the hypothesis that there exists a bottleneck in transmission and raises the need for further investigation.

Researchers have also cited a bottleneck in transmission while linking methods that mitigate the mucosal barrier to inflammatory genital infections (Haaland et al., 2009). This information provides a direct link to mucosa as an agent that inhibits HIV transmission.

In October of 2011, researchers published a study directly focused on the transmission bottleneck that may be due to genetic diversity (Boeras et al., 2011). The research involved

a group of couples where one person per pair was infected and the other subject was expected to become infected. This study provided scientists with a unique opportunity to capture near incident transmission variants of the virus. In at least one case, researchers were able to identify the original variant that crossed the barrier and established a *de novo* infection.

With an identified transmitted founder (a variant in the body less than 6 months) and source variant (the donor virus that crossed the transmission barrier), an opportunity to search for methods of analysis with a potential to predict HIV variants that could support crossing the transmission barrier may be possible.

Broadly Neutralizing Antibodies

As previously stated, bnAbs are a biological response to intrusion of the body by antigens associated with pathogens, such is the case with viral infections. While the process culminating in this response is outside the scope of this research, the structures (bnAbs) produced by this response are of pivotal importance. These structures are specific to the antigen present and therefore unique to specific types of infection, such as HIV.

Current research in the field of HIV has largely focused on bnAbs. Figure 2.2 displays a gp120 (top) bound to a bnAb (bottom). The bnAb is a multi-chain molecule represented in red and gray to differentiate the heavy and light chains, respectively, that comprise the protein. Fragments of gp120 selected via evolutionary sequence analysis and computational optimization to potentially invoke the production of bnAbs have previously been employed in work on vaccine production (Fischer et al., 2007). More recently, our understanding of how bnAbs interact with gp120 has improved; scientist have identified several approaches to these interactions involving regions of gp120 at the: CD4 binding site (CD4bs), variable loop 2 (V2) apex, variable loop 3 (V3) glycan regions, and membrane-proximal external region (MPER) of sub-unit protein gp41, which lies beneath gp120 in unbound conformations.

Figure 2.2: A gp120 (top) bound to a broadly neutralizing antibody (bottom). The bnAb is colored in red for the heavy chain and gray for the light chain.

Previous research provided an in-depth analysis of gp120 binding functions and acknowledges antibodies that recognize conserved and discontinuous gp120 epitopes (folded amino acid chains) , which experience greater exposure after CD4 binding and are potent inhibitors of gp120 (Wyatt, 1998). Furthermore, Wyatt et al. suggest that disassembled gp120 proteins elicit most of the antibody response to these viral components, but are unable to bind properly with the substructures and therefore cannot inhibit infection. Later research clarifies (Wyatt, 1998) in that recombinant monomeric gp120 induced antibodies are not effective against circulating primary viruses and do not prevent transmission (Schultz and Bradac, 2001).

Schultz and Bradac (2001) goes on to express that laboratory strains of the virus were much more sensitive to neutralization than wild types. Schultz et al. also points out that

neutralizing antibodies induced by laboratory HIV strains were ineffective against wild types of the virus and suggest that individuals who develop bnAbs as a result of infection may provide more potent bnAb variants. This research also suggests that induced variations may provide a key to developing more effective bnAbs.

Deletion of variable loops V1/V2 and V3 from gp120 as a means to increase binding efficacy of antibodies suggest that these variable loops shield the CD4bs, thereby limiting the ability of antibody recognition (Pantophlet and Burton, 2006; Sullivan et al., 1998; Cao et al., 1997; Wyatt et al., 1995, 1993).

Restrictions on the effectiveness a bnAb may have against a broad range of gp120 variations has reduced the field from which a potent bnAb may emerge. Two such bnAbs, 3BNC117 and VRC01, have been the subject of studies showing forward direction towards a vaccine (Kwon et al., 2018; Bar et al., 2016; Caskey et al., 2015; Scheid et al., 2016; Schoofs et al., 2016).

Work to engineer bnAbs that target gp120 at CD4bs that address the shortcoming of wild types, is ongoing (LaBranche et al., 2019; Kwon et al., 2018). Variants of bnAb 10E8 engineered to posses hydrophobic or positively charged amino acids have increased potency and undiminished functional breadth (Kwon et al., 2018). LaBranche et al. (2019) also provides examples of wild types and engineered variants of CH235 that display increased potency and undiminished breadth.

Electrostatic analysis may inform work, to engineer bnAbs, by providing mechanistic insights into the binding function of variations in the structure of bnAbs.

CHAPTER III : CALCULATING ELECTROSTATICS

This research extends upon a pipeline previously developed to determine a so called electrophoretic fingerprint of the gp120 trimer. Scientists hypothesized that electrophoretic mobility (EM) (Mehrishi and Bauer, 2002; Richmond and Fisher, 1973) could be applied to study proteins across pH titrations and salinities of mucosal and other fundamental compartments (Stieh et al., 2013). Stieh et al. performed laboratory experiments and developed a computational approach to compare against. The resulting pipeline produced electrostatic charge data for the surface of bound and unbound conformations of gp120 as tables of 61 titrations by the number of models generated, as shown in Figure 3.1 (top and middle respectively). The difference of results as $b - ub$, where b the bound data and ub is the unbound data, produce the electrophoretic fingerprint (bottom) of Figure 3.1.

Electrostatic Surface Charge Pipeline

The pipeline of (Stieh et al., 2013) is enhanced through the use of full structure assemblies, energy minimization, advanced compression techniques and a fully automatic execution environment based on standard message passing interface (MPI) functions and direct system calls to general public licensed (GPL) third party tools. Figure 3.2 shows the path taken through the pipeline and all third party utilities invoked. An explanation of each step in the pipeline, including significant parameters, will be followed by an explanation of the execution environment and configuration methods.

Structure Modeling

Structure modeling is facilitated through comparative modeling by Modeller (Sali and Blundell, 1993). Comparative modeling uses experimentally determined protein structures, which may be generated by X-Ray Crystallography, Nuclear Magnetic Resonance Spectroscopy or Three Dimensional Electron Microscopy (Drenth and Mesters, 2007; Cavanagh, 2007; Frank, 2006) to predict variable loops and other ligands. The predicted

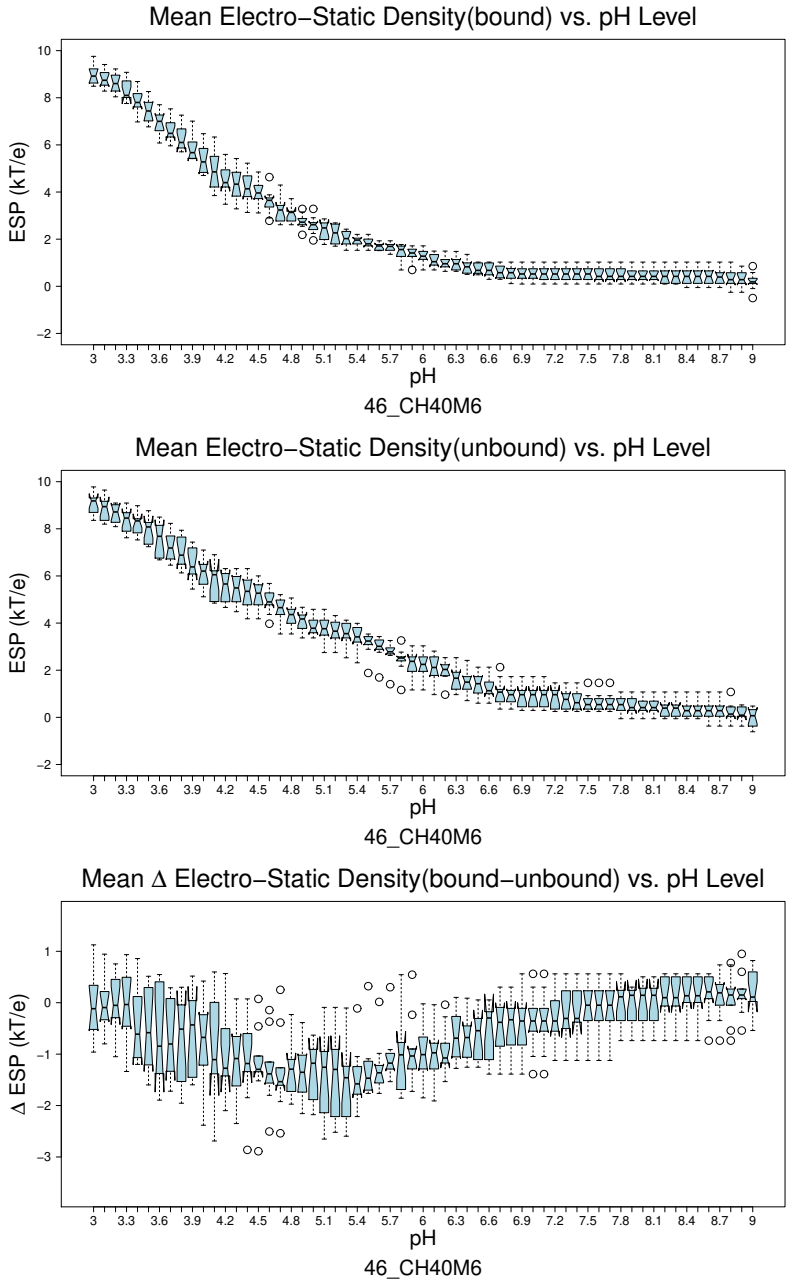

Figure 3.1: The electrostatic potential map of a bound (top) and unbound (top middle) gp120 showing the slight variations from approximately pH 4.5 to 6.0. The electrophoretic fingerprint (bottom) is the result of subtracting the unbound data from the bound data.

protein is possible through comparison of one or more closely related template proteins to produce a useful theoretical model of the undetermined sequence (Eswar et al., 2002; Martí-Renom et al., 2000; Fiser, 2004; Misura and Baker, 2005; Petrey and Honig, 2005;

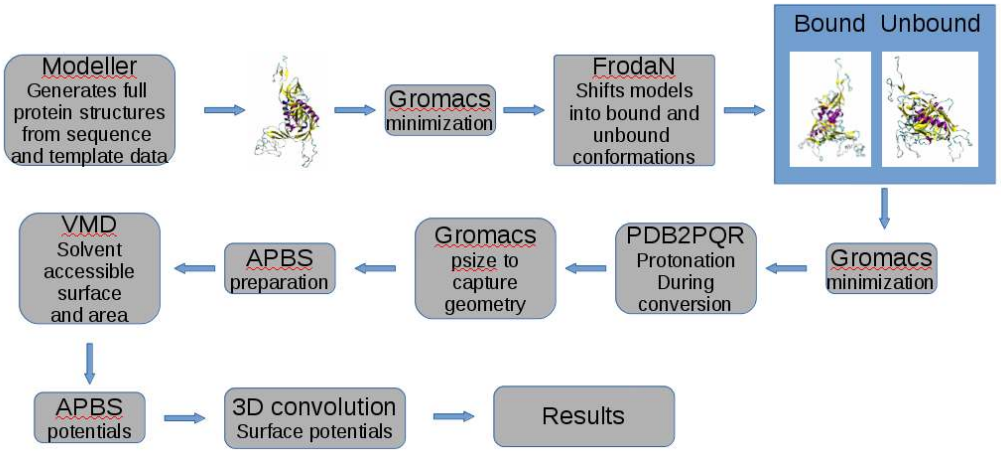

Figure 3.2: The electrostatic surface charge pipeline has many steps to process utilizing an array of external utilities.

Misura et al., 2006). Closely related template proteins are one of three factors involved in the selection of source templates. Environmental aspects such as solvent and pH combined with physical aspects such as completeness, resolution and quality should also be considered (Martí-Renom et al., 2000). Comparative modeling produces theoretical structural models with root mean square (rms) errors of approximately 1Å (Sali and Blundell, 1993), where the previous considerations have been applied.

For each model produced, Modeller evaluates the structure based on the discrete optimized protein energy (DOPE) (Shen and Sali, 2006) to provide a score value. In principal, native structures have the lowest free energy state in natural conditions, but calculations of this type are computationally expensive (Shen and Sali, 2006). Shen et al. describe a variety of works involving statistical potentials, the simplicity and accuracy of these types of methods and the wide range of uses to substantiate DOPE. Scoring by DOPE via discrete optimization requires treating the reference state as a uniformly dense sphere based on the native structure size from which the statistical potential is determined (Shen and Sali, 2006). The pipeline uses DOPE scoring to select a quality set of models for processing.

Template proteins used to generate all gp120 models are: 1G9M, 1RZK, 2B4C, 2BF1, 2NY7, 3JWD, 3JWO, and 3LQA (Kwong et al., 2000; Huang et al., 2004, 2005; Chen et al., 2005; Zhou et al., 2007; Pancera et al., 2010; Diskin et al., 2010). Template proteins used to generate all bnAb models are: 3U4B, 4FQ1, 4FQ2, 4OD1, and 2NY7 (McLellan et al., 2011; Mouquet et al., 2012; Doria-Rose et al., 2014; Zhou et al., 2007). CD4 was generated from a single template protein, 1G9M (Kwong et al., 2000). All templates and models are stored in PDB format, an atomic coordinate storage method used for proteins (Berman, 2000; Berman et al., 2003).

Gromacs Minimization

Once theoretical models are generated by Modeller, the structures are energy minimized by Gromacs (Berendsen et al., 1995; Lindahl et al., 2001) to ensure that proteins are not manipulated under stress from atoms that are unnaturally close that may alter folding characteristics or even break covalent bonds in simulation. Gromacs (Berendsen et al., 1995; Lindahl et al., 2001) was chosen to perform this operation because of its wide acceptance, open source, and mature options set. Conjugate gradient algorithm is the selected integrator and the procedure is limited in the number of iterations to ensure completion. The assembly is prepared with Gromacs' pdb2gmx utility using the Amber99SB-ILDN force field (Lindorff-Larsen et al., 2010), ignore hydrogen, tip3p water, and add terminator options. Other force fields are not considered, since only structure energy minimization is desired.

Frodan

Frodan manipulates proteins geometrically by exploring angular and torsional limits of bonded atoms to maintain stereo-chemically correct states (Farrell et al., 2010). By supplying a template protein in a desired state, such as bound or unbound conformations, the source protein can be shifted to the target proteins conformation to the supported limits of the source protein. Figure 3.3 shows examples of a model structure shifted into the bound (left) and unbound (right) conformations. Frodan performs protein folding based on atom

overlap detection, and geometric limitations of covalent bonds, torsion points, hydrophobic points of contact, and hydrogen bonds. Frodan performs targeted protein folding simulation approximately one thousand times faster than typical molecular dynamics simulations. Accuracy of Frodan can be expressed by before and after energy minimization data, as shown in Figure 3.4. Post Modeller energy minimization is expressed by red in Figure 3.4, while post Frodan minimization is expressed by blue. These data indicate that Frodan generally maintains a lower energy state during manipulation of the structure.

Figure 3.3: A gp120 protein in the unbound (left) and bound (right) conformations. Orientation of the protein is identical between the two images allowing an assessment of the conformational changes performed by Frodan.

Target states for gp120, both bound and unbound, are represented by 1RZK in respective conformations. This gp120 structure is only available bound to a CD4 protein (1RZK) or antibody structure (2NY7). 2B1F is the only available putative unbound gp120 at the time of this writing, and is from the Simian Immunodeficiency Virus (SIV) gp120 core (Chen et al., 2005). By utilizing 2B1F as a target for 1RZK, Frodan is capable of manipulating from the bound to the unbound state. Targets for bnAbs are provided by 2NY7 and for CD4, targets are provided by 1G9M.

PDB2PQR

In order to calculate electrostatic charges in a pH solvent, the PDB file must be converted to PQR, which allows for a broader information base over PDB. During the conversion process, a source PDB is copied, the coordinate systems is inserted into a containing

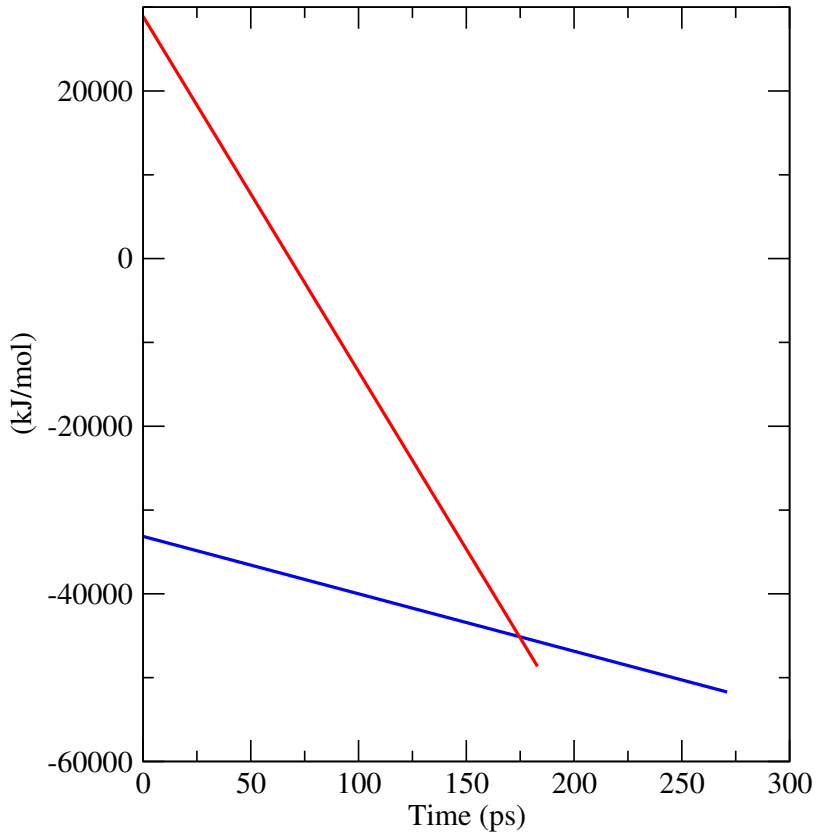

Figure 3.4: Graph expressing typical energy minimization data after Modeller (red) and Frodan (blue). These data indicate that Frodan generally maintains a lowered energy state during manipulation of the structure.

box in which a solvent of specific pH concentration is applied. Titrations of pH from 3.0 to 9.0 is performed in this manner to generate 61 new files using 0.1 increments. The conversion uses: the AMBER (Hornak et al., 2006) force field, for its wide acceptance, and PROPKA, one of the most commonly used predictors of pH state (Olsson et al., 2011). PDB2PQR is a required tool for the setup of Adaptive Poisson-Boltzmann Solver processing.

Gromacs PSIZE Utility

Each structure generated requires the determination of grid points, center of mass, fine and coarse mesh lengths, all provided by Gromacs' psize.py (Baker et al., 2001). Again, Gromacs is a common tool, with wide acceptance of use.

APBS Preparation

Each structure must be prepared for the Adaptive Poisson-Boltzmann Solver (APBS) (Baker et al., 2001; Jurrus et al., 2018). Preparation involves providing access to an APBS input file, with all required parameters determined and applied. APBS is a widely used, open source software with a mature code base that performs well and is very stable in parallel environments.

VMD Solvent Accessible Surface Area

This function executes a Visual Molecular Dynamics (VMD) (Humphrey et al., 1996) script that determines the solvent accessible surface (SAS) and surface area (SASA) using VMD's measure sasa function at 1.4 Å resolution, to return two files containing the required information for later calculations. VMD is a widely used, open source software utility that provides scripted operations which are used extensively in the pipeline and analysis processes.

APBS

This step represents the largest execution requirements in terms of CPU, memory, and disk. APBS generates data proportional to the number of atoms contained in the molecule(s), and the box volume containing the structures being processed. Multiply this by the number of entries being examined, and a very large data set appears in the target directory structure. The data returned is all floating point data of $x >= 1.00000$, in scientific notation for each interesting point within the grid (i.e. atoms). For solvents or empty space, values approach 0. This format has a higher tolerance for minor floating point errors introduced by lossy compression routines such as ZFP (Lindstrom, 2014). ZFP is incorporated into the pipeline using Cython methods (Virtanen et al., 2020; Behnel et al., 2011).

ZFP

The initial estimates of the total data to be produced during one study of 252 sequences was estimated at approximately 130TB. The largest producer of data is APBS, from which all charge data is stored in DX format, which is textual based, consisting of descriptive and numeric content. Basic methods of encapsulation (eg. GZIP) typically achieve 2:1 compression ratios and is a callable method in APBS, but, ratios of 2:1 are entirely inadequate for large scale analysis of structures in the manner presented by this example.

To overcome this limitation ZFP is employed to provide floating point data compression with a 0.1 acceptable loss setting. ZFP works exclusively with radix based exponential data by ingesting binary arrays and compressing them through signal processing methods (Lindstrom, 2014). A typical operation in our study produced compression ratios of 75:1, a maximum error of 0.016 kT/e, and a peak signal to noise ratio of 113:1. This compression method reduced overall data storage requirements down to an easily manageable size of approximately 7TB, that preserves the work for future analysis.

APBS output is stored in a temporary location, compressed and written to disk with a modified DX file that contains all descriptive information and the location of the ZFP compressed data file. Figure 3.5 shows a typical modified DX file used in this process.

3 Dimensional Convolution

To determine electrostatic charges on the surface area of any given protein or complex of proteins, a 3 dimensional convolution of charge data per surface point is performed. This method involves using the eight surrounding points in a lattice grid that encompass a central point to calculate the average charge at that central point.

An MPI Based Execution Environment

Message Passing Interface (MPI) is a standard used for the communications and interactions of multiple host systems, performing multi-process functions commonly used in

```
# Data from 1.4.2
#
# POTENTIAL (kT/e)
#
object 1 class gridpositions counts 193 193 257
origin 1.606700e+01 -3.164700e+01 -5.223200e+01
delta 5.171771e-01 0.000000e+00 0.000000e+00
delta 0.000000e+00 4.630000e-01 0.000000e+00
delta 0.000000e+00 0.000000e+00 4.776641e-01
object 2 class gridconnections counts 193 193 257
object 3 class array type double rank 0 items 9572993 data follows

---
prot-0033.dx.ZDX
---

attribute "dep" string "positions"
object "regular positions regular connections" class field
component "positions" value 1
component "connections" value 2
component "data" value 3
```

Figure 3.5: Typical output of a ZFP compressed DX file showing the location of compressed data and all required notations and parameters needed to reconstruct the original DX file format.

high performance computing environments. MPI facilitates the implementation of multiple program multiple data (MPMD) parallelism and is utilized in Python (Van Rossum and Drake Jr, 1995), which simplifies rapid modification of the program during prototyping. A producer-consumer modeled approach is utilized, which involves indexing the number of activities to be performed for any given step, sending the next index to each sub-process involved, and deriving the work unit from the index in real time.

JSON (ISO/IEC JTC 1/SC 22, 2017) was chosen for configuration of the pipeline because of its simplicity, ease of use. Its popularity has spread from JavaScript to other communities, e.g. Python, where it has largely replaced other markup languages such as

XML. Indicative of this popularity is that Python has a JSON module that is part of the standard distribution.

The pipeline driver was modeled using VCML2, which is a Docker (Rodrigues and Druschel, 2010) and Linux container based (Linux Containers, 2008) virtual cluster method developed in Python (Morton et al., 2016). VCML2 also uses JSON as a configuration method to define virtual clusters. The software is published and available at: https://zenodo.org/badge/latestdoi/270651434

The source code for the ESSC pipeline is published and available at: https://zenodo.org/badge/latestdoi/271174094

CHAPTER IV : METHODS

Background

Dynamic Electrophoretic Fingerprinting

Electrophoretic mobility (EM) is an experimental measure of surface charge used to characterize and separate micro-organisms (Mehrishi and Bauer, 2002; Richmond and Fisher, 1973). Researchers hypothesized the method could be applied across saline and pH ranges relevant to mucosal environments where transmission is common and results in systemic infection. The method was employed to study trimeric gp120/gp41 from clade B HIV-1 strain BX08 (Stieh et al., 2013) in the bound and unbound conformations by evaluating the difference between the two states. The results described surface charge variations across titrations indicating decreased gp120 surface charge in mucosal environments, complementing the positive charge of the CD4 receptor surface. This potentially could be the result of variations in gp120 protein structure and the interactions of surrounding solvent where blood plasma and mucous vary in pH and saline levels. This technique is used to validate the pipeline process in the methods that follow.

Bio-molecular Electrostatic Indexing

Bio-molecular ElectroStatic Indexing (BESI) is a machine learning method of classification, loosely based on Latent Semantic Indexing (LSI) (Deerwester et al., 1990). The goal is to determine if gp120 has distinguishable characteristics, in terms of electrostatics, that could be used to compare against a variant known to have caused an infection. This knowledge would provide the ability to predict variants of gp120 more likely to cross the transmission barrier. This method involves both principal component analysis and cosine similarity analysis. The data used to derive this method was produced from sequences provided by (Boeras et al., 2011).

Principal component analysis (PCA) is a common method of dimensional reduction (Pearson and Lipman, 1988; Hotelling, 1933) used in a wide range of fields. The method is useful for exploratory analysis and predictive modeling, where it provides low dimensional representations of high dimensional, multivariate data better suited for visualization.

The method utilizes cosine similarity analysis (CSA) as a means of comparing vectors on an R^x coordinate system, where the cosine of the angle between two vectors is an indicator of the similarity, where $cos(0) = 1$ indicates the vectors are on the same line. This holds true for $cos(180) = -1$, where the direction of the ray is reversed, the line on which the vector exist is still identical. The calculation is:

$$cos(\theta) = \frac{a \cdot b}{||a||_2||b||_2} = \frac{\sum_{i=1}^{x} a_i b_i}{\sqrt{\sum_{i=1}^{x} a_i^2}\sqrt{\sum_{i=1}^{x} b_i^2}}$$

where a and b are of the target and control sequences respectively.

BESI combines PCA and CSA in order to quantify the similarity between gp120 variants using three data sources: unbound, bound conformational ESP vectors and the difference between the two, as previously described. Initially, the first principal component for each sequence and conformation was compared to that of the control sequence, using CSA. Figures 4.1 - 4.3 express the results of each analysis. Based on the information present in EFP, Figure 4.1 displayed unexpected results. One can observe that the distribution of scores produces a large number of highly related sequences, which contrast the statistical results of (Boeras et al., 2011). The same description holds true in Figure 4.2, where bound data is presented. In contrast, Figure 4.3 demonstrates that the unbound conformation data provides a more discerning means of identifying sequences that are similar to the control.

To further refine the method, BESI uses the first two principal components to represent each data point, as two components are sufficient to describe at least 50% of the variance in the data, on average. Figure 4.4 displays the percentage of variance in each sequence that is explained by the first two principal components. Observe that the lowest returned score is 0.4933. Finally, BESI uses the absolute values of the CSA scores, as show in Figure 4.5

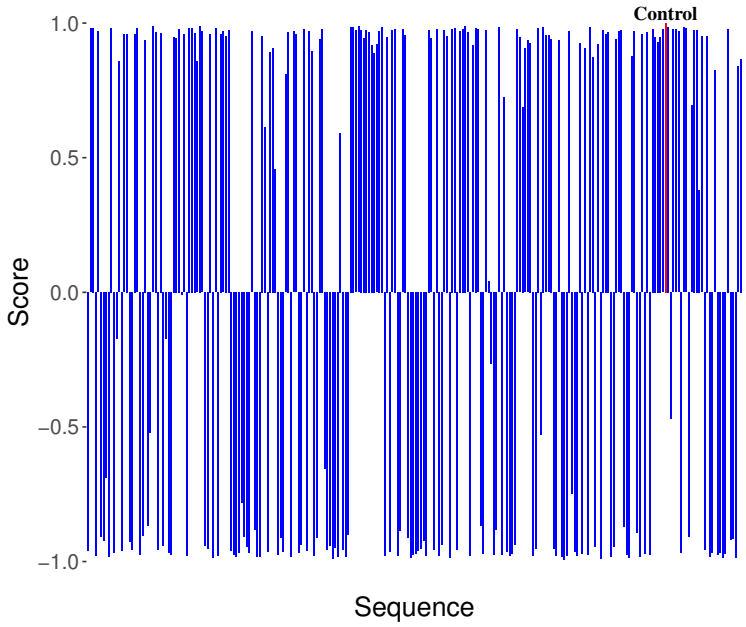

Figure 4.1: BESI scores for the electrostatic fingerprint data showing an unusually high number of similar sequences.

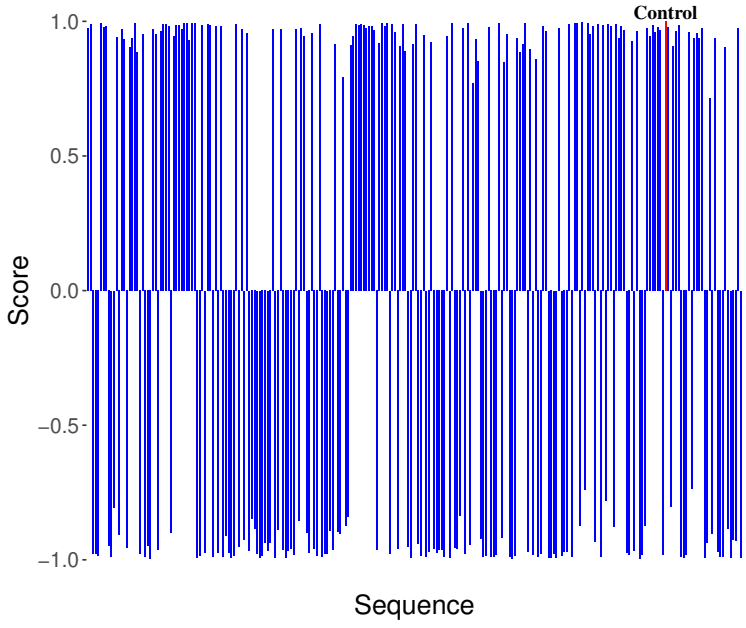

Figure 4.2: BESI scores of bound conformation data showing an unusually high number of similar sequences.

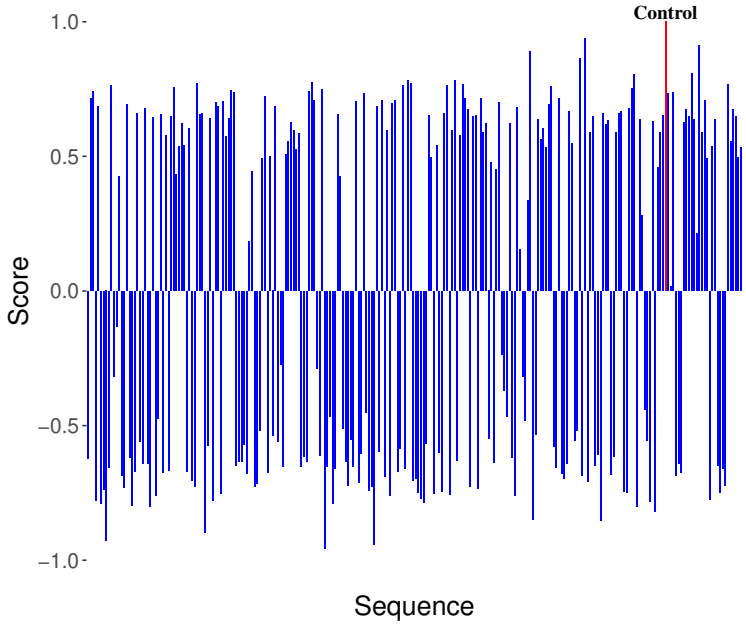

Figure 4.3: BESI scores of unbound conformational data showing a significant signal can be determined.

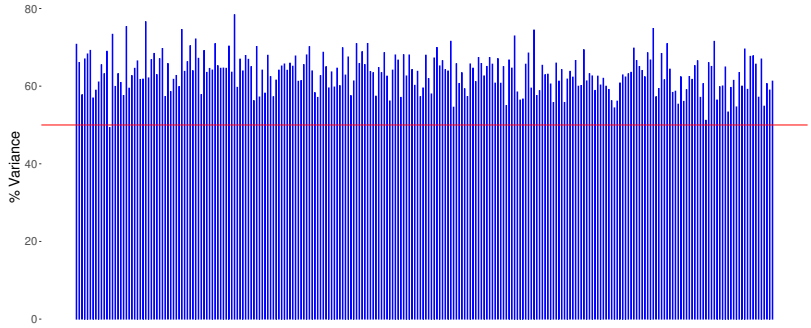

Figure 4.4: The minimum number of principal components to obtain an average minimum variance greater than 50% requires the use of the first two principal components. By this standard, the lowest value returned is 0.4933.

To verify that two principal components are a good fit, we reverse the PCA to reconstruct the bound, unbound, and EFP electrostatic data. Figures 4.6, 4.7, and 4.8 display original (left) and PCA reconstructed (right) data for bound, unbound, and EFP respectively.

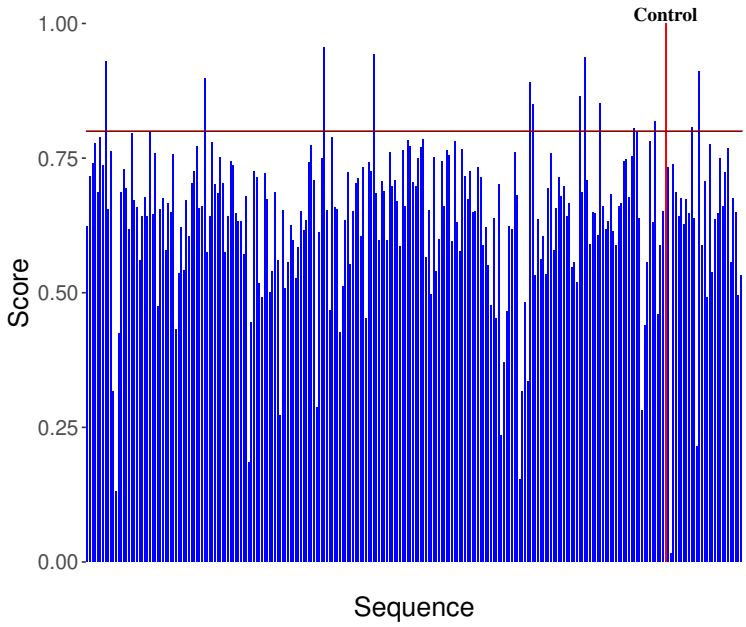

Figure 4.5: BESI scores taken as the absolute value. Horizontal line intersects the y-axis at 0.80 to distinguish predicted sequences that exhibit characteristics of the control.

Figure 4.6: Original (left) and PCA reconstruction (right) of typical bound electrostatic data show reconstructed data from PCA is valid.

Figure 4.7: Original (left) and PCA reconstruction (right) of typical unbound electrostatic data show reconstructed data from PCA is valid.

Figure 4.8: Original (left) and PCA reconstruction (right) of typical EFP shows reconstructed data from PCA is valid.

BESI computes the cosine similarity between the principal components of each variant sequence and the target sequence, and averages the values together to return a BESI score for each variant. To visualize BESI as a search space, Table 4.1 expresses the method as a list of sequence model data where PCA data is generated from the model data computed for each sequence, BESI searches by comparison of results to a control variant to return similar sequences. BESI has a tendency to select the most likely candidate variant within the top 3 scores returned and in no predictable scoring order. It is important to keep in mind that the variations taking place upon infection of a new cell are not predictable, BESI only predicts electrostatic characteristics that match those of the control variant used.

BESI scores can then be applied as a color gradient to leaves in a phylogenetic tree as a visual comparison method. Figure 4.9 represents a typical overlay of BESI to a phylogenetic tree, with donor and recipient classes of sequences being represented as two color gradients. Recipient scores are applied white to red, donor scores are light green to blue.

Phylogenetic trees were constructed as follows: Sequences were separated by subject, and aligned with MAFFT v7.273 using the L-INS-i strategy(Katoh and Standley, 2013). A maximum likelihood (ML) phylogenetic tree was constructed using the RAxML software, version 8.2.11 (Stamatakis, 2014) with the HIVW amino acid model of substitution (Nickle et al., 2007) and 100 bootstrap replicates. Trees were midpoint-rooted and rendered using APE version 5.0 (Paradis et al., 2004)..

Table 4.1: Visualization of the search space imposed by BESI. The process encompasses PCA of the sequence model set of surface charge data, CSA comparison of the first 2 principal components of the control and target data, which is loosely based on latent semantic indexing.

Unbound Data for Each Sequence						
Model/pH	3.0	3.1	3.2	3.3	. . .	9.0
Model_1	0.008658	-1.246752	0.441558	1.229436	. . .	-1.290042
Model_2	0.017316	1.25541	-0.017316	0.580086	. . .	-1.16883
. . .	. . .	. . .	. . .	. . .	. . .	. . .
Model_N	0.019243	1.142856	-1.55844	1.549782	. . .	1.090908

Seq_1 Seq_2 . . . Seq_N	PCA PCA . . . PCA	CSA of PC 1,2	**BESI**	BESI compares the first 2 principal components of each sequence against the first 2 principal components of the control.

Applying this method to several different studies give a unique perspective of the transmission bottleneck previously described.

Results

These results are a product of the sequence set of 252 gp120 proteins sourced from Boeras et al. (2011), Trask et al. (2002), Li et al. (2006a), Rong et al. (2009), Kawashima et al. (2009), and Carlson et al. (2014). As previously stated, these sequences provide a unique opportunity to investigate the transmission bottleneck related to HIV. The investigations involved couples where one subject was infected with HIV and the other was expected to contract the disease from their partner.

The sequence data is coded to include the source country, subject pair numbers, gender, and extraction characteristics as shown in Table 4.2. The source countries are Rwanda and Zambia, the couples are heterosexual. Details of the studies and results can be obtained from the referenced papers.

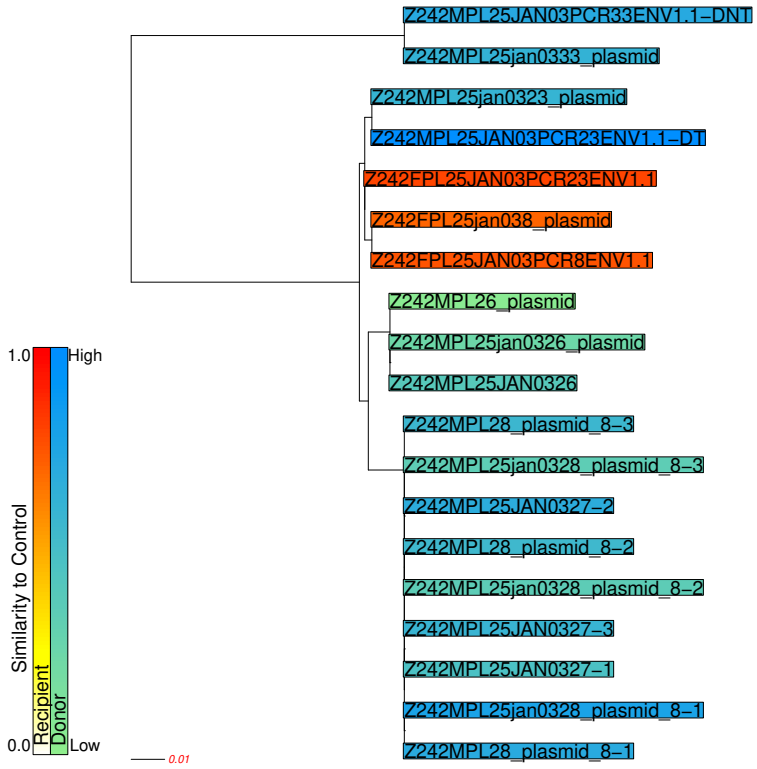

Figure 4.9: Typical phylogenetic tree with BESI scores overlaid as a color gradient on the leafs. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table 4.2: List of sequence donors. Subject indicates country of origin, couple identifier and gender respectively. D/R indicates the subjects status as the donor and communication recipient, respectively. Total is the number of variants provided.

Subject	**D/R**	**Total**	**Subject**	**D/R**	**Total**
R56F	R	4	R56M	D	13
Z153F	D	11	Z153M	R	10
Z185F	D	10	Z185M	R	10
Z201F	D	42	Z201M	R	14
Z205F	D	5	Z205M	R	7
Z216F	D	24	Z216M	R	1
Z221F	D	26	Z221M	R	10
Z238F	D	20	Z238M	R	2
Z242F	R	3	Z242M	D	16
Z292F	D	18	Z292M	R	6

Table 4.3: Couple Z242 details sequence name, HIV clade, donor/recipient classification and BESI scores. This set contains the control gp120 variant with a score of 1.

Sequence	Clade	Donor/Recipient	Score
Z242FPL25jan038_plasmid	C	R	0.7165
Z242FPL25JAN03PCR23ENV1.1	C	R	0.8074
Z242FPL25JAN03PCR8ENV1.1	C	R	0.7678
Z242MPL25jan0323_plasmid	C	D	0.6160
Z242MPL25jan0326_plasmid	C	D	0.2545
Z242MPL25JAN0326	C	D	0.4055
Z242MPL25JAN0327-1	C	D	0.4842
Z242MPL25JAN0327-2	C	D	0.6960
Z242MPL25JAN0327-3	C	D	0.6077
Z242MPL25jan0328_plasmid_8-1	C	D	0.7817
Z242MPL25jan0328_plasmid_8-2	C	D	0.3643
Z242MPL25jan0328_plasmid_8-3	C	D	0.3562
Z242MPL25jan0333_plasmid	C	D	0.6345
Z242MPL25JAN03PCR23ENV1.1-DT	C	D	1
Z242MPL25JAN03PCR33ENV1.1-DNT	C	D	0.7255
Z242MPL26_plasmid	C	D	0.0567
Z242MPL28_plasmid_8-1	C	D	0.7180
Z242MPL28_plasmid_8-2	C	D	0.5666
Z242MPL28_plasmid_8-3	C	D	0.5867

From this large pool of HIV sequences, Boeras et al. (2011) provides a set of predictions regarding two gp120 variants, a Donor Transmitted (DT) Z242MPL25JAN03PCR23ENV-1.1-DT and a Donor Non Transmitted (DNT) Z242MPL25JAN03PCR33ENV1.1-DNT gp120 (Boeras et al., 2011). The DT variant is used as the control for BESI in this and all other studies involving BESI.

Figure 4.9 represents BESI versus phylogenetic tree containing the two predicted variants previously described. The tree classifies the DT variant as a child of the DNT variant from the donor and closely relates recipient variant Z242FPL25JAN03PCR23ENV1.1 to DT; in fact, the two variants differ by a single residue (Boeras et al., 2011). Z242FPL-25JAN03PCR23ENV1.1 has a high BESI score, meaning that it has similar electrostatic characteristics as the DT variant indicating it has the potential to cross the transmission bottleneck. BESI scores for couple Z242 can be viewed in Table 4.3.

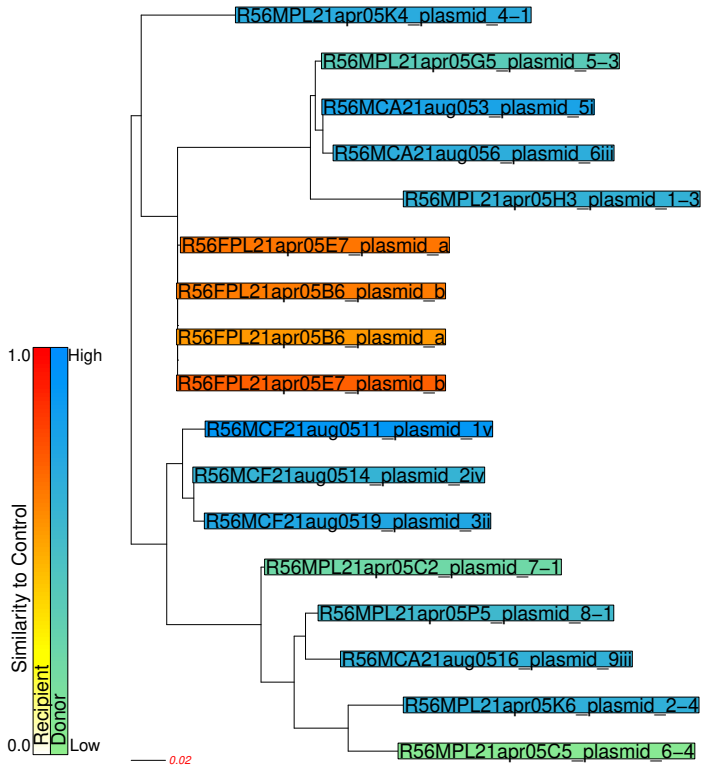

Figure 4.10: BESI versus phylogenetic tree for couple R56. This tree requires the scores in Table 4.4 and the understanding that BESI typically includes the candidate gp120 in the top 3 scores. This information indicates R56MCA21aug053_plasmid_5i is the transmitted founder for this couple and is the second highest score returned for this donor.

BESI scores returned for couple R56 requires deeper interpretation. As previously stated, BESI has a tendency to pick the top 3 contenders in no particular scoring order. R56 is a good example of this behavior, where the recipients are descendants of R56MPL21-apr05K4_plasmid_4-1 along with the second highest donor scored sequence R56MCA21-aug053_plasmid_5i. This represents a good example of the potential predictive power of BESI.

BESI scores returned for couple Z153 follow the same interpretation requirements as couple R56. The phylogenetic tree indicates that recipient variants of gp120 are descendants of Z153FPL13MAR02ENV1.1. The candidate picked by BESI as the transmitted variant is Z153FPL13MAR02ENV3.1, with the second highest donor BESI score and is

Table 4.4: Couple R56 details sequence name, HIV clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
R56FPL21apr05B6_plasmid_a	A1	R	0.5639
R56FPL21apr05B6_plasmid_b	A1	R	0.6491
R56FPL21apr05E7_plasmid_a	A1	R	0.6681
R56FPL21apr05E7_plasmid_b	A1	R	0.7265
R56MCA21aug0516_plasmid_9iii	A1	D	0.6510
R56MCA21aug053_plasmid_5i	A1	D	0.7849
R56MCA21aug056_plasmid_6iii	A1	D	0.6874
R56MCF21aug0511_plasmid_1v	A1	D	0.9149
R56MCF21aug0514_plasmid_2iv	A1	D	0.6241
R56MCF21aug0519_plasmid_3ii	A1	D	0.7520
R56MPL21apr05C2_plasmid_7-1	A1	D	0.2402
R56MPL21apr05C5_plasmid_6-4	A1	D	0.0690
R56MPL21apr05G5_plasmid_5-3	A1	D	0.3853
R56MPL21apr05H3_plasmid_1-3	A1	D	0.6418
R56MPL21apr05K4_plasmid_4-1	A1	D	0.7099
R56MPL21apr05K6_plasmid_2-4	A1	D	0.6610
R56MPL21apr05P5_plasmid_8-1	A1	D	0.5628

also under the same sub-tree as the recipients. The complete list of scores for couple Z153 is in Table 4.5.

Couple Z185 presents a selection of sequences, where ten variations were selected out of twenty-three (counts obtained from(LANL, 2020)), see Figure 4.12. BESI selects a candidate sequence Z185MPB17AUG02ENV1.2 with a score of 0.7578 as the highest donor score, see Table 4.6. According to the phylogenetic tree, sequence Z185MPB17AUG02-ENVC18 is the potential transmitted variant, but the score is too low to be considered a positive match. Hence, in this example BESI was unable to infer the transmitted variant. Note, however, it is possible that the gp120 group selected does not contain the actual donor variation or the process falls short under certain circumstances yet to be determined that resulted in an indeterminate.

Figure 4.13 provides a potential view of BESI performing poorly in that a strong score of 0.844 for variant Z221FPL7MAR03ENV3.3 exists in the same clade as the recipient

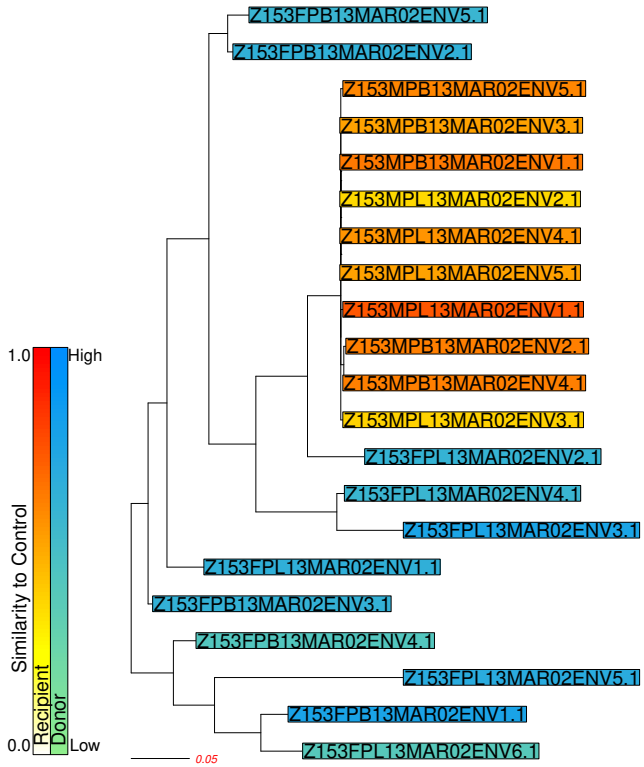

Figure 4.11: Couple Z153 also follows through with the second highest donor BESI score (see Table 4.5), Z153FPL13MAR02ENV3.1, and being in the correct sub-tree as a candidate to cross the transmission barrier.

variations, but is below the recipient branch. For this donor (LANL, 2020) reports 75 sequence variations were extracted. Again this presents the potential that the transmitted variant is not present.

The remaining five BESI versus phylogenetic trees can be observed in Appendix A. All remaining comparisons display similar results to those presented above.

Discussion

BESI is a machine learning method that can be used to predict which gp120 variants have a high potential to cross the HIV transmission bottleneck. BESI has provided favorable results where 70% of the couples evaluated indicate a valid selection of the potential transmitted variant. BESI should be evaluated against traditional laboratory methods to

Table 4.5: Couple Z153 details sequence name, HIV clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z153FPB13MAR02ENV1.1	C	D	0.7805
Z153FPB13MAR02ENV2.1	C	D	0.6534
Z153FPB13MAR02ENV3.1	C	D	0.6418
Z153FPB13MAR02ENV4.1	C	D	0.4343
Z153FPB13MAR02ENV5.1	C	D	0.5929
Z153FPL13MAR02ENV1.1	C	D	0.6646
Z153FPL13MAR02ENV2.1	C	D	0.6235
Z153FPL13MAR02ENV3.1	C	D	0.7729
Z153FPL13MAR02ENV4.1	C	D	0.5975
Z153FPL13MAR02ENV5.1	C	D	0.7057
Z153FPL13MAR02ENV6.1	C	D	0.4000
Z153MPB13MAR02ENV1.1	C	R	0.6535
Z153MPB13MAR02ENV2.1	C	R	0.6395
Z153MPB13MAR02ENV3.1	C	R	0.5342
Z153MPB13MAR02ENV4.1	C	R	0.6366
Z153MPB13MAR02ENV5.1	C	R	0.6169
Z153MPL13MAR02ENV1.1	C	R	0.7532
Z153MPL13MAR02ENV2.1	C	R	0.3788
Z153MPL13MAR02ENV3.1	C	R	0.3973
Z153MPL13MAR02ENV4.1	C	R	0.5760
Z153MPL13MAR02ENV5.1	C	R	0.5313

determine accuracy. Efforts of this nature would substantiate the method and bring forward emerging technologies that help to understand biological functions.

Additionally, BESI should be used with complete experimentally derived structures as control variants, which would limit structural fluctuations based on generation of homology models built from templates. Furthermore, sensitivity of the method to proteins of various clade may also present the need to make clade specific comparisons.

Electrostatic Variance Masking

Selection of residues that show surface charge response to pH shifts involves calculating the electrostatic potential variance of each residue across all sequence aligned variants, vertically. All proteins are aligned to HXB2CG, as described in (Korber-Irrgang et al.,

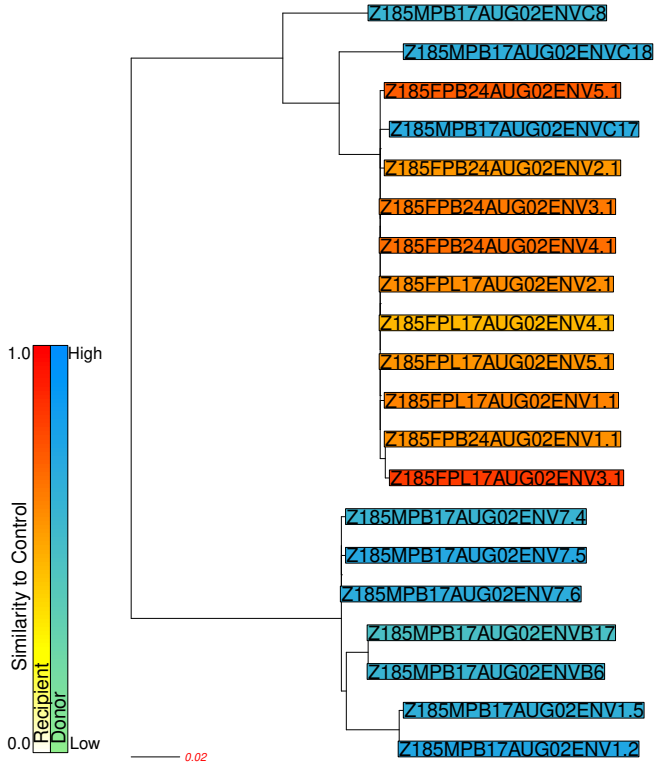

Figure 4.12: BESI versus phylogenetic tree for couple Z185 show BESI failing to correctly identify a plausible donor variant that matches the evolutionary tree with the third highest score is the only selection under the proper clade, but in a sub-tree below the plausible transmission point of Z185MPB17AUG02ENVC18.

1998), using MAFFT (Katoh and Standley, 2013) with einsi and a gap penalty of 2.0. This provides a common numbering scheme for residues and allows describing those residues that EVM selects in a concise manner.

This process to derive residue charge variance is graphically expresses in Figure 4.14, and described in detail as follows: Each sequence model is analyzed for each pH value at the residue level to create a 3 dimensional array of X residues by Y models by Z pH values. The arrays are stacked to align residues where, model 1 at pH 3.0, residue 1 is aligned with model 2 at pH 3.0, residue 1 and so on. The column median on the Y axis (models) at pH 3.0 is taken. This is repeated for each residue for pH 3.0 to 9.0 in 0.1 increments until all residues for this sequence have been processed to reduce the array to 2 dimensions.

Table 4.6: Couple Z185 details sequence name, HIV clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z185FPB24AUG02ENV1.1	C	R	0.5886
Z185FPB24AUG02ENV2.1	C	R	0.5685
Z185FPB24AUG02ENV3.1	C	R	0.6572
Z185FPB24AUG02ENV4.1	C	R	0.6877
Z185FPB24AUG02ENV5.1	C	R	0.7378
Z185FPL17AUG02ENV1.1	C	R	0.6261
Z185FPL17AUG02ENV2.1	C	R	0.5967
Z185FPL17AUG02ENV3.1	C	R	0.8361
Z185FPL17AUG02ENV4.1	C	R	0.4797
Z185FPL17AUG02ENV5.1	C	R	0.5713
Z185MPB17AUG02ENV1.2	C	D	0.7578
Z185MPB17AUG02ENV1.5	C	D	0.6644
Z185MPB17AUG02ENV7.4	C	D	0.6073
Z185MPB17AUG02ENV7.5	C	D	0.7367
Z185MPB17AUG02ENV7.6	C	D	0.6901
Z185MPB17AUG02ENVB17	C	D	0.4992
Z185MPB17AUG02ENVB6	C	D	0.6004
Z185MPB17AUG02ENVC17	C	D	0.7019
Z185MPB17AUG02ENVC18	C	D	0.6798
Z185MPB17AUG02ENVC8	C	D	0.6053

The sequence alignment with HXB2CG is then referenced to expand the array by the X dimension (residue) to match the sequence alignment, so that where gaps in the sequence alignment exist, a charge value of zero is assigned to the gap position.

This is then repeated for the next sequence and so on, until a 3 dimensional array is created consisting of X residues by Y sequences by Z pH. The 3 dimensional array is then reduced to 2 dimensions again by the means of the Y axis across all residues (X axis) so that each residue position has values across the pH range. The variance across the pH range for each residue position is then determined and stored as a single dimension array and graphed as in Figure 4.15. This method allows effective filtering of residues with small variations in mean surface charge across the pH shift.

Table 4.7: Couple Z221 details sequence name, HIV clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z221FPB7MAR03ENV10.3	C	D	0.5970
Z221FPB7MAR03ENV11.3	C	D	0.4638
Z221FPB7MAR03ENV6.4	C	D	0.4345
Z221FPB7MAR03ENV9.1	C	D	0.5074
Z221FPL08mar0335	C	D	0.4239
Z221FPL08mar0344	C	D	0.5674
Z221FPL08mar0348	C	D	0.3801
Z221FPL08mar0351	C	D	0.6545
Z221FPL08mar0355	C	D	0.2047
Z221FPL08mar0371	C	D	0.3647
Z221FPL08mar0380	C	D	0.4614
Z221FPL35_plasmid_7-1	C	D	0.6257
Z221FPL44_plasmid_4-1	C	D	0.5208
Z221FPL48_plasmid_5-1	C	D	0.7250
Z221FPL51_plasmid_2-2	C	D	0.6522
Z221FPL55_plasmid_6-2	C	D	0.0882
Z221FPL71_plasmid_9-1	C	D	0.3137
Z221FPL7MAR03ENV1.2	C	D	0.4447
Z221FPL7MAR03ENV10.4	C	D	0.1714
Z221FPL7MAR03ENV2.3	C	D	0.8690
Z221FPL7MAR03ENV3.3	C	D	0.8440
Z221FPL80_plasmid_8-3	C	D	0.5000
Z221FSW08mar0314H16iii	C	D	0.6183
Z221FSW08mar0314H16iv	C	D	0.5307
Z221FSW14H16_plasmid_6iii	C	D	0.5581
Z221FSW14H16iv_plasmid_6iv	C	D	0.5209
Z221MPB7MAR03ENV4.1	C	R	0.6787
Z221MPB7MAR03ENV5.4	C	R	0.7317
Z221MPB7MAR03ENV6.4	C	R	0.4948
Z221MPL08mar0375a	C	R	0.6477
Z221MPL08mar0375f	C	R	0.6811
Z221MPL75_plasmid_a	C	R	0.6570
Z221MPL75_plasmid_f	C	R	0.6478
Z221MPL7MAR03ENV2.1	C	R	0.6140
Z221MPL7MAR03ENV4.2	C	R	0.5595
Z221MPL7MAR03ENV6.4	C	R	0.4602

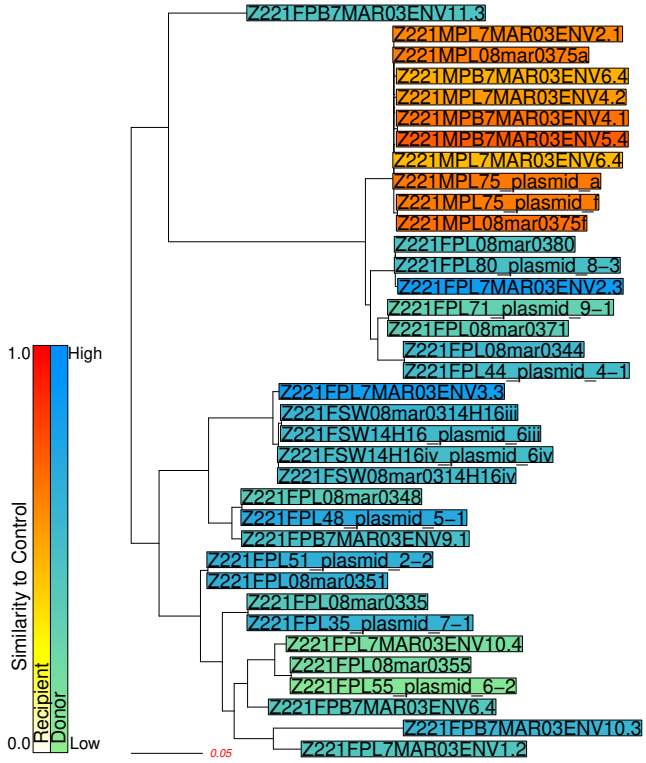

Figure 4.13: BESI versus phylogenetic tree for couple Z221. BESI scores in this set indicate BESI performing poorly or potentially that the transmitting variant is not included in the list of studied variants out of the 75 extracted from the patient (LANL, 2020). Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

The sequence alignment is then referenced again to provide the selection criteria for high variance residues. Alignment gaps allows the determination of a cutoff value for variance where gaps in some determined sequence can easily be detected. To determine a starting value for selection, the ceiling of one-half the standard deviation is calculated for the variance data and used as the cutoff value. Assuming a gap is selected, the cutoff value is incremented by one until a uniform selection across all sequences can be determined.

The selected residues of the gp120 protein are then applied to a VMD representation (Humphrey et al., 1996) to display the residues with high variance across the pH range. For each of the sequences, the first model of the unbound conformation is loaded into

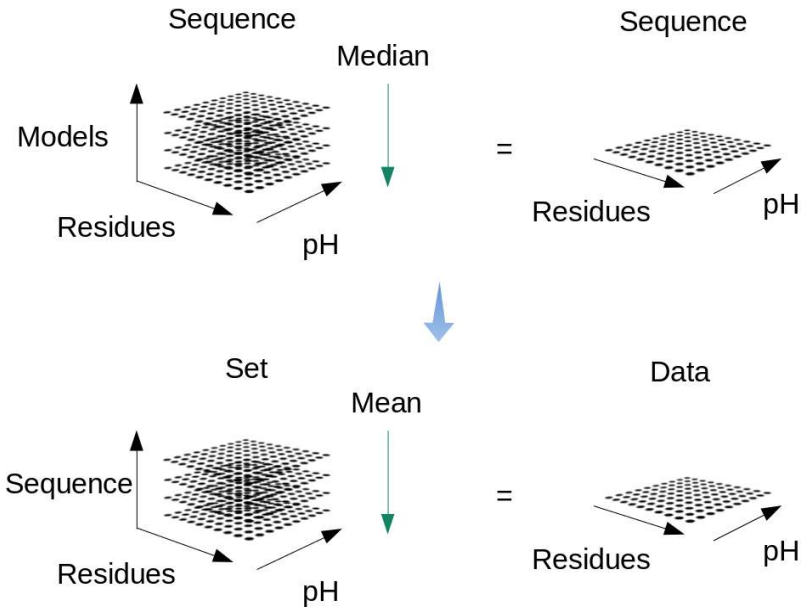

Figure 4.14: Visualization of the process to extract variance data from residue electrostatics. Models of residue data are reduced to 2 dimensions by taking the median of the set across models/residue to eliminate the effects of outliers on the data. Once each seqence/-model set is processed, the mean across the sequence set is taken to produce the residue data for which the variance is extracted for each residue across the pH range.

VMD and an additional representation of the protein created and set to only display EVM selected residues. The primary representation is presented as cartoon, colored by secondary structure. For the residue selection, a red transparent surface is used to allow viewing the alpha helix's, beta sheets and other ligands with high pH variance.

This method of imaging residue structures participating in the mechanistic functions of the binding process is called Electrostatic Variance Masking (EVM). Figure 4.16 is an annotated representation of EVM applied as an overlay in red to a gp120 structure to display residues of high variance. The $\alpha 2$ helix is marked to orient the view of the protein so the CD4bs is facing outward. All other representations of EVM in this manner will be oriented with the $\alpha 2$ helix left of the CD4 binding interface.

EVM also allows a representation of conserved residues to be presented via Weblogo (MCrooks et al., 2004), as shown in Figure 4.17. Letters represent the single character

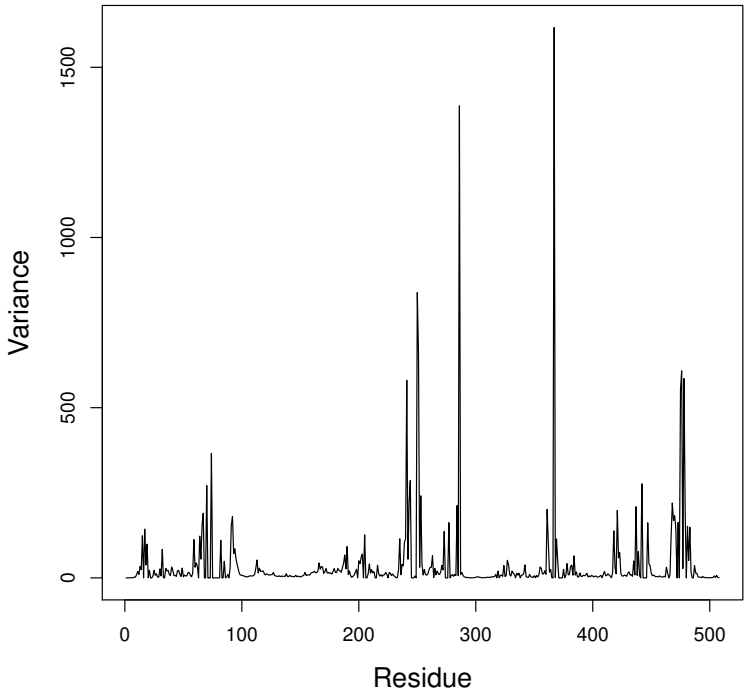

Figure 4.15: Example of EVM results for a typical set of gp120 proteins.

residue identifier, common residues among sequences are shown with taller lettering, and stacking indicates differences among sequences.

Finally, the alignment of each sequence to HXB2 allows examination of selected residues against the map of HXB2 as provided in Korber-Irrgang et al. (1998). This allows for a concise description of residues selected by EVM.

Results

The first set of results comes from a sequence set of 24 gp120 pairs consisting of one transmitted founder (TF) and one chronic control (CC) structure from clade B, and C, with 18 and 6 pairs of TF and CC sequences respectively. An additional TF sequence from clade B is also included in the study. A complete list of accession numbers can be found in Howton (2017). B clade gp120 were acquired from Keele et al. (2008), Bar

Figure 4.16: Typical EVM overlay to visualize the selected high variance residues. The image is annotated showing the $\alpha 2$ helix to orientate the CD4 binding site as facing forward.

Figure 4.17: Typical Weblogo representation of EVM select residues. Letters represent the single character residue identifier, commonality is indicated by taller lettering in the graph indicating to indicate the level of conservation among sequences and stacking indicates differences between sequences.

et al. (2012), Salazar-Gonzalez et al. (2009), Dacheux et al. (2004), Turnbull et al. (2009), Bunnik et al. (2008), Wei et al. (2003), and Li et al. (2006b). C clade gp120 were obtained from Kothe et al. (2006), Abrahams et al. (2009), and Liu et al. (2012). This study broke the sequence set into groupings to compare against clade and TF versus CC sequences as designated by (Parrish et al., 2013). Sequence Z242MPL25JAN03PCR23ENV1.1-DT is the control variant from (Morton et al., 2017). Table 4.8 displays the complete list of sequences including clade and sub-class designations.

Table 4.8: Complete list of sequences showing clade and sub-class information from (Morton et al., 2018)

Sequence	Clade	Sub-Class	Sequence	Clade	Sub-Class
03_CH40TF	B	TF	1997.133-L-10	B	CC
46_CH40M6	B	CC	1997.159-L-1	B	CC
47_CH58TF	B	TF	1999.153-L-7	B	CC
48_CH58M6	B	CC	2000.309-L-7	B	CC
49_CH77TF	B	TF	2004.MM42d22_GN1	B	TF
50_CH77M6	B	CC	2005.MM42d324_GN1	B	CC
51_CH470TF	B	TF	1985.H2_5_12E3	B	TF
52_CH470M6	B	CC	1985.H5_4	B	TF
53_CH569TF	C	TF	1986.H1_7_2D5	B	TF
54_CH569M6	C	CC	1986.H4_007_1C11	B	TF
55_CH42TF	C	TF	1987.H3_12_7D5	B	TF
56_CH42M6	C	CC	1995.H2_114_8F6	B	CC
57_CH236TF	C	TF	1996.H1_62_1A8	B	CC
58_CH236M6	C	CC	1996.H5_75_7G12	B	CC
59_CH850TF	C	TF	1997.H3_110_8G7	B	CC
60_CH850M6	C	CC	1998.H4_146_2H10	B	CC
61_CH264TF	C	TF	BORI556_49	B	CC
62_CH264M6	C	CC	HOBRd16_20	B	TF
63_CH164M6	C	CC	SUMA736_59	B	CC
64_CH164TF	C	TF	1990.BORId9_3F12	B	TF
3w.21dps	B	TF	1990.WEAUd15_B2	B	TF
1992.133-7	B	TF	1991.HOBR0961_A21	B	CC
1993.153-10	B	TF	1991.SUMAd4_A32	B	TF
1993.159-4	B	TF	1993.WEAU1166_39	B	CC
1994.309-2	B	TF	Z242MPL25JAN03PCR23ENV1.1-DT	C	CC

For all studies involving the pipeline, a sample EFP is taken to ensure the pipeline has processed data as expected. Figure 4.18 shows typical results for bound, unbound and

difference data to confirm the pipeline has processed as expected. These results open the door to continue the analysis of these structures using EVM.

EVM selects residues that are predicted to be pH sensitive and therefore have variations in charge across the physiological range of pH. For the referenced sequences, Figure 4.19 displays a map of the variance in the average ESP value of each residue. The selection process takes the minimum variance required to obtain a consistent set of residues as described in the methods section. Figure 4.20 displays a screeplot of the variance data and the selected cutoff value, shown as a red horizontal line. Table 4.9 shows the statistical data returned from EVM for the selected residues. The standard deviation is for the variance data set. The cutoff value is the minimum variance value that excludes all sequence alignment gaps and the % of variance selected is the percentage of the variance in the data set.

Table 4.9: Statistics for selection of high variance residues. The standard deviation is for the variance data set. The cutoff value is the minimum variance value that excludes all sequence alignment gaps and the % of variance selected is the percentage of the variance in the data set.

Standard Deviation	*101.0*
1/2 Standard Deviation	*50.5*
Number of Selected Residues	*64.0*
Variance cutoff selected	*51.0*
% of variance selected	*75.5%*
% of residues selected	*11.3%*

The EVM selected residues across the set of sequences contains a conserved set of residues, as shown in the Weblogo graph of Figure 4.21. This study provided additional information involving HIV clade (clades B and C) and sub-class based on Keele et al. (2008) that identify transmitted founder (TF) and chronic control (CC) sub-classes *in vivo*. This information allows those divisions to be presented in separate graphs.

Figure 4.22 breaks the selection across clade to reveal an alternative perspective of residue conservation. One can observe that clade C, which has only six sequences in the

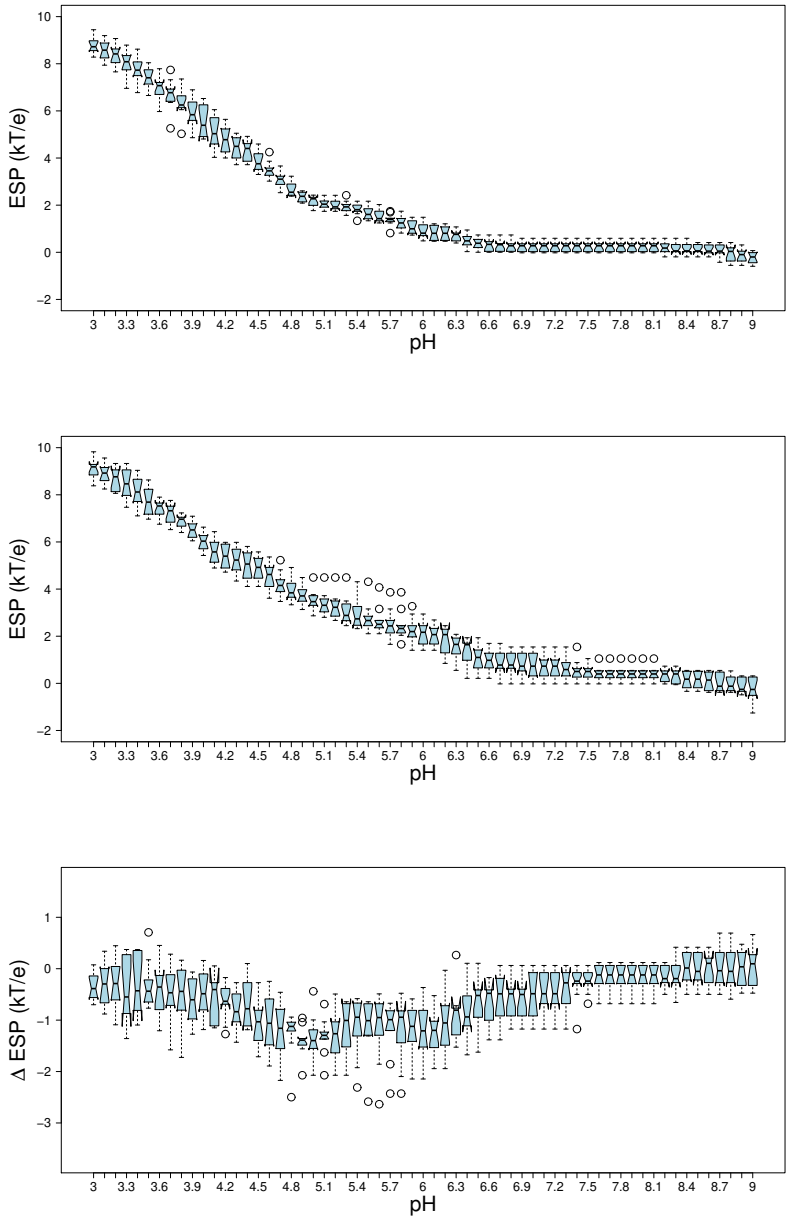

Figure 4.18: Electrostatics data for sequence 03_CH40TF displays a normal descent of charge from low to high pH for bound (top) and unbound (middle) conformations. Bound less unbound charge data (bottom) displays the signature EFP typical of this protein structure that confirms the pipeline has processed accordingly.

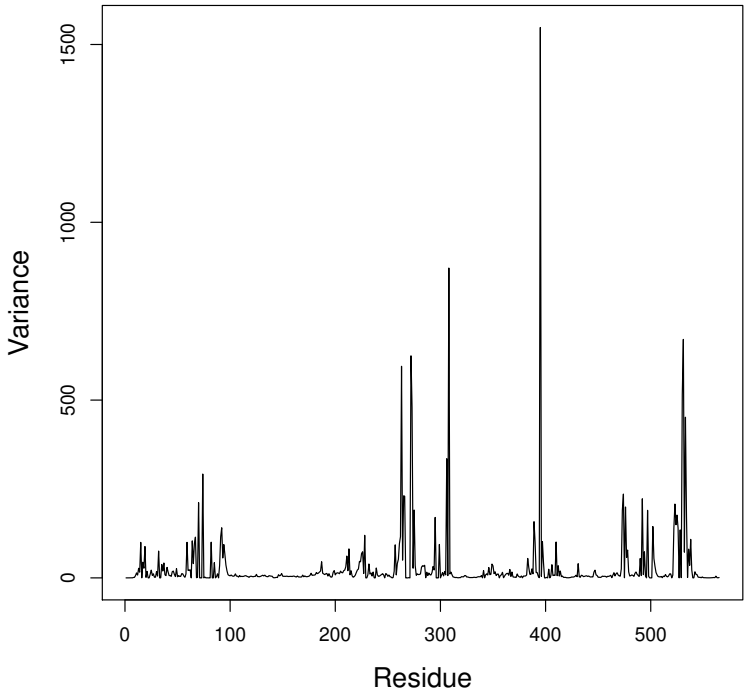

Figure 4.19: Variance map of all sequences based on the method described presents a clear signal.

set, has a wider variation of conserved residues than clade B, which has eighteen sequences in the set.

Figure 4.23 breaks the selection across sub-classes to show an additional view of residue conservation. This separation is nearly indistinguishable, which leads us to predict that either the measure of sub-class delineation is incorrect or else no differences are developed over time that distinguishes the two subclasses in terms of these selected residues.

EVM also produces actual residue numbering lists (not aligned to HXB2). The following are selection lists for two typical sequences:

- 56_CH42M6: length - 64

  14 18 31 58 63 65 66 69 73 81 90 91 92 93 94 170 172 184 185 187 216 219 220 221 222 224 225 231 232 234 254 258 265 267 333 339 340 345 347 360 394 395

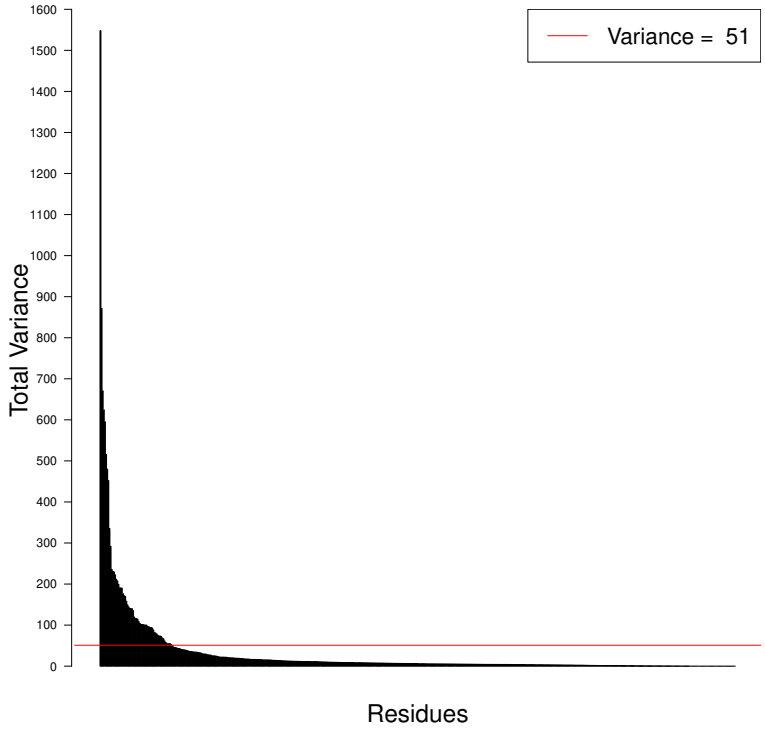

Figure 4.20: Screeplot of the variance data with the cutoff value shown as a red horizontal line.

Figure 4.21: Weblogo representation of the EVM selected residues for sequences in Morton et al. (2018). The graph displays a high level of predicted conservation among the set.

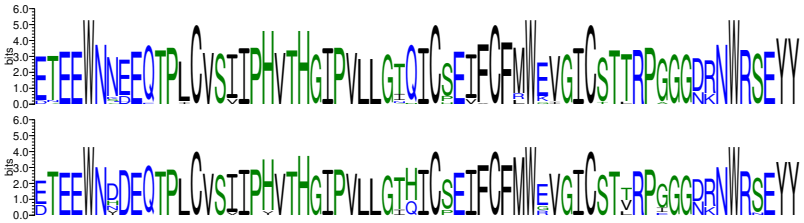

Figure 4.22: Weblogo representations, separating sequences across clade B (top) and C (bottom). clade C, having only 6 sequences in the set, shows a wider variation of selected residues versus the 18 sequences of clade B.

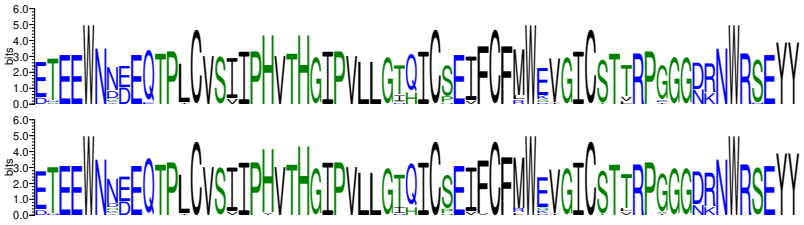

Figure 4.23: Weblogo representations of sequences separated across subclasses. Subclass CC (top) and TF (bottom) are nearly indistinguishable, predicting that either the measure of subclass delineation is incorrect or no differences are developed over time that distinguishes the two in terms of the variability in these selected residues.

397 398 399 411 413 415 418 423 424 440 441 442 443 444 446 448 449 450 451 452 454 456

- 1996.H1_62_1A8: length - 64

  14 18 31 58 63 65 66 69 73 81 90 91 92 93 94 176 178 190 191 193 222 225 226 227 228 230 231 237 238 240 260 264 271 273 339 345 346 351 353 366 414 415 417 418 419 431 433 435 438 443 444 461 462 463 464 465 467 469 470 471 472 473 475 477

Applying the imaging method previously described we produce Figures 4.24 and 4.25. Note the similarities across gp120 structures. In particular we see that residues of the CD4bs are highly conserved between the two variants.

Finally, this method produces the following HXB2 alignments using the previously described methods to express the residues selected in terms of HXB2 sequence alignment:

- 47 51 64 91 96 98 99 102 106 114 123 124 125 126 127 199 201 213 214 216 245 248 249 250 251 253 254 260 261 263 283 287 294 296 364 370 371 376 378 391 426 427 429 430 431 443 445 447 450 455 456 470 471 472 473 474 476 478 479 480 481 482 484 486

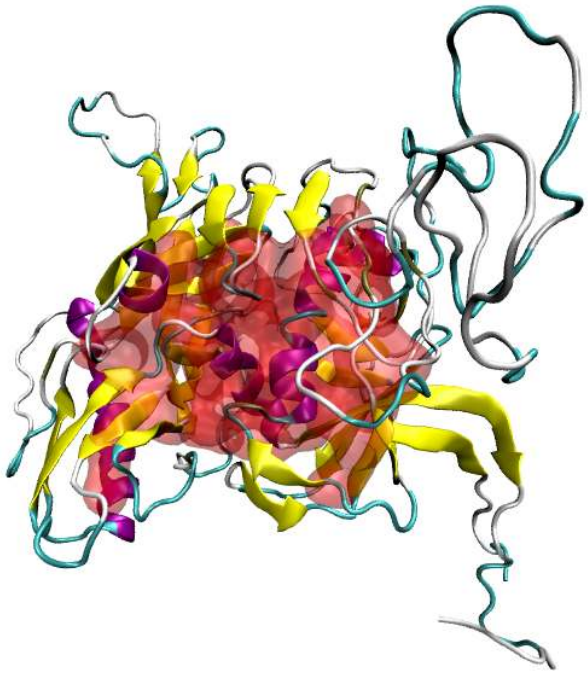

Figure 4.24: EVM imagery displaying the selected residues for sequence 56_CH42M6 in red.

For all 48 structures in this simulation, 41 presented identical residue selections, while the remaining seven structures varied by a single identical selection. The alternate list of selected residues with the difference in red bold-faced font are:

- 47 51 64 91 96 98 99 102 106 114 123 124 125 126 127 199 201 213 214 216 245 248 249 250 251 253 254 260 261 263 283 287 294 296 364 370 371 376 378 **396** 426 427 429 430 431 443 445 447 450 455 456 470 471 472 473 474 476 478 479 480 481 482 484 486

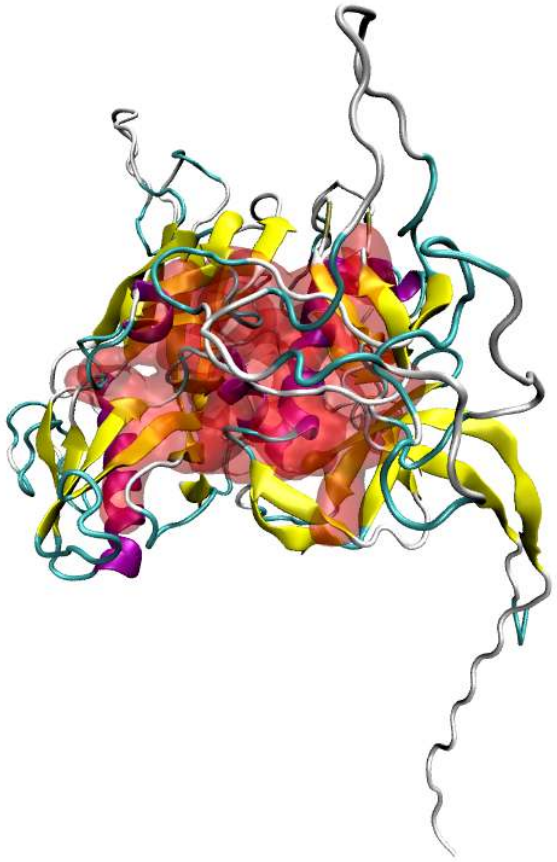

Figure 4.25: EVM imagery displaying the selected residues for sequence 1996.H1_62_1A8 in red.

The seven structures with the alternate selection were evenly distributed, to the extent possible, across TF/CC classes. Five of these variants were of clade B, the dominant subspecies of this study.

Most notably, EVM selected amino-acids 124-127, 283, 364, 370, 371, 426-431, 455, 456, 470-474, 476 which are CD4 contact residues. Other pertinent selections are as follows: Residues 64, and 91 are adjacent to interface contacts with gp41. Residue 123 is a co-receptor binding site outside of V3. Residues 199, 201, 251 are co-receptor sites specific R5/X4. Residues 261 and 263 are adjacent to glycosite 262. Residue 294 is adjacent to

Table 4.10: Sequence clade sources from (Morton et al., 2017)

Sequence	Clade
R56MCF21aug0511_plasmid_1v	A1
R56MPL21apr05C5_plasmid_6-4	A1
Z153FPB13MAR02ENV1.1	C
Z153FPL13MAR02ENV6.1	C
Z185MPB17AUG02ENVB17	C
Z185MPB17AUG02ENV1.2	C
Z201FPL7FEB03ENV2.1	C
Z201FCF07feb03DNA13C18	C
Z205MPB27MAR03ENV9.1	C
Z205MPB27MAR03ENV6.1	C
Z216FPL17jan0485f	C
Z216FPB98_plasmid_e	C
Z221FPL55_plasmid_6-2	C
Z221FPL7MAR03ENV2.3	C
Z238FSW29oct0215A6v	C
Z238FCF29oct0215A39	C
Z242MPL25JAN03PCR23ENV1.1-DT	C
Z242MPL26_plasmid	C
Z292FCF24may0512E26_plasmid_10iv	A1
Z292FCF24may0512D18_plasmid_4i	A1

glycosite 295. Residue 296 is the start of the V3 loop. Residue 391 is adjacent to glycosite 392. Residue 396 is at the V4 hyper-variable hot spot. Residue 447 is adjacent to glycosite 448. The previous descriptions are per the HXB2 Annotated Spreadsheet (Bette T. Korber et al., 2017).

For the complete list of EVM figures and selections from this study, please refer to Appendix B.

The second set of results come from a subset of sequences from (Morton et al., 2017). Table 4.10 provides the sequence names and source clades for the 20 gp120 proteins analyzed.

Sample EFP graphs, to confirm proper processing of data, are show in Figure 4.26 .

For the referenced sequences, Figure 4.27 display the predicted pH sensitivity map across residues of this study. Figure 4.28 provides the associated screeplot with the cutoff

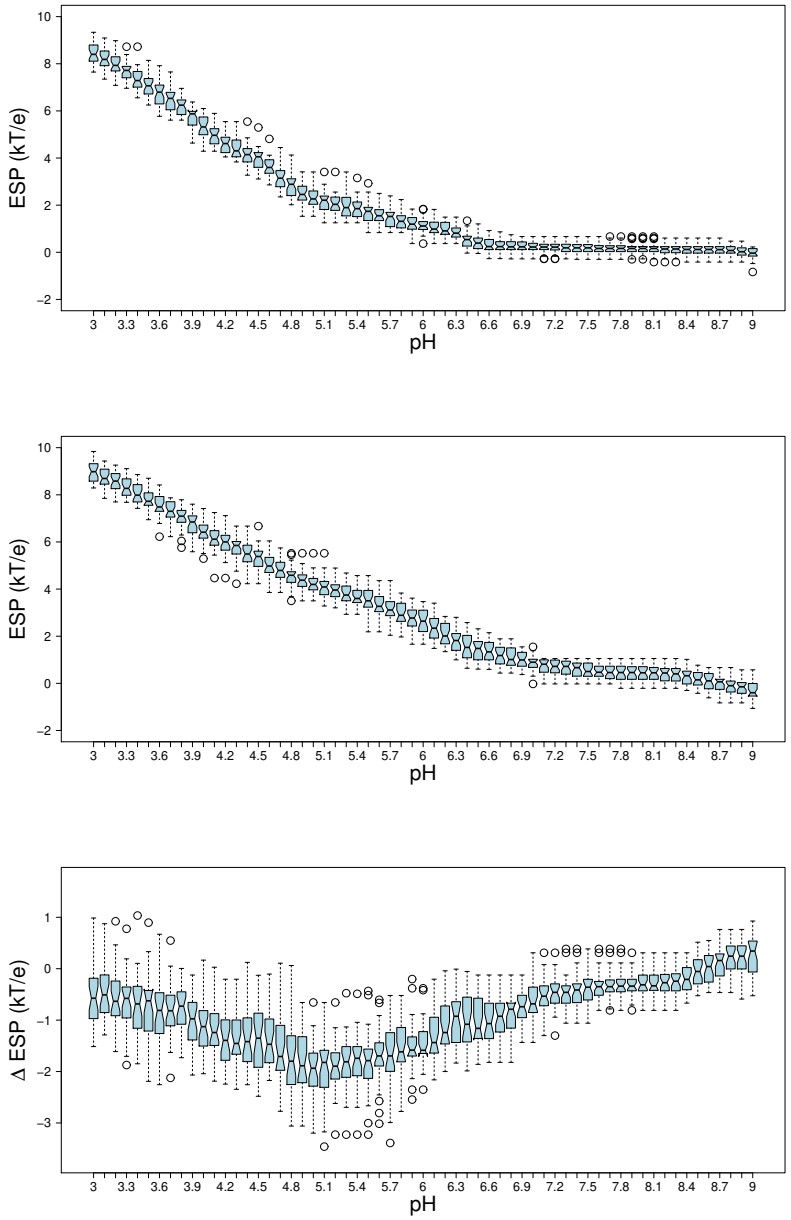

Figure 4.26: Electrostatics data for sequence R56MCF21aug0511_plasmid_1v displays a normal descent of charge from low to high pH for bound (top) and unbound (middle) conformations. Bound less unbound charge data (bottom) displays the signature EFP typical of this protein structure, indicating that the pipeline has processed accordingly.

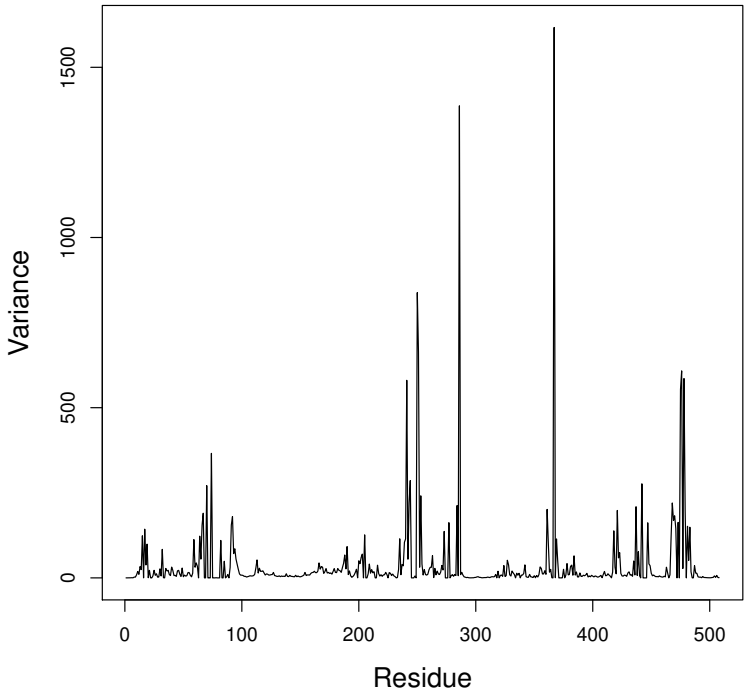

Figure 4.27: Variance map of all sequences based on the method described presents a clear signal.

value shown by the red horizontal line. Table 4.11 shows the statistical data returned from EVM for the selected residues.

Table 4.11: Statistics for selection of high variance residues. The standard deviation is for the variance data set. The cutoff value is the minimum variance value that excludes all sequence alignment gaps and the % of variance selected is the percentage of the variance in the data set.

Standard Deviation	*123.7*
1/2 Standard Deviation	*61.8*
Number of Selected Residues	*56.0*
Variance cutoff selected	*65.0*
% of variance selected	*73.6%*
% of residues selected	*11.0%*

The following Weblogo representation shows all selected sequences in Figure 4.29. Note the selected residues are highly conserved across variations of gp120 in (Morton

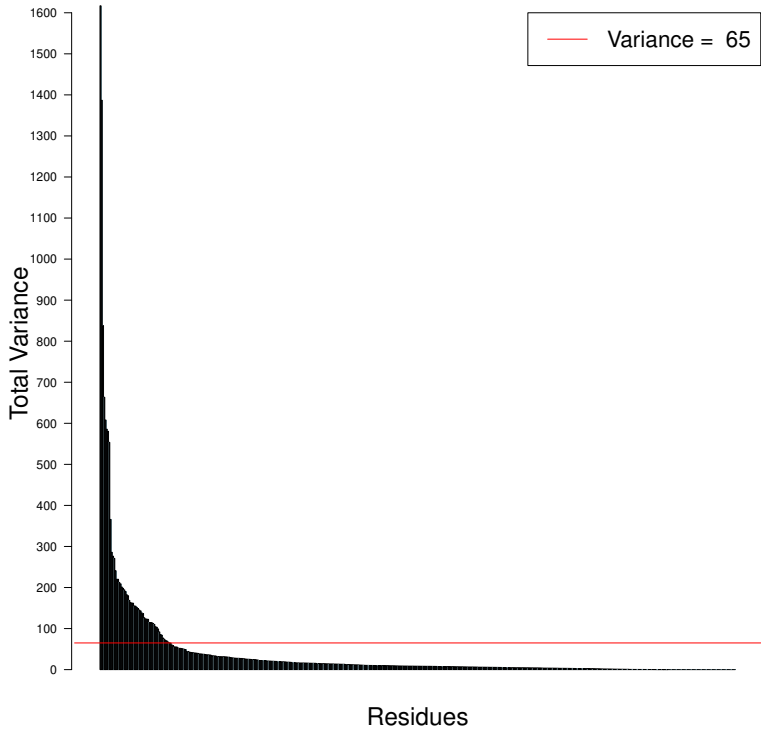

Figure 4.28: Screeplot of the variance data with the cutoff value shown as a red horizontal line.

et al., 2018). This data was further separated by clade to provide two additional Weblogo representations, Figure 4.30 allows us to visualize clade A1 sequence residue variability, and Figure 4.31 allows us to visualize clade C sequence residue variability. Clade A1 sequences have a smaller representation in this study that explains the low amplitude of the graph.

Figure 4.29: Weblogo representation of EVM selected residues for sequences. note that these residues are highly conserved across all 20 gp120 variants. Commonality is indicated by taller lettering and stacking indicates differences.

Figure 4.30: Weblogo representation of EVM selected residues for sequences. showing the conservation of residues among clade A1 variations of gp120. Clade A1 has fewer sequences analyzed (4) as an explanation of the lower amplitude observed. Commonality is indicated by taller lettering and stacking indicates differences.

Figure 4.31: Weblogo representation of EVM select residues for sequences showing the conservation among clade C variations of gp120. Commonality is indicated by taller lettering and stacking indicates differences.

EVM produces actual residue numbering lists (not aligned with HXB2). The following are selection lists for two typical sequences:

- R56MCF21aug0511_plasmid_1v: length - 56

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 161 163 176 178 207 211 212 213 215 216 222 223 225 235 245 249 256 258 330 331 336 338 378 381 383 397 399 402 407 423 424 425 426 427 429 431 432 434 435 437 439

- Z201FPL7FEB03ENV2.1: length - 56

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 174 176 189 191 220 224 225 226 228 229 235 236 238 248 258 262 269 271 343 344 349 351 395 398 400 414 416 419 424 437 438 439 440 441 443 445 446 448 449 451 453

Applying the imaging method previously described we produce Figures 4.32 and 4.33. In particular we see that residues of the CD4bs are highly conserved between the two variants.

Figure 4.32: EVM imagery displaying the selected residues for sequence R56MCF21aug-0511_plasmid_1v in red.

Finally, for this set of sequences, residues are aligned to HXB2 as previously described to express the residues selected in terms of HXB2 sequence alignment. Processing returned the following list for all sequences in this set:

- 47 49 51 64 91 96 98 99 102 106 114 123 124 125 126 199 201 214 216 245 249 250 251 253 254 260 261 263 273 283 287 294 296 370 371 376 378 426 429 431 445 447 450 455 470 471 472 473 474 476 478 479 481 482 484 486

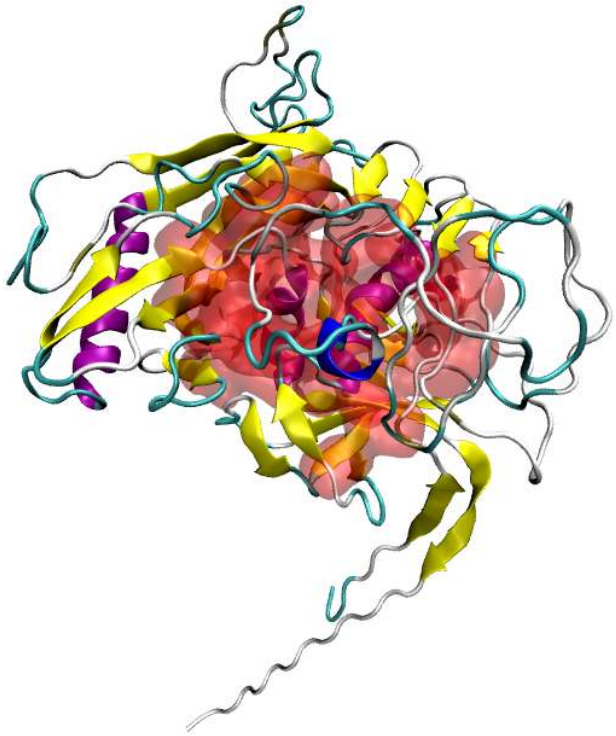

Figure 4.33: EVM imagery displaying the selected residues for sequence Z201FPL7-FEB03ENV2.1 in red.

Selected residues are described as follows: Residues 64 and 91 are adjacent to 65 and 92, respectively, which are interface contacts with gp41; 123 which is a co-receptor binding site outside of V3 and adjacent to 122 of the same function; 124-126 are CD4 contact residues; 199 is a co-receptor specific (R5/X4) site; 201 is adjacent to 202 is a co-receptor binding site outside of the V3 loop; 249-251 where 251 is co-receptor specific (R5/X4) site; 253 is adjacent to 252, which is a interface contact with gp41; 261 and 263 are adjacent to glycosite 262; 283 is a CD4 contact residue; 294 is adjacent to glycosite 295; 296 is the beginning of V3 loop; 370 is a CD4 contact residue and 371 is adjacent; 376 is adjacent

to 377, a co-receptor binding site outside of V3; 378 is Cysteine linked to a counter part at 445; 426, 429, 431 are CD4 contact residues; 445 is Cysteine linked to a counter part at 378; 447 is adjacent to glycosite 448; 455 is a CD4 contact residue; 470 is the V5 loop end and adjacent to CD4 contact residue 469; 471-476 are CD4 contact residues.

For the complete list of EVM figures and selections from this study, please refer to Appendix B.

Discussion

The data presented in this section displays EVM in a single mode of operation that represents the predicted conservation of residues across entire sets of gp120 structures. We observe that, when applied to different sequence sets, EVM selected residues differ in number, but in every case, EVM selected CD4bs residues responsible for the mechanistic binding function. These data indicate EVM is a powerful tool for understanding the binding function and how environmental factors such as pH affect the binding process. It may be interesting to perform EVM selection on individual clade, sub-class and single gp120 protein structures.

Some considerations of this potential use would involve adjusting the initial cutoff value to a larger number, most likely to one standard deviation, to avoid capturing too many residues. Currently, EVM attempts to capture the largest amount of variance in the fewest possible residues. As calibrated, EVM captures approximately 50% of the total variance. Although the total variance captured is dynamic, it depends on the current set of structures being analyzed. The consistency of core residues captured suggests the method is robust. Future studies will be devised that exploit alternate implementations, as describe above, that may expose more granular details of surface charge modulation due to environmental pH.

Binding Energies

Experimental and computational studies have shown pH to alter gp120 conformation and impact binding to CD4 (Mason and Jensen, 2008); it stands to reason pH would also impact gp120 to bnAb binding as well. Protein interactions involving gp120 and CD4 were previously modeled using solved gp120 structures (Stieh et al., 2013), and led to predictions that lower pH enhanced attraction between the positive charge of CD4 to the negative charge of gp120; these modifications were not elaborated on until investigated in Howton (2017), and Howton and Phillips (2017). Interactions involving BE of gp120 to bnAb have not previously been investigated using the ESSC pipeline.

The Adaptive Poisson-Boltzmann Solver (APBS) (Baker et al., 2001; Jurrus et al., 2018) provides charge data at the molecule and atomic levels, but has additional features that allow calculating binding free energy as well. The "free energy cycle" (Baker et al., 2001; Jurrus et al., 2018) is determined by the "elecEnergy" values for the complex of primary and secondary structures (which are returned by APBS), where the bound binding energy value 'ΔBE_b' of the complex structure 'ΔG_c', less the value of bound gp120 conformation 'ΔG_{eb},' less the charge of a secondary structure (CD4 or bnAb) 'ΔG_a' such that the formula:

$$\Delta BE_b = \Delta G_c - \Delta G_{eb} - \Delta G_a.$$

is satisfied for each possible combination. The same logic holds true for the unbound gp120 combinations, as expressed in this formula:

$$\Delta BE_{ub} = \Delta G_c - \Delta G_{eub} - \Delta G_a.$$

However, this is counter intuitive to the molecular process as the primary structure (gp120) would not naturally be bound to a counter part molecule in the unbound state, and hence would quickly transition to a natural bound conformation. This data, where calculated,

will be presented for completeness. In order to represent this data in the clearest possible manner, alternative graph presentations are employed.

For the purposes of expressing BE, all theoretical data will be normalized for clarity using:

$$x' = \frac{x - set_{min}}{set_{max} - set_{min}}$$

Where x is the theoretical value, set_{min} is the minimum value of all theoretical data produced, set_{max} is the maximum value of all theoretical data produced, and x' is the normalized value returned.

This method presents the general hypothesis that *binding energies increase as pH increases*.

Results

Source sequences and process confirmation for this study were as presented previously. This study differs from the previous in that it involves the CD4 protein, the sequence for which was sourced from PDB 1RZK_2.

Here we reproduce the work in Howton (2017), and the work in Howton and Phillips (2017), to determine if the enhanced protocols of the modernized ESSC pipeline will alter the results of those studies. Howton et al. hypothesized that *differences in the transmissibility of the TF variant as compared to the CC variant of gp120 could be determined in binding energy characteristics*. In both studies, the results were inconclusive in consideration of TF and CC, however, both studies indicated that binding energies of gp120 and CD4 in complex increased with increases in pH.

Originally, the representation of data involved taking mean values of data across the set to produce predictions among clade and sub-class variants of gp120. Figure 4.34 displays an aggregation of all binding energy motifs. The graph displays a range of pH (approximately pH 5.1 to 8.9 indicated by red shading) where the general hypothesis that *binding energies increase as pH increases* applies.

The data provides clade sources for each gp120 to allow for comparison of clade. Figure 4.35 reveals little in overall differences between the two clade. Similarly, we see few differences between BE's of TF and CC subclades of clades B and C in Figure 4.36.

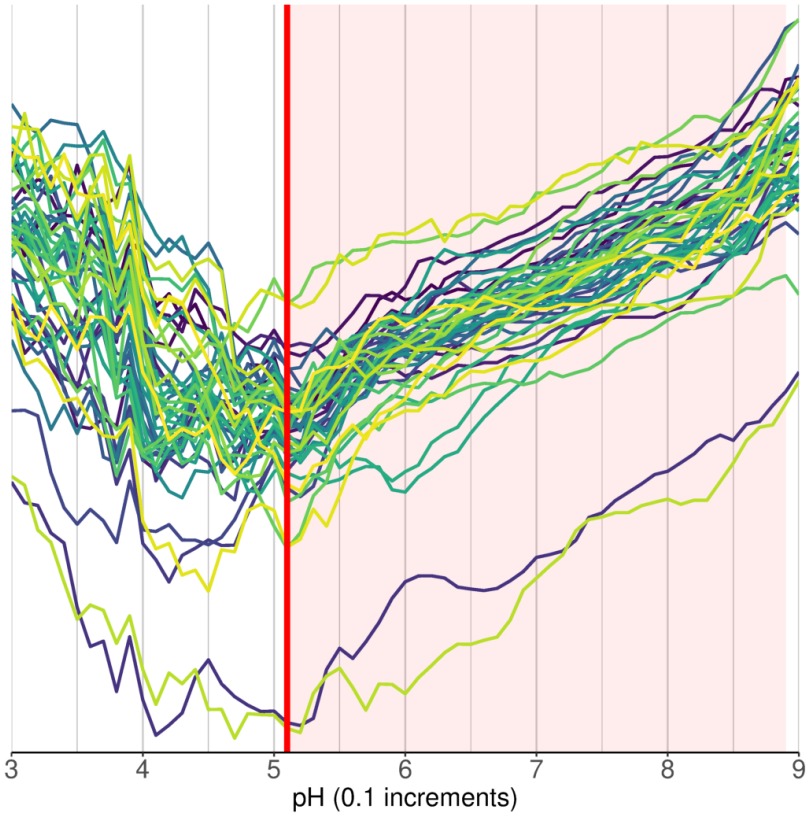

Figure 4.34: Aggregation of all binding energy motifs for sequences of this set displays a range (approximately pH 5.1 to 8.9) where BE moves more positive as pH increases. Red shading indicates the approximate range of agreement with the general hypothesis (*binding energies increase as pH increases*).

One should note the small percentage of gp120 variants that have enhanced potential to bind at low pH values, as indicated by their binding energies (see Figure 4.37).

We use BESI scores to further analyze the small percentage of gp120 variants which displayed interesting differences in binding energies. From clade C, we examine 53_CH569TF, 55_CH42TF, and 56_CH42M6 with BESI scores of 0.6678, 0.8212, and 0.9548 respec-

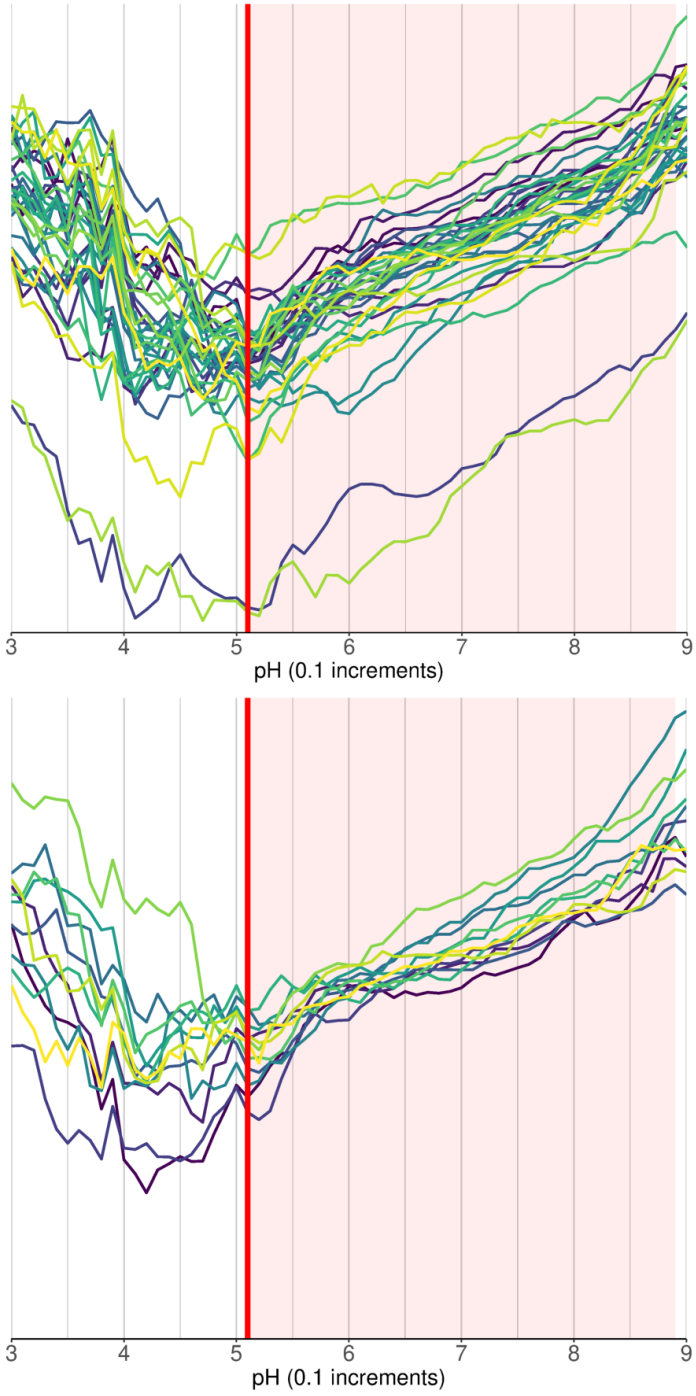

Figure 4.35: Comparison of clade B (top) and clade C (bottom). The number of gp120 variations in clade B allow for a broader representation of predicted BE versus clade C, however, the two clades display similar characteristics overall. Red shading indicates the approximate range of agreement with the general hypothesis (*binding energies increase as pH increases*).

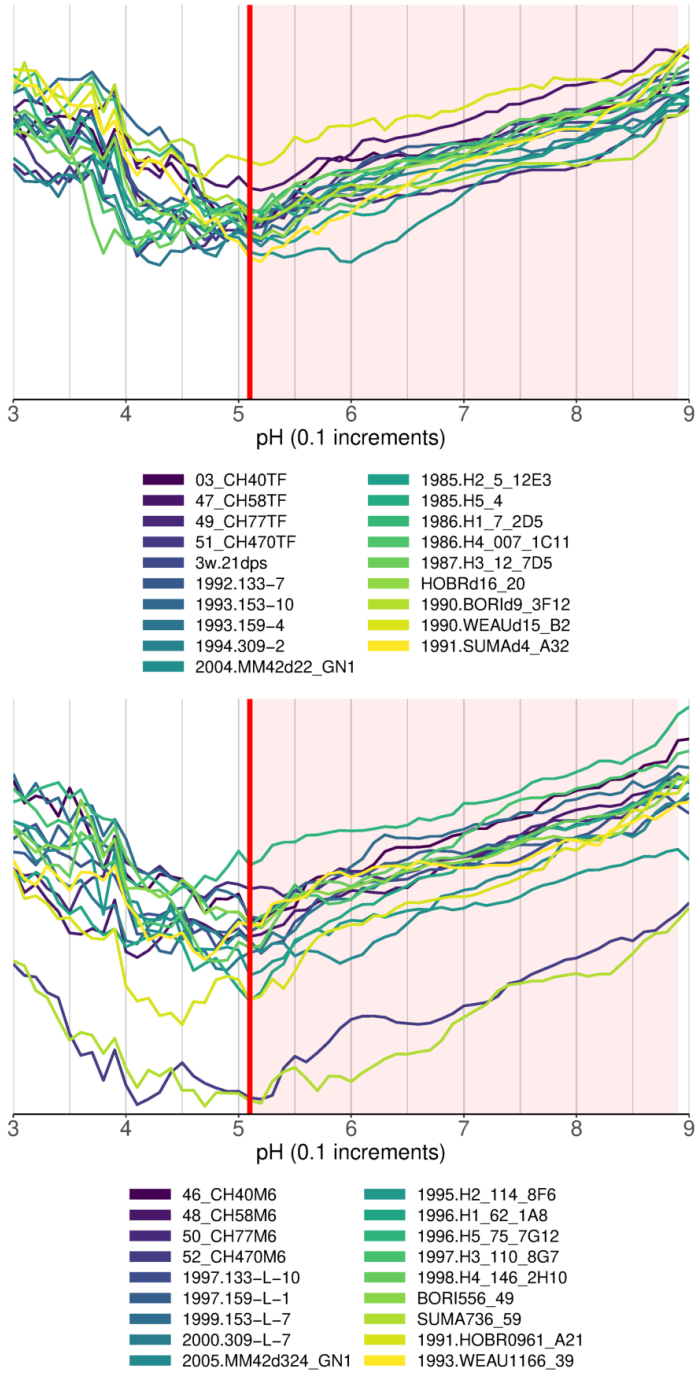

Figure 4.36: From clade B a comparison of sub-class TF (top) versus CC (bottom). No discernible differences standout in predicted BE across the two sub-classes. Red shading indicates the approximate range of agreement with the general hypothesis *binding energies increase as pH increases*.

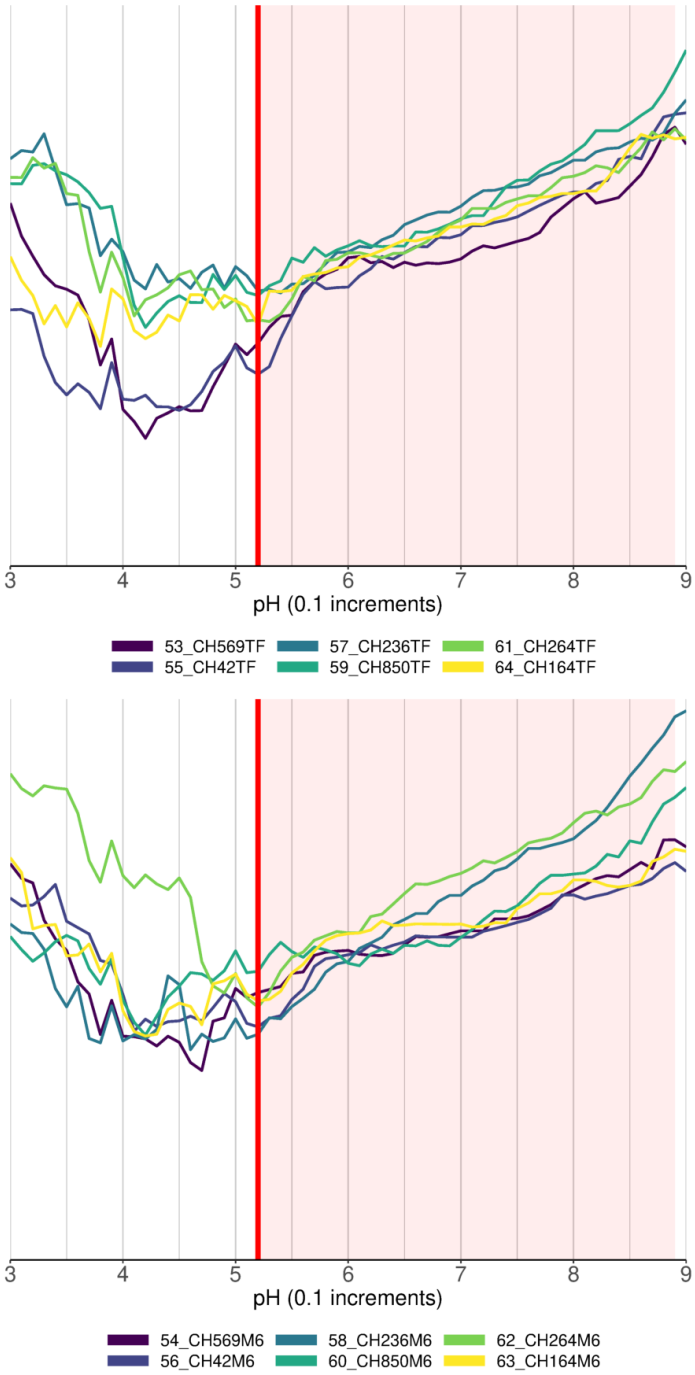

Figure 4.37: From clade C a comparison of sub-class TF (top) versus CC (bottom). No overall differences standout in predicted BE across the two sub-classes.

tively. Two out of three scores indicate a strong potential to cross the transmission boundary.

These next results are from a collaborative effort between traditional biology methods and methods of computational science intended to validate computer simulations against laboratory results. The laboratory results are from four separate experiments evaluating multiple known antibody binding locations: CD4bs, V2/V3/Glycan, N332 Glycan, Glycan, MPER and Polyclonal. Eleven different gp120 proteins were screened against fifteen bnAbs at pH 5.5 and pH 7.4. Limitations in available crystal structures restrict the computational pipeline to assessment of CD4bs at this time, but further studies could be performed on alternate binding targets when suitable structures become available. As a result, we only process 4 of the 15 bnAbs with the computational pipeline: 3BNC117, B12, CH31, and VRC01. These antibodies are bound to eleven monomer gp120 variants:

- 1056_10 (EU289186)
- 6101_1 (AY835434)
- 6535_3 (AY835438)
- CAAN_A2 (AY835452)
- PVO_4 (AY835444)
- RHPA_7 (AY835447)
- THRO_18 (AY835448)
- TRJO_58 (AY835450)
- TRO_11 (AY835445)
- WEAU_d15 (EU289202)

- WITO_33 (AY835451)

Experimental results are compiled in Tables 4.12, 4.13, and 4.14. Duplication of experiment 1 was used for confirmation of the methods. Results are viewable in Appendix C. The data in each table is laid out in three sections horizontally. The top section contains the experimental data, the mid and lower sections contain the simulation data for bound and unbound gp120 conformations, respectively. The experimental data is focused on specific regions of protein-protein interactions indicated under the column Specificity. Each table is shaded where the values increase from lower to higher pH.

Inspection of experimental data reveals greater than 50% of results exhibit binding energies that increase with pH (see Figure 4.38). V2/V3/Glycan regions being the only binding location to contradict the general hypothesis that *binding energies increase as pH increases*. Looking at the bottom of Figure 4.38, we see that most variations conform to the general hypothesis that *binding energies increase as pH increases*. However, RHPA_7 and WITO_33 are exceptions.

Lab experiments produce eighty unique complexes (four sets did not include bnAb B12) and BE's for each complex were analyzed at low and high pH to determine if BE varies significantly with pH. Some gp120 proteins (65535_3, CAAN_A2, TRO_11, WITO_33) were analyzed twice, in which case the most significant results are selected to reduce statistically indeterminate values. The number of indeterminate from experiments was 23.75%. This leaves 30 comparable complexes with determinate experimental results for which computational/theoretical results are also available. We note that simulations resulted in no statistically indeterminate computational/theoretical results.

Figure 4.39 compares lab results with theoretical data. The two sub panels represent lab experiments (top) and theoretical predictions (bottom) with blue and red indicating pH 5.5 and pH 7.4, respectively. The $+/-$ markers represent the direction of change in binding energy from lower to higher pH. Agreement, disagreement and indeterminate experimental

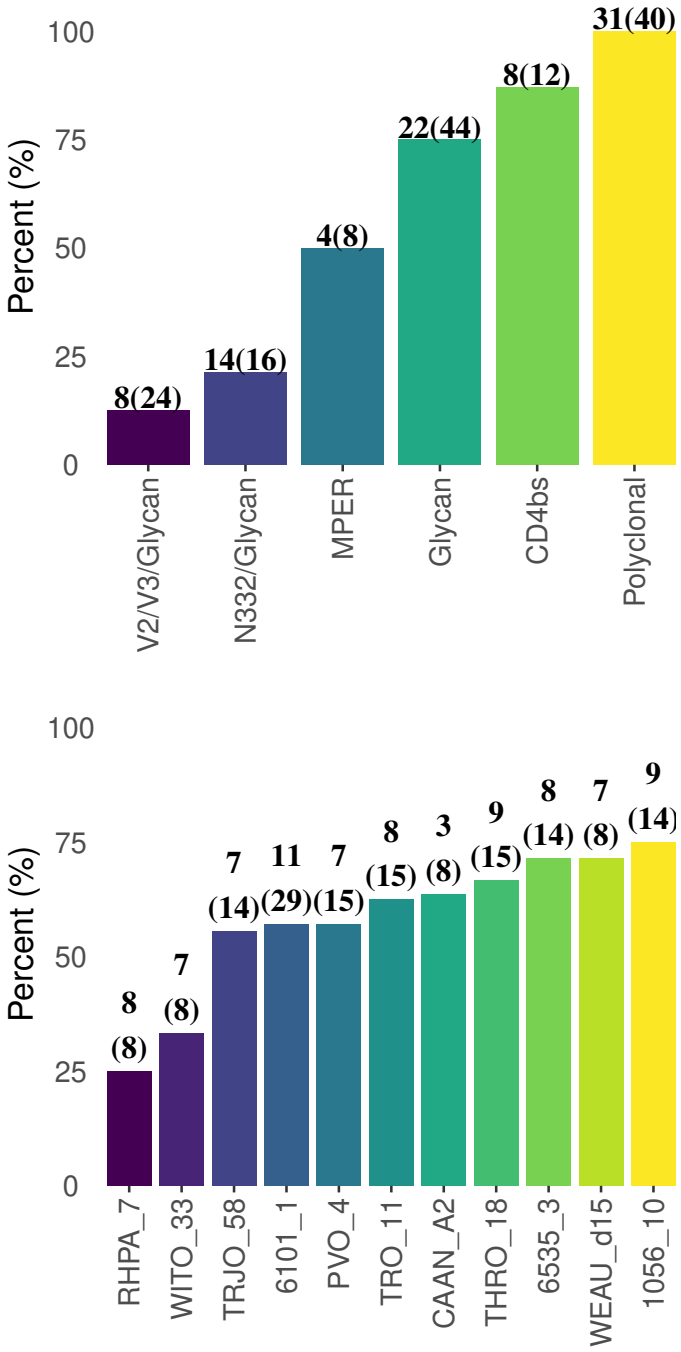

Figure 4.38: (top) Aggregation of all binding energy data grouped by binding location from experimental results. (bottom) Aggregation of all binding energy data grouped by gp120 from experimental results. Columns represent the percentage of entries where binding energies increase as pH rises. Label values, *x (y)*, represent the number of entries used for calculation (x) and the number of experimental entries including statistically indeterminate values (y).

Table 4.12: Binding energies for various combinations of gp120 and bnAb interactions. Data is from laboratory experiment 1 (a), theoretical simulations for bound (b),and unbound (c) conformations. Shading indicates a positive shift from pH 5.5 and pH 7.4.

(a) Experimental Data		WITO_33		TRO_11		CAAN_A2		6535_3	
Mab	Specificity	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**
VRC01	CD4bs	0.2	0.57	0.81	1.81	0.48	4.05	2.25	10.3
3BNC117	CD4bs	0.06	0.11	0.08	0.09	0.64	1.86	0.5	1.49
CH31	CD4bs	0.39	0.36	0.14	0.46	1.67	>25	>25	>25
CH01	V2/V3/Glycan	0.22	0.18	>25	>25	>25	>25	1.08	3.15
PG9	V2/V3/Glycan	0.07	0.07	>5	>5	>5	>5	2.35	1.3
PG16	V2/V3/Glycan	0.04	0.01	>5	0.85	>5	>5	>5	>5
PGT121	N332 Glycan	4.53	2.65	0.02	0.06	0.03	0.06	0.01	0.03
PGT128	N332 Glycan	>5	>5	0.08	0.04	0.29	0.26	0.02	0.02
B12	CD4bs	>25	>25	>25	>25	>25	>25	>25	11.86
2G12	Glycan	3.43	1.91	0.45	0.43	>25	>25	>25	10.51
2F5	MPER	3.13	>25	>25	>25	19.48	>25	>25	>25
4E10	MPER	8.13	>25	>25	15.47	5.14	>25	>25	>25
10E8		1.02	1.6	0.67	0.51	>5	>5	>5	1.55
HIVIG-C	Polyclonal	>625	>625	11.59	305.24	104.34	418.5	22.33	85.04
(b) Theoretical Data (Bound Conformation) (kJ/mol)									
VRC01	CD4bs	-738.842	134.421	690.625	724.008	-577.448	525.957	-995.402	-21.477
3BNC117	CD4bs	-577.435	-276.686	267.382	-311.158	-1467.642	-982.648	-258.319	-352.997
CH31	CD4bs	-614.313	22.164	-431.106	-695.141	-980.249	-164.143	-511.745	-419.132
B12	CD4bs	253.089	805.424	1052.882	850.447	672.888	957.792	467.298	326.361
(c) Theoretical Data (Unbound Conformation) (kJ/mol)									
VRC01	CD4bs	-246.47	-254.385	611.395	246.099	253.613	180.33	-600.143	-168.005
3BNC117	CD4bs	-110.491	-650.74	361.673	-603.765	-807.513	-1521.939	88.97	-569.129
CH31	CD4bs	125.187	-105.794	-62.67	-724.235	-24.053	-416.22	243.916	-192.191
B12	CD4bs	570.553	272.138	898.522	296.377	1542.047	648.147	675.992	20.267

data are indicated with green, yellow and gray shading, respectively. Complexes are noted as bnAb/gp120 along the horizontal axis.

Binding energies increased for $26/30$ complexes, or 86.67% of experimental complexes (top sub-panel of Figure 4.39). Theoretical predictions indicate a similar pattern with $31/40$ complexes, or 77.5% exhibiting binding energies that increase with pH (see the lower sub-panel of Figure 4.39). Looking at experimental bnAb binding energies individually we find that 90.9% of 3BNC117 complexes, 63.6% of CH31 complexes, 0.0% of B12 complexes, and 81.8% of VRC01 complexes exhibit binding energies that increase with pH. Meanwhile, theoretical predictions present 72.7% of 3BNC117 complexes , 42.8% of B12 complexes, 81.8% of CH31 complexes, and 100% of VRC01 complexes exhibit binding energies that increase with pH. Overall, the two methods are in 80% agreement; ten comparisons are indeterminate.

Table 4.13: Binding energies for various combinations of gp120 and bnAb interactions. Data is from laboratory experiment 3 (a), theoretical simulations for bound (b),and unbound (c) conformations. Shading indicates a positive shift from pH 5.5 and pH 7.4.

(a) Experimental Data		CAAN_A2		PVO_4		RHPA_7		TRJO_58	
Mab	Specificity	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**
VRC01	CD4bs	2.46	3.93	1.79	2.5	0.18	0.18	0.3	0.36
3BNC117	CD4bs	1.52	1.4	0.19	0.26	0.06	0.06	0.2	0.28
CH31	CD4bs	>25	>25	0.68	1.21	0.26	0.46	0.2	0.68
CH01	V2/V3/Glycan	>25	>25	>25	>25	>25	>25	>25	>25
PG9	V2/V3/Glycan	>5	>5	>5	>5	>5	>5	1.52	1.38
PG16	V2/V3/Glycan	>5	>5	>5	>5	2.99	2.54	3.71	0.91
PGT121	N332 Glycan	0.10	0.07	1.3	1.25	0.12	0.10	>5	>5
PGT128	N332 Glycan	1.31	0.47	0.08	0.06	0.14	0.12	0.08	0.08
B12	CD4bs	>25	>25	>25	>25	0.82	0.71	>25	>25
2G12	Glycan	>25	>25	8.85	5.41	>25	>25	>25	>25
2F5	MPER	17.38	>25	>25	>25	>25	>25	>25	>25
4E10	MPER	>25	>25	>25	>25	>25	>25	>25	>25
10E8	MPER	4.92	>5	>5	>5	>5	>5	4.28	3.21
DH512	MPER	8.91	15.8	27.51	36.69	24.2	32.69	8.79	13.97
HIVIG-C	Polyclonal	247	452	>625	>625	>625	>625	277	406
(b) Theoretical Data (Bound Conformation) (kJ/mol)									
VRC01	CD4bs	-577.448	525.957	-1655.079	-337.444	-365.739	-265.319	-15.73	749.636
3BNC117	CD4bs	-1467.642	-982.648	-1214.111	-724.872	998.961	47.708	232.436	257.113
CH31	CD4bs	-980.249	-164.143	-1802.553	-1171.902	-101.613	-542.056	-528.289	129.654
B12	CD4bs	672.888	957.792	143.964	1038.709	488.658	349.589	1079.109	850.867
(c) Theoretical Data (Unbound Conformation) (kJ/mol)									
VRC01	CD4bs	253.613	180.33	-338.03	-217.35	463.952	-142.728	-86.706	106.753
3BNC117	CD4bs	-807.513	-1521.939	-259.281	-978.416	1799.105	120.993	55.545	-462.854
CH31	CD4bs	-24.053	-416.22	-684.752	-1228.829	637.608	-528.427	-380.298	-314.488
B12	CD4bs	1542.047	648.147	729.007	440.539	1204.27	299.473	1305.875	546.177

Figure 4.40 shows four panels with binding patterns of the selected bnAbs to gp120: (A) 3BNC117, (B) B12, (C) CH31, and (D) VRC01. Clearly, bnAbs dictate binding potential motifs of the complexes across varying pH levels. The red vertical bar is conservatively placed at the approximate point where unpredictable binding energies end and the red shaded background predicts the functional binding range of each bnAb where the general hypothesis that *binding energies increase as pH increases* is supported. Additionally, our data supports the hypothesis that *fluctuations in the starting pH and range of expected binding predictability controls breadth of bnAb effectiveness in vivo*, as suggested by (Stieh et al., 2013; Morton et al., 2017). Additional studies are required to identify where each bnAb denatures at the extremes of pH to provide further evidence to support this hypothesis.

Table 4.14: Binding energies for various combinations of gp120 and bnAb interactions. Data is from laboratory experiment 3 (a), theoretical simulations for bound (b),and unbound (c) conformations. Shading indicates a positive shift from pH 5.5 and pH 7.4.

(a) Experimental Data		1056_10_TA11_1826		6101_1		THRO_18		WEAU_d15_410_5017	
Mab	Specificity	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**
VRC01	CD4bs	1.48	2.31	0.14	0.3	>25	>25	0.22	0.52
3BNC117	CD4bs	0.44	0.55	0.06	0.11	5.83	12.48	0.15	0.31
CH31	CD4bs	0.61	1.62	0.21	1.02	>25	>25	0.17	0.49
2F5	MPER	0.91	2.02	>25	>25	>25	>25	2.31	5.18
4E10	MPER	4.77	5.77	1.58	1.34	>25	>25	5.24	4.95
10E8	MPER	1.32	0.85	0.13	0.05	2.56	2.72	>5	>5
DH512	MPER	0.93	0.84	0.53	0.31	5.99	4.55	1.22	1.17
HIVIG-C	Polyclonal	224.38	359.12	222.8	563.15	554.02	>625	188.78	466.44
(b) Theoretical Data (Bound Conformation) (kJ/mol)									
VRC01	CD4bs	-329.054	493.316	-95.681	765.584	-349.488	1025.603	397.848	1166.446
3BNC117	CD4bs	-522.442	-88.206	-760.644	-519.292	-1083.553	-661.214	184.225	452.75
CH31	CD4bs	-66.364	494.52	-719.118	-610.734	-304.668	720.046	761.281	1573.397
B12	CD4bs	-422.101	419.317	328.044	335.167	286.265	1325.066	440.151	1110.685
(c) Theoretical Data (Unbound Conformation) (kJ/mol)									
VRC01	CD4bs	-321.629	112.58	-43.598	145.586	611.395	246.099	-375.303	38.83
3BNC117	CD4bs	-449.077	-409.769	-565.985	-1007.794	361.673	-603.765	-316.893	-479.116
CH31	CD4bs	-180.928	5.75	-108.169	-652.381	-62.67	-724.235	-121.937	269.598
B12	CD4bs	-188.829	277.933	735.001	94.994	898.522	296.377	248.421	488.277

Mascola et al. compiled data of various site functional bnAbs including breadth and potency specific information (Mascola and Haynes, 2013). Our observations mostly agree with Mascola et al. in terms of breadth and potency versus starting pH and functional range. These results are visible in Figure 4.40 and also expressed here with the following notation: wide, good, moderate, and low are denoted as '++++', '+++', '++', and '+,' respectively:

- 3BNC117 has good breadth with good potency (+++, +++) and good starting pH (5.9) with good range of 2.6 (5.9 - 8.5)
- B12 has low breadth with moderate potency (+, ++) and high starting pH (6.2) with good range of 2.7 (6.2 - 8.9)
- VRC-CH30-34 lineages have good breadth with good potency (+++, +++) and high starting pH with low range 2.2 (6.3–8.5)
- VRC01-03 lineages have wide breadth with good potency (++++, +++) and low starting pH (5.6) with wide range 2.9 (5.6–8.5)

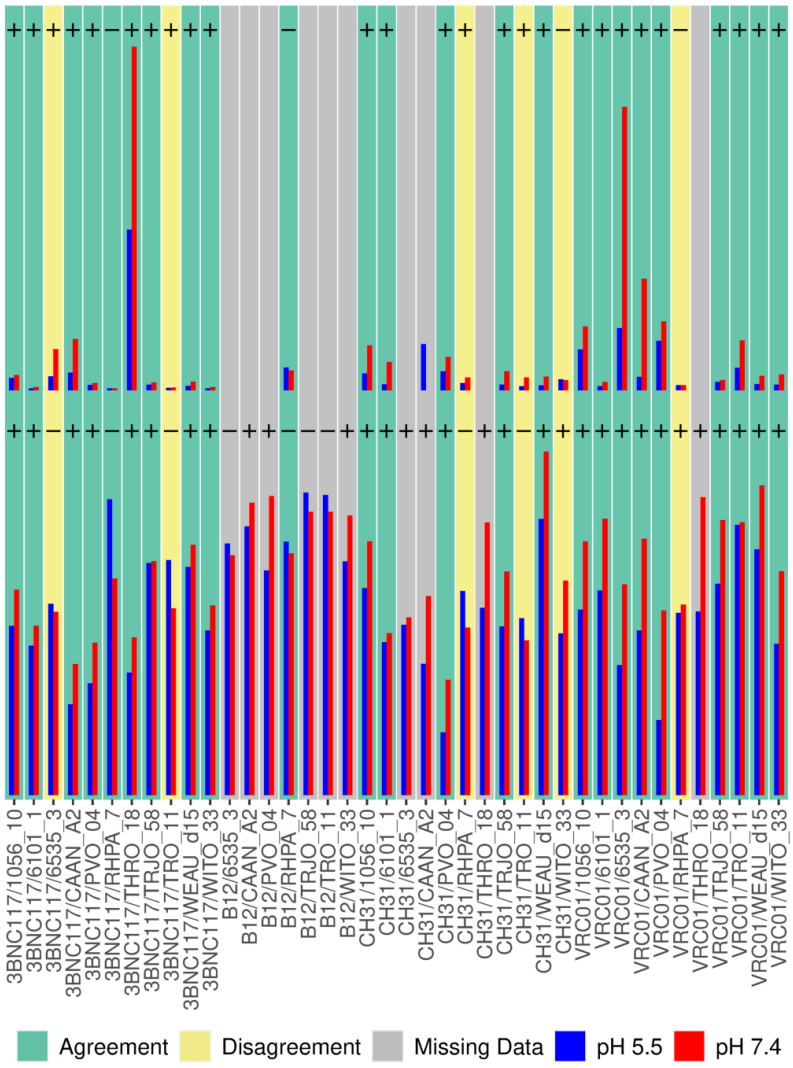

Figure 4.39: Graph showing the comparison of lab results to theoretical data. The two sub panels represent lab experiments (top) and theoretical results (bottom). Blue represents pH 5.5 and red indicates pH 7.4. Markers $(+/-)$ present the direction of change from lower to higher pH. The background color for each method set of results indicates agreement between theory and experiment using a green shade, disagreement using yellow shading while gray indicates indeterminate lab results. Complexes are represented as bnAb/gp120 along the horizontal axis.

Neutralization assays Neutralizing antibodies were measured with Env-pseudoyped viri using TZM-bl cells as targets for infection essentially as described Montefiori (2009), and Li et al. (2005) with minor modification. Briefly, TZM-bl cells were pre-seeded at a density of 8,000 cells/well in 96-well culture plates and incubated overnight at 37°C. In a separate

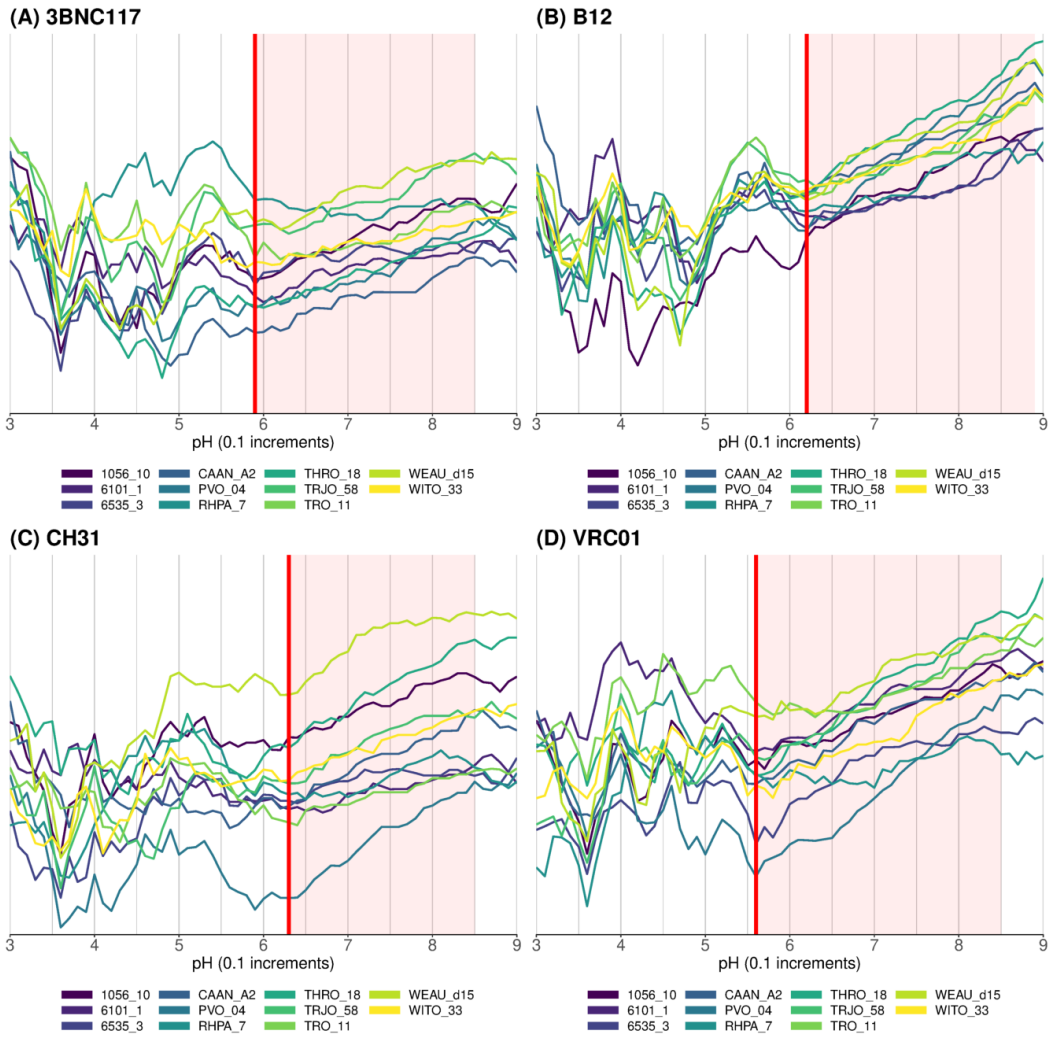

Figure 4.40: Broad spectrum binding energy motifs of bnAbs (A) 3BNC117, (B) B12, (C) CH31, (D) VRC01 displaying the affinity each has binding to the eleven Env proteins analyzed. The red vertical bar is conservatively placed at the approximate pH value where, to the right, outcomes become predictable in their positive movement as pH rises. The shaded background indicates the functional range of predictable activity. Data is the normalized mean of ten models per Complex.

plate, a pre-titrated dose of pseudovirus was incubated for 1 hr at 37°C with serial 3-fold dilutions of test sample in duplicate in a total volume of 150 μl of standard growth medium (DMEM, 10% fetal bovine serum, gentamicin 50 μg/ml, HEPES, pH 7.4) and the same growth medium (-HEPES) adjusted to pH 5.5 using 1N HCl. During this incubation period, all growth medium in the plates containing cells was replaced with 150 μl of either

pH 7.4 or pH 5.5 growth medium containing 75 μg/ml DEAE dextran. After the incubation, the virus/sample mixtures were transferred to the cell plates. One set of control wells received cells + virus (virus control) and another set received cells only (background control). After 48 hours of incubation, 100 μl of cells was transferred to a 96-well black solid plate (Costar) for measurements of luminescence using the Britelite Luminescence Reporter Gene Assay System (PerkinElmer Life Sciences). Neutralization titers are the dilution (serum/plasma samples) or concentration (mAbs) at which relative luminescence units (RLU) are reduced by 50% compared to virus control wells after subtraction of background RLUs. Assay stocks of molecularly cloned Env-pseudotyped viruses will be prepared by transfection in 293T/17 cells (American Type Culture Collection) and titrated in TZM-bl cells as described Montefiori (2009), Li et al. (2005).

Discussion

Binding energy motifs provide no evidence that the range of pH values for which binding energies increase with pH varies between subclasses or clades (Morton et al., 2018). The data does provide some interesting observations about the breadth of pH values where functional binding between gp120 and CD4 occurs, where the general hypothesis that *binding energies increase as pH increases*, applies across a very broad range of pH. The significance of this point becomes apparent on further investigation. Figure 4.40 shows wide variations in binding energies predicted computationally below pH 6.0. These variations may explain, in part, sporadic indicators of negative movement in binding energies as pH increases. In the case of B12, the motif clearly explains the reversal of predicted positive movement in binding energies for most B12 complexes where binding energies exhibit a local maximum at pH 5.5 with a value exceeding that at pH 7.4. B12 is also observed to be centered vertically across complexes more so than others. A detailed analysis of B12 versus other strains of bnAbs would be required to properly understand this phenomenon.

While no data from experiments or theoretical predictions indicate the potential for any bnAb to protect or eliminate HIV at the wide range of physiological pH, theoretical predictions provide an indicator of why HIV may be able to escape the immune response to HIV infection. Looking at the predicted functional range of gp120 to CD4 binding, this range exceeds and in some cases is predicted to extend into the pH range of mucosa where bnAbs investigated to date do not show any predicted functional capabilities.

Comparing BESI to Supervised Machine Learning

This section is presented as validation of the unsupervised methods employed by BESI for the purposes of seeking agreement from more than one technique. For this method, simple artificial neural networks are employed as a binary and tertiary classifier. The methods employ Tensorflow (Abadi et al., 2016) and Keras (Chollet, 2015) as displayed in code snippets. All training data is augmented through derivation of source data by averaging individual model results of target data. With thirty models of a specific target, such as the control variant gp120, model zero is averaged with model one to produce a new set of results, then with model two etc. Then model one results are averaged with model two results etc., until all model results have been cycled through to produce the augmented training data. This method is employed across all specific requirements for each supervised method.

Binary Classifier

Figure 4.41 shows the model construct for the binary classifier using sixty-one inputs with a one hundred and twenty eight node hidden layer to a single node output. Figure 4.42 provides a visual representation of the binary classifier. Dropout layers are only used during training to prevent over-fitting the data and are not depicted in the figure. The neural network is dense (fully meshed) and utilizes Rectified Linear Unit (ReLU) activation for the input and hidden layers. Output is sigmoid activated to complete the binary classifier.

```
model = keras.Sequential(
    [
    layers.Dense(128, input_dim=61, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(128, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(1, activation='sigmoid')
    ]
    )

model.compile(loss='binary_crossentropy',
              optimizer='rmsprop',
              metrics=['accuracy'])

history = model.fit(x_train, y_train,
                    epochs=100, batch_size=50,
                    validation_data=(x_val, y_val),
                    verbose=1)

test_data = np.genfromtxt('inputdata.txt', delimiter=' ')
test_data_labels = np.genfromtxt('sequence.list', delimiter=' ')

results = model.predict(test_data)
```

Figure 4.41: Model construct of a binary classifier using 61 inputs tied to a 128 node hidden layer that feeds a single output node.

The network is trained with one hundred epochs using augmented training and validation data as previously described. Figures 4.43 and 4.44 express training and validation loss and accuracy respectively.

Results

Scoring by the binary classifier can be see in Table 4.15 for all 252 sequences used in (Morton et al., 2017). The results indicate fitting issues are experienced by the neural network. Some parameter exploration took place with marginal changes to the results (data not shown).

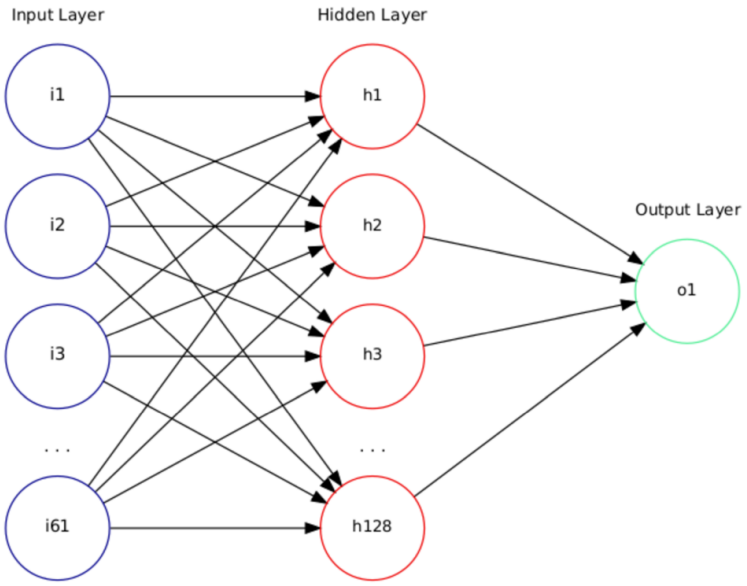

Figure 4.42: Graphic representation of binary classifier showing individual layers. Dropout layers are not expressed and are only used during training to control over and under fitting.

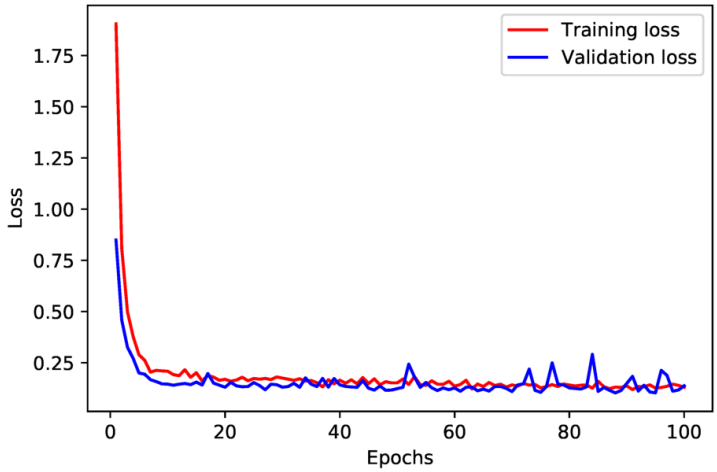

Figure 4.43: Graph showing training and validation losses for the binary classifier in Figure 4.41.

Table 4.15 – *Continued on next page*

Table 4.15 – *Continued from previous page*

Sequence	Score	Sequence	Score

Table 4.15: Scoring from binary classifier showing under-fitting by the neural network.

Sequence	Score	Sequence	Score
Z238FSW15A6_plasmid_6v	0.9819	Z201FPL07feb03102-1	1
Z238FSW15G4_plasmid_4i	0.9999	Z201FPL07feb03103-1	1
Z238FSW15H8_plasmid_3ii	1	Z201FPL07feb03105-1	1
Z238FSW29oct0215A11	1	Z201FPL07feb0350-2	1
Z238FSW29oct0215A6v	0.9767	Z201FPL07feb0351-1	0.8955
Z238FSW29oct0215G4	0.9999	Z201FPL07feb0368-2	1
Z238FSW29oct0215H8	1	Z201FPL07feb0372-1	1
Z238MPL17_plasmid_a	0.8942	Z201FPL07feb0390-1	1
Z238MPL9_plasmid_c	0.3827	Z201FPL100_plasmid_8-1	1
Z242FPL25JAN03PCR23ENV1.1	0.9998	Z201FPL102_plasmid_7-1	1
Z242FPL25JAN03PCR8ENV1.1	1	Z201FPL103_plasmid_4-1	1
Z242FPL25jan038_plasmid	1	Z201FPL105_plasmid_3-1	1
Z242MPL25JAN0326	0.8888	Z201FPL50_plasmid_5-2	1
Z242MPL25JAN0327-1	1	Z201FPL51_plasmid_1-1	0.913
Z242MPL25JAN0327-2	0.9999	Z201FPL68_plasmid_6-2	1
Z242MPL25JAN0327-3	1	Z201FPL72_plasmid_9-1	1
Z242MPL25JAN03PCR23ENV1.1-DT	1	Z201FPL7FEB03ENV1.8	1
Z242MPL25JAN03PCR33ENV1.1-DNT	0.0016	Z201FPL7FEB03ENV2.1	1
Z242MPL25jan0323_plasmid	1	Z201FPL7FEB03ENV3.3	1
Z242MPL25jan0326_plasmid	0.9648	Z201FPL7FEB03ENV4.1	1
Z242MPL25jan0328_plasmid_8-1	1	Z201FPL7FEB03ENV5.2	1
Z242MPL25jan0328_plasmid_8-2	1	Z201FPL7FEB03ENV6.1	1
Z242MPL25jan0328_plasmid_8-3	0.9998	Z201FPL7FEB03ENV7.1	0.8914
Z242MPL25jan0333_plasmid	0.0018	Z201FPL90_plasmid_2-1	1
Z242MPL26_plasmid	0.9942	Z201FSW07feb03DNA13D1	1

Table 4.15 – *Continued on next page*

Table 4.15 – *Continued from previous page*

Sequence	Score	Sequence	Score
Z242MPL28_plasmid_8-1	1	Z201FSWDNA13D1_plasmid_4i	1
Z242MPL28_plasmid_8-2	1	Z201MPB7FEB03ENV2.1	1
Z242MPL28_plasmid_8-3	1	Z201MPB7FEB03ENV4.1	1
Z292FCA12A52_plasmid_9v	1	Z201MPB7FEB03ENV5.1	0.9999
Z292FCA24may0512A52	1	Z201MPL07feb0352a	1
Z292FCA24may0512A52_plasmid_9v	1	Z201MPL07feb0352aa	0.9995
Z292FCA24may0512A58_plasmid_6v	0.9656	Z201MPL07feb0352e	0.9998
Z292FCA24may0512D10_plasmid_5iii	1	Z201MPL07feb0384c	0.9999
Z292FCF12E26_plasmid_10iv	1	Z201MPL52_plasmid_a	0.9999
Z292FCF24may0512D18_plasmid_4i	1	Z201MPL52_plasmid_aa	0.9999
Z292FCF24may0512E26	1	Z201MPL52_plasmid_e	0.9998
Z292FCF24may0512E26_plasmid_10iv	1	Z201MPL7FEB03ENV2.1	0.9998
Z292FPL24may05105_plasmid_5-1	1	Z201MPL7FEB03ENV3.1	1
Z292FPL24may05136_plasmid_7-1	1	Z201MPL7FEB03ENV4.1	1
Z292FPL24may05152_plasmid_1-3	0.9999	Z201MPL84_plasmid_c	0.9957
Z292FPL24may05160_plasmid_4-1	1	Z205FPB27MAR03ENV1.1	0.8165
Z292FPL24may05164_plasmid_9-2	1	Z205FPB27MAR03ENV4.2	0.7305
Z292FPL24may05172_plasmid_6-1	1	Z205FPL27MAR03ENV4.1	0.019
Z292FPL24may0535_plasmid_3-3	1	Z205FPL27MAR03ENV5.2	1
Z292FSW24may0512E12_plasmid_3v	0.9996	Z205FPL27MAR03ENV6.3	0.6004
Z292FSW24may0512E20_plasmid_2i	1	Z205MPB27MAR03ENV4.1	1
Z292MPL113_plasmid_e	1	Z205MPB27MAR03ENV6.1	1
Z292MPL150_plasmid_b	1	Z205MPB27MAR03ENV9.1	0.0816
Z292MPL24may05113_plasmid_e	1	Z205MPL27MAR03ENV1.1NF	1
Z292MPL24may05113e	1	Z205MPL27MAR03ENV2.3	0.341
Z292MPL24may05150_plasmid_b	1	Z205MPL27MAR03ENV3.1NF	0.9999
Z292MPL24may05150b	1	Z205MPL27MAR03ENV6.3	1
R56FPL21apr05B6_plasmid_a	0.9777	Z216FC17jan04RNAB37	0.9996

Table 4.15 – *Continued on next page*

Table 4.15 – *Continued from previous page*

Sequence	Score	Sequence	Score
R56FPL21apr05B6_plasmid_b	0.9807	Z216FCF17jan04RNAB44	0.0097
R56FPL21apr05E7_plasmid_a	0.5807	Z216FCFRNA11B44_plasmid_2iv	0.0064
R56FPL21apr05E7_plasmid_b	0.9906	Z216FCRNA11B37_plasmid_7i	0.9998
R56MCA21aug0516_plasmid_9iii	0.9999	Z216FPB112_plasmid_e	1
R56MCA21aug053_plasmid_5i	1	Z216FPB85_plasmid_f	0.9289
R56MCA21aug056_plasmid_6iii	1	Z216FPB98_plasmid_e	0.9447
R56MCF21aug0511_plasmid_1v	1	Z216FPL129_plasmid_6-1	1
R56MCF21aug0514_plasmid_2iv	1	Z216FPL138_plasmid_8-3	0.9963
R56MCF21aug0519_plasmid_3ii	1	Z216FPL17jan04112e	1
R56MPL21apr05C2_plasmid_7-1	1	Z216FPL17jan04129	1
R56MPL21apr05C5_plasmid_6-4	1	Z216FPL17jan04138	0.9872
R56MPL21apr05G5_plasmid_5-3	1	Z216FPL17jan04190	0.9994
R56MPL21apr05H3_plasmid_1-3	1	Z216FPL17jan046	0.0008
R56MPL21apr05K4_plasmid_4-1	1	Z216FPL17jan0483	0.0069
R56MPL21apr05K6_plasmid_2-4	1	Z216FPL17jan0485f	0.9793
R56MPL21apr05P5_plasmid_8-1	1	Z216FPL17jan0492	1
Z153FPB13MAR02ENV1.1	0.828	Z216FPL17jan0498e	0.8215
Z153FPB13MAR02ENV2.1	0.9979	Z216FPL190_plasmid_5-1	0.9999
Z153FPB13MAR02ENV3.1	0.3384	Z216FPL6_plasmid_4-4	0.0028
Z153FPB13MAR02ENV4.1	0.9996	Z216FPL83_plasmid_7-2	0.1336
Z153FPB13MAR02ENV5.1	0.9989	Z216FPL92_plasmid_1-1	0.9999
Z153FPL13MAR02ENV1.1	0.9033	Z216FSW17jan04DNA15	0.975
Z153FPL13MAR02ENV2.1	0.2734	Z216FSWDNA11I5_plasmid_5v	0.9916
Z153FPL13MAR02ENV3.1	0.0002	Z216MPL133_plasmid	0.9964
Z153FPL13MAR02ENV4.1	0.9996	Z221FPB7MAR03ENV10.3	1
Z153FPL13MAR02ENV5.1	0.0015	Z221FPB7MAR03ENV11.3	1
Z153FPL13MAR02ENV6.1	0.0014	Z221FPB7MAR03ENV6.4	1
Z153MPB13MAR02ENV1.1	0.9998	Z221FPB7MAR03ENV9.1	0.935

Table 4.15 – *Continued on next page*

Table 4.15 – *Continued from previous page*

Sequence	Score	Sequence	Score
Z153MPB13MAR02ENV2.1	0.8576	Z221FPL08mar0335	0.9796
Z153MPB13MAR02ENV3.1	0.9998	Z221FPL08mar0344	0.0004
Z153MPB13MAR02ENV4.1	1	Z221FPL08mar0348	0.4223
Z153MPB13MAR02ENV5.1	0.9999	Z221FPL08mar0351	0.9989
Z153MPL13MAR02ENV1.1	0.9996	Z221FPL08mar0355	0.0619
Z153MPL13MAR02ENV2.1	0.9995	Z221FPL08mar0371	0.9597
Z153MPL13MAR02ENV3.1	0.9631	Z221FPL08mar0380	1
Z153MPL13MAR02ENV4.1	0.9997	Z221FPL35_plasmid_7-1	0.8827
Z153MPL13MAR02ENV5.1	0.9996	Z221FPL44_plasmid_4-1	0.0008
Z185FPB24AUG02ENV1.1	0	Z221FPL48_plasmid_5-1	0.3918
Z185FPB24AUG02ENV2.1	0	Z221FPL51_plasmid_2-2	0.9996
Z185FPB24AUG02ENV3.1	0	Z221FPL55_plasmid_6-2	0.0259
Z185FPB24AUG02ENV4.1	0	Z221FPL71_plasmid_9-1	0.9746
Z185FPB24AUG02ENV5.1	0	Z221FPL7MAR03ENV1.2	0.9677
Z185FPL17AUG02ENV1.1	0.0003	Z221FPL7MAR03ENV10.4	0.9999
Z185FPL17AUG02ENV2.1	0	Z221FPL7MAR03ENV2.3	1
Z185FPL17AUG02ENV3.1	0.0018	Z221FPL7MAR03ENV3.3	0.9987
Z185FPL17AUG02ENV4.1	0	Z221FPL80_plasmid_8-3	1
Z185FPL17AUG02ENV5.1	0	Z221FSW08mar0314H16iii	0.7743
Z185MPB17AUG02ENV1.2	0.0038	Z221FSW08mar0314H16iv	0.9668
Z185MPB17AUG02ENV1.5	0.0021	Z221FSW14H16_plasmid_6iii	0.9665
Z185MPB17AUG02ENV7.4	0	Z221FSW14H16iv_plasmid_6iv	0.9941
Z185MPB17AUG02ENV7.5	0	Z221MPB7MAR03ENV4.1	0.9832
Z185MPB17AUG02ENV7.6	0	Z221MPB7MAR03ENV5.4	0.8951
Z185MPB17AUG02ENVB17	0	Z221MPB7MAR03ENV6.4	0.2915
Z185MPB17AUG02ENVB6	0	Z221MPL08mar0375a	0.8709
Z185MPB17AUG02ENVC17	0	Z221MPL08mar0375f	0.9939
Z185MPB17AUG02ENVC18	0	Z221MPL75_plasmid_a	0.9284

Table 4.15 – *Continued on next page*

Table 4.15 – *Continued from previous page*

Sequence	Score	Sequence	Score
Z185MPB17AUG02ENVC8	0.0034	Z221MPL75_plasmid_f	0.9941
Z201FCA07feb0313C8	0.9999	Z221MPL7MAR03ENV2.1	0.9861
Z201FCA07feb03DNA13G10	1	Z221MPL7MAR03ENV4.2	0.9195
Z201FCA13C8_plasmid_2iii	0.9999	Z221MPL7MAR03ENV6.4	0.8982
Z201FCADNA13G10_plasmid_6i	1	Z238FCA15C6_plasmid_1v	0.8263
Z201FCF07feb03DNA13C18	1	Z238FCA29oct0215C6	0.8369
Z201FCF07feb03DNA13G13	1	Z238FCF15A39_plasmid_9ii	1
Z201FCF07feb03DNA13H13	0.9999	Z238FCF15C13_plasmid_2ii	1
Z201FCF07feb03DNA13H9	0.4086	Z238FCF29oct0215A39	1
Z201FCFDNA13C18_plasmid_3ii	1	Z238FCF29oct0215C13	1
Z201FCFDNA13G13_plasmid_7i	1	Z238FPL12_plasmid_1-2	1
Z201FCFDNA13H13_plasmid_10i	0.9991	Z238FPL16_plasmid_2-3	1
Z201FCFDNA13H9_plasmid_8v	0.3429	Z238FPL29nov0212	1
Z201FPB7FEB03ENV1.1	1	Z238FPL29nov0216	1
Z201FPB7FEB03ENV5.1	1	Z238FPL29nov024	1
Z201FPB7FEB03ENV6.1	0.9996	Z238FPL4_plasmid_6-1	1
Z201FPL07feb03100-1	1	Z238FSW15A11_plasmid_7ii	1

Three Class Neural Network

The evaluation of supervised methods was continued with a three output classifier in an attempt to identify transmitted, non-transmitted and recipient variations of gp120. Figure 4.45 provides the model construct of the neural network. Figure 4.46 provides a graphical representation of the neural network. This method was trained with augmented data in similar fashion as the binary classifier with similar accuracy characteristics and displayed similar issues with fitting (data not shown).

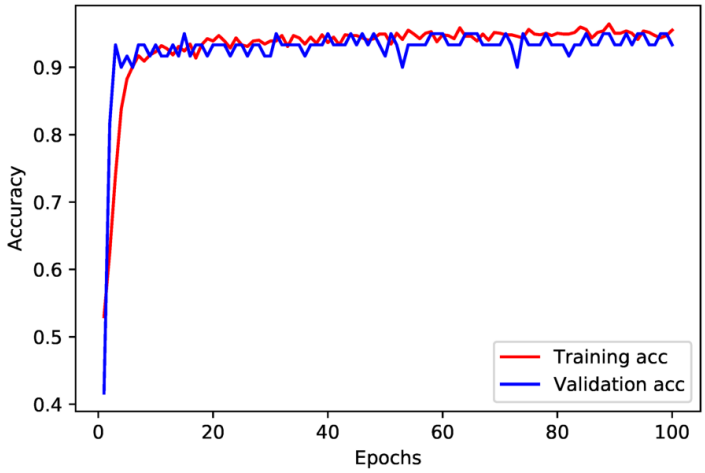

Figure 4.44: Graph show training and validation accuracy for the binary classifier in Figure 4.41.

```
model = tf.keras.Sequential([
        tf.keras.layers.Dense(128,input_dim=61,
                              kernel_regularizer=tf.keras.regularizers.
    l2(0.001),
                              activation='relu'),
        tf.keras.layers.Dropout(0.2),
        tf.keras.layers.Dense(128,
                              kernel_regularizer=tf.keras.regularizers.
    l2(0.001),
                              activation='relu'),
        tf.keras.layers.Dense(3, activation='softmax')])

model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

history = model.fit(x_train, y_train,
                    epochs=1000, batch_size=50,
                    validation_data=(x_val, y_val),
                    verbose=1)
```

Figure 4.45: Model construct of a three class neural network using 61 inputs tied to a 128 node hidden layer that feeds into a three output layer.

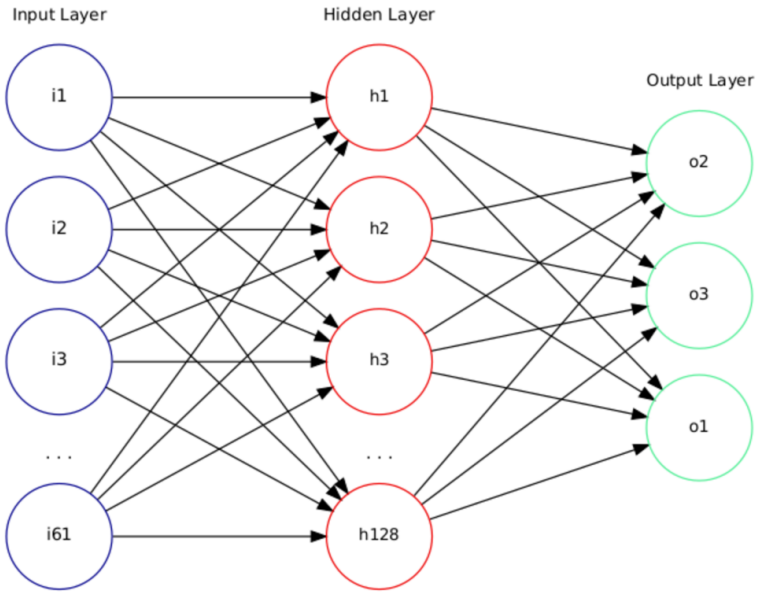

Figure 4.46: Graphic representation of three state classifier showing individual layers. Dropout layers are not expressed and are only used during training to control over and under fitting.

Discussion

Comparing BESI to artificial neural networks designed to classify ESSC pipeline data in relation to a control variant produced results showing fitting issues with the network models in both classifier strategies. Potentially, a compromise could be made by separating the data out by clade to investigate the sensitivity neural networks may have to variations across primary HIV strains. This would require known transmitted variants for each clade to be identified and processed. Additionally, data augmentation may own a portion of fault with fitting issues for both networks in that the augmentation is extracted form a single source, also requiring more identified transmitted variants.

Comparison of BESI Scores to Variable Loop Lengths

In 2004, Derdeyn et al. observed and suggested a correlation that early recipient variants of gp120 had shorter V1 through V4 loop lengths than was statistically predicted

(Derdeyn et al., 2004). In (Morton et al., 2019) a comparison of BESI against variable loop lengths was performed against a subset of sequences from (Morton et al., 2017). BESI is a clustering method, therefore the analysis of loop lengths in comparison to BESI would suggest that clusters of sequences with similar scores would gather around the vicinity of a control variant on a Cartesian coordinate system. The original data indicated potential correlation in variable loops 2 and 5 as is shown in Figures 4.48 and 4.51. Variable loops 1, 3, and 4 indicate no potential correlation as shown in Figures 4.47, 4.49, and 4.50. The results from this analysis indicated the need to perform the comparison against a larger population of gp120.

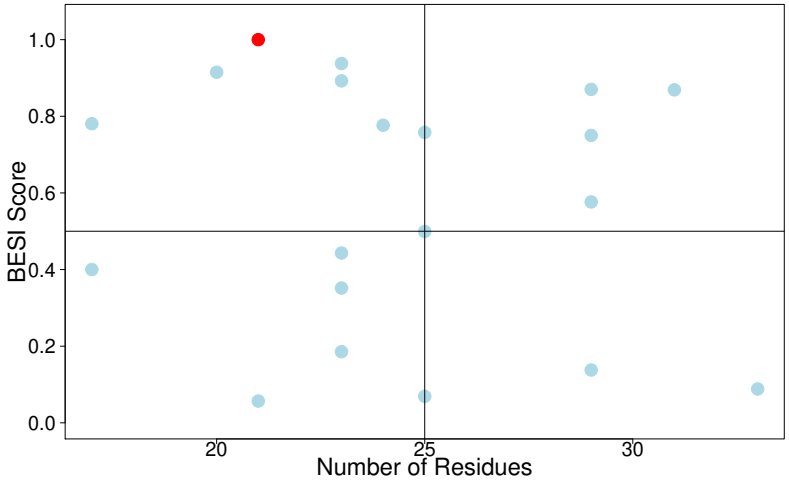

Figure 4.47: BESI control (red) versus Variable loop 1 length and score.

Results

The same comparative method is applied to the full set of sequences from (Morton et al., 2017) to produce Figures 4.52 through 4.56. These data indicate no correlation with observations made Derdeyn et al. in regards to BESI versus variable loop lengths.

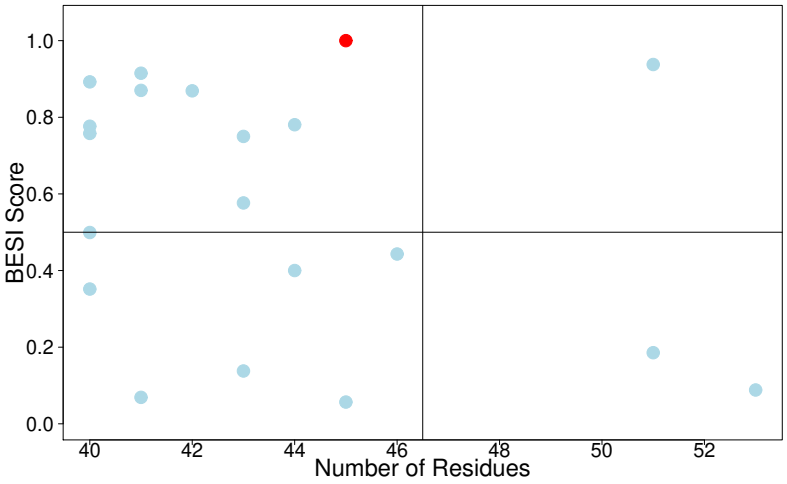

Figure 4.48: BESI control (red) versus Variable loop 2 length and score.

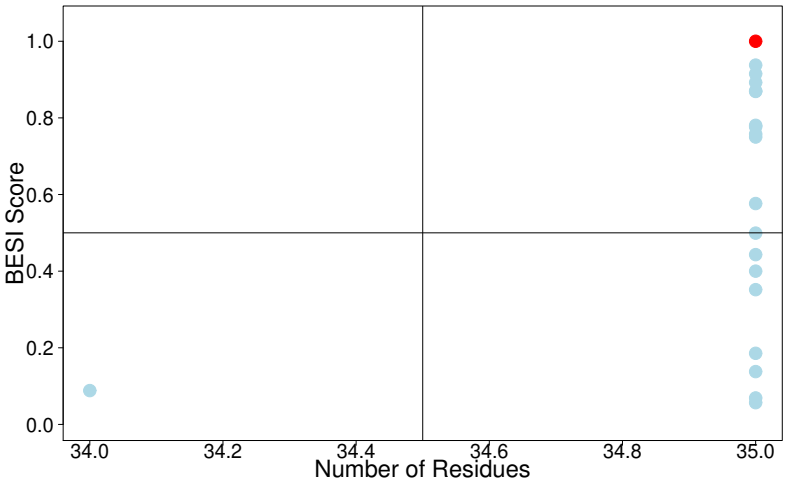

Figure 4.49: BESI control (red) versus Variable loop 3 length and score.

Discussion

Variable loop lengths provided no usable correlation with BESI as the scatter plots for all loops of 252 sequences had broad ranges of scores for each loop length. Based on these data, the conclusion is that variable loop lengths have no bearing on binding function that can be determined by BESI, but this does imply that variable loop lengths do not play a role in binding efficacy generally.

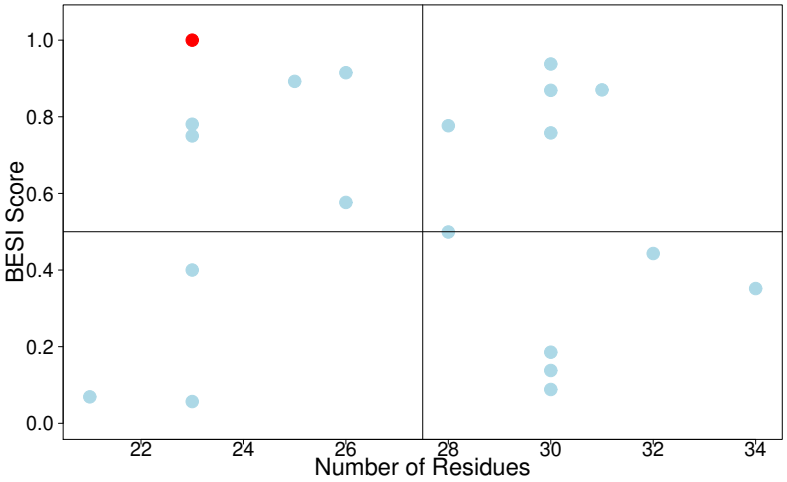

Figure 4.50: BESI control (red) versus Variable loop 4 length and score.

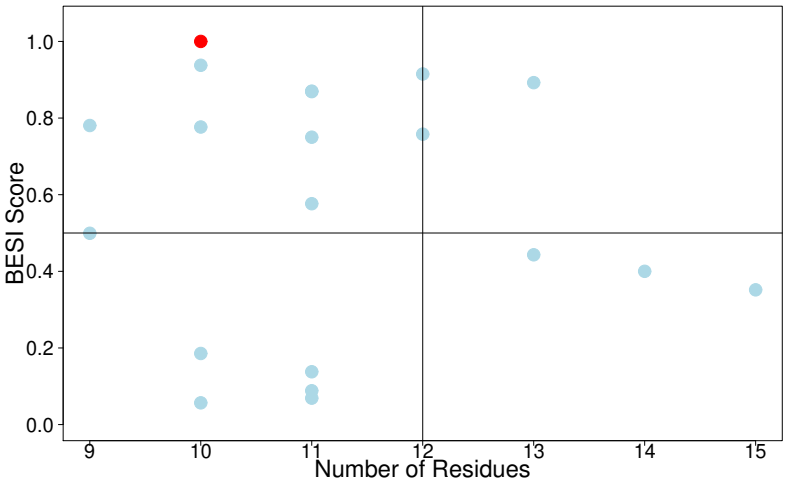

Figure 4.51: BESI control (red) versus Variable loop 5 length and score.

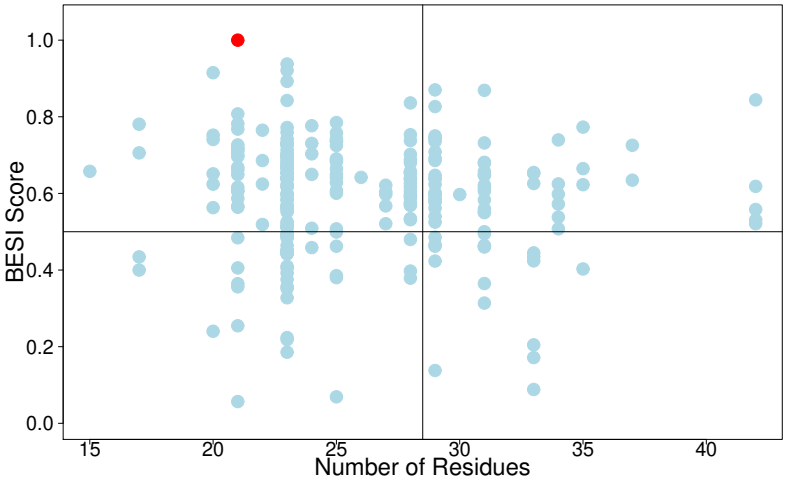

Figure 4.52: BESI control (red) versus Variable loop 1 length and score for all sequences.

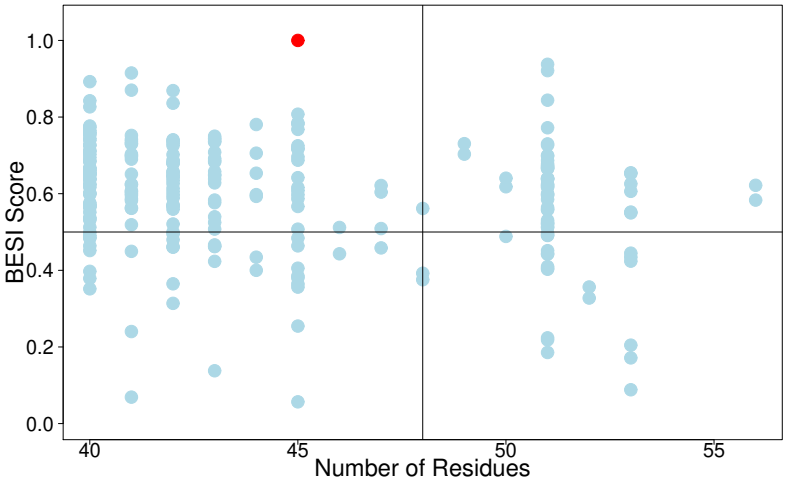

Figure 4.53: BESI control (red) versus Variable loop 2 length and score for all sequences.

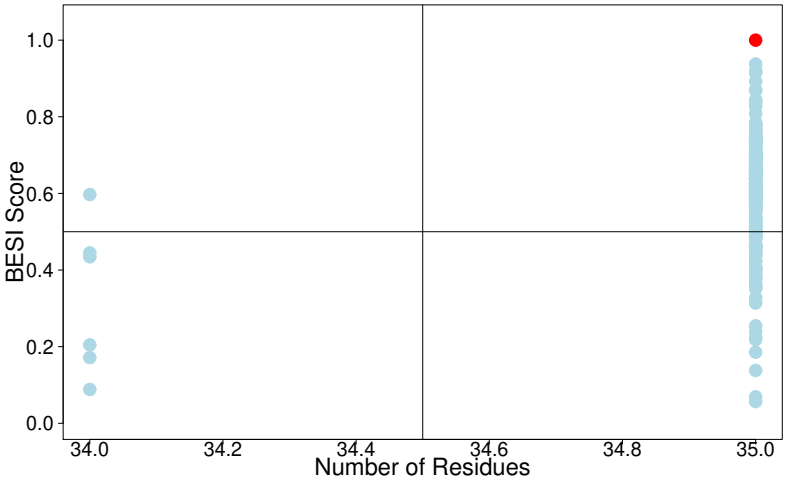

Figure 4.54: BESI control (red) versus Variable loop 3 length and score for all sequences.

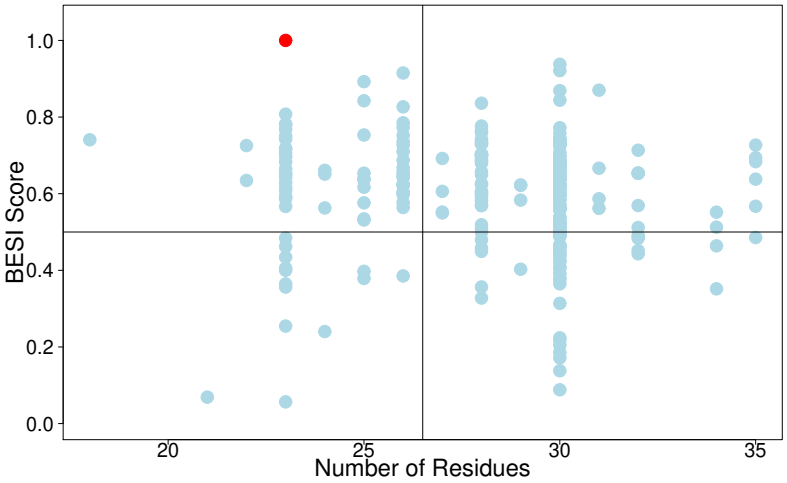

Figure 4.55: BESI control (red) versus Variable loop 4 length and score for all sequences.

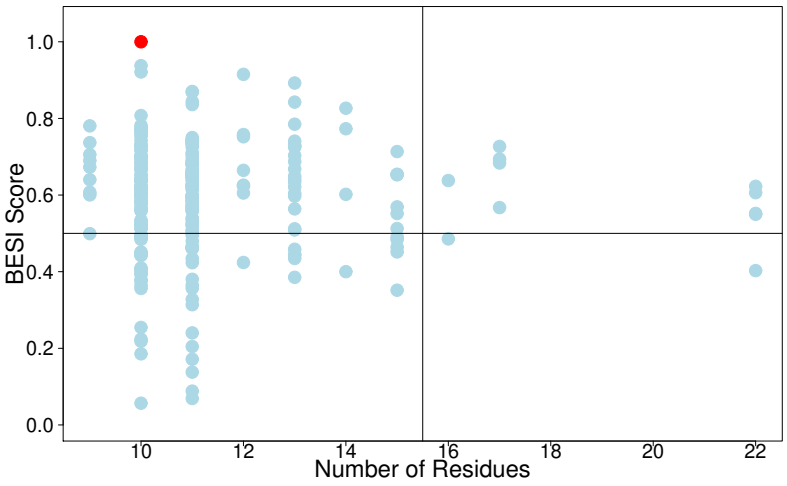

Figure 4.56: BESI control (red) versus Variable loop 5 length and score for all sequences.

CHAPTER V : DISCUSSIONS

BESI predicts a cyclical nature of gp120 variations indicating that variants of HIV cycle through the ability to cross the transmission barrier in genital tract mucosa, barring inflammation or open sores. EVM suggest that sequence variations of residues outside the CD4 binding site are the primary mechanism for modulating the potential transmission rate of the virus since residues outside of the conserved binding site are primarily where sequence differences arise. Allosteric interactions must drive much of the process in this case, but BESI cannot determine this definitively since it only works on the static endpoints of the binding process. It might be possible to model these transitions using molecular dynamics or Frodan, but the increase in APBS calculations needed to sample the intermediate states is still computationally prohibitive with current software.

Binding energy data suggests that gp120 interactions with CD4 are predicted to function below pH 5.0 in a small fraction of gp120/CD4 interactions and easily into the low pH 5.0 range for all interactions evaluated. This is in contrast to predicted working ranges of gp120/bnAb interactions that predict a functional range down to pH 5.6. These results suggest that a potential vaccine solution would be to engineer bnAbs that have a gp120 to bnAb BE motif similar to that of gp120 to CD4 motifs. These data further suggest that research across larger gp120/bnAb interactions are required to substantiate these observations and validate the computational predictions.

The comparative methods evaluated against BESI (supervised learning and variable loop lengths) provided no useful correlation with BESI. The use of artificial neural networks needs to be further investigated to determine any potential value; these particular instances were initial exploratory methods to determine feasibility of process.

Fauci et al. suggest that “enormous intellectual leaps beyond present day knowledge” are required to design a vaccine that blocks HIV infection, but continues to suggests that “laboratory, non-human primate testing and clinical research” are required to do so (Fauci

et al., 2008). While the first statement is true, the solution lies in computational methods capable of evaluating interactions in broader and more expedient experiments and use these results to guide laboratory experimentation in a more effective manner, this is the intellectual value provided by the methods of analysis presented in this dissertation.

BIBLIOGRAPHY

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Others (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Abrahams, M.-R., Anderson, J. A., Giorgi, E. E., Seoighe, C., Mlisana, K., Ping, L.-H., Athreya, G. S., Treurnicht, F. K., Keele, B. F., Wood, N., Salazar-Gonzalez, J. F., Bhattacharya, T., Chu, H., Hoffman, I., Galvin, S., Mapanje, C., Kazembe, P., Thebus, R., Fiscus, S., Hide, W., Cohen, M. S., Karim, S. A., Haynes, B. F., Shaw, G. M., Hahn, B. H., Korber, B. T., Swanstrom, R., Williamson, C., CAPRISA Acute Infection Study Team, Center for HIV-AIDS Vaccine Immunology Consortium, CAPRISA Acute Infection Study Team, f. t. C. A. I. S. T., the Center for HIV-AIDS Vaccine Immunology, and Center for HIV-AIDS Vaccine Immunology Consortium (2009). Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *Journal of Virology*, 83(8):3556–3567.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037–10041.

Bar, K. J., Sneller, M. C., Harrison, L. J., Justement, J. S., Overton, E. T., Petrone, M. E., Salantes, D. B., Seamon, C. A., Scheinfeld, B., Kwan, R. W., Learn, G. H., Proschan, M. A., Kreider, E. F., Blazkova, J., Bardsley, M., Refsland, E. W., Messer, M., Clarridge, K. E., Tustin, N. B., Madden, P. J., Oden, K., O'Dell, S. J., Jarocki, B., Shiakolas, A. R., Tressler, R. L., Doria-Rose, N. A., Bailer, R. T., Ledgerwood, J. E., Capparelli, E. V., Lynch, R. M., Graham, B. S., Moir, S., Koup, R. A., Mascola, J. R., Hoxie, J. A., Fauci, A. S., Tebas, P., and Chun, T.-W. (2016). Effect of HIV antibody VRC01 on viral rebound after treatment interruption. *New England Journal of Medicine*, 375(21):2037–2050.

Bar, K. J., Tsao, C.-y., Iyer, S. S., Decker, J. M., Yang, Y., Bonsignori, M., Chen, X., Hwang, K.-K., Montefiori, D. C., Liao, H.-X., Hraber, P., Fischer, W., Li, H., Wang, S., Sterrett, S., Keele, B. F., Ganusov, V. V., Perelson, A. S., Korber, B. T., Georgiev, I., McLellan, J. S., Pavlicek, J. W., Gao, F., Haynes, B. F., Hahn, B. H., Kwong, P. D., and Shaw, G. M. (2012). Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. *Public Library of Science Pathogens*, 8(5):e1002721.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. (2011). Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39.

Berendsen, H. J., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, 10(12):980–980.

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Bette T. Korber, Brian T. Foley, Carla L. Kuiken, Satish K. Pillai, and Joseph G. Sodroski (2017). HXB2 annotated spreadsheet. Technical report, Los Alamos National Laboratory.

Boeras, D. I., Hraber, P. T., Hurlston, M., Evans-Strickfaden, T., Bhattacharya, T., Giorgi, E. E., Mulenga, J., Karita, E., Korber, B. T., Allen, S., Hart, C. E., Derdeyn, C. a., and Hunter, E. (2011). Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proceedings of the National Academy of Sciences*, 108(46):E1156–E1163.

Bunnik, E. M., Pisas, L., van Nuenen, A. C., and Schuitemaker, H. (2008). Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. *Journal of Virology*, 82(16):7932–7941.

Cao, J., Sullivan, N., Desjardin, E., Parolin, C., Robinson, J., Wyatt, R., and Sodroski, J. (1997). Replication and neutralization of human immunodeficiency virus type 1 lacking the V1 and V2 variable loops of the gp120 envelope glycoprotein. *Journal of Virology*, 71(12):9808–12.

Carlson, J. M., Schaefer, M., Monaco, D. C., Batorsky, R., Claiborne, D. T., Prince, J., Deymier, M. J., Ende, Z. S., Klatt, N. R., DeZiel, C. E., Lin, T.-H., Peng, J., Seese, A. M., Shapiro, R., Frater, J., Ndung'u, T., Tang, J., Goepfert, P., Gilmour, J., Price, M. A., Kilembe, W., Heckerman, D., Goulder, P. J. R., Allen, T. M., Allen, S., and Hunter, E. (2014). HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science (New York, N.Y.)*, 345(6193):1254031.

Caskey, M., Klein, F., Lorenzi, J. C. C., Seaman, M. S., West, A. P., Buckley, N., Kremer, G., Nogueira, L., Braunschweig, M., Scheid, J. F., Horwitz, J. A., Shimeliovich, I., Ben-Avraham, S., Witmer-Pack, M., Platten, M., Lehmann, C., Burke, L. A., Hawthorne, T., Gorelick, R. J., Walker, B. D., Keler, T., Gulick, R. M., Fätkenheuer, G., Schlesinger, S. J., and Nussenzweig, M. C. (2015). Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature*, 522(7557):487–491.

Cavanagh, J. (2007). *Protein NMR Spectroscopy : Principles and Practice*. Academic Press.

Chen, B., Vogan, E. M., Gong, H., Skehel, J. J., Wiley, D. C., and Harrison, S. C. (2005). Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature*, 433(7028):834–841.

Chollet, F. (2015). Keras.

Dacheux, L., Moreau, A., Ataman-Onal, Y., Biron, F., Verrier, B., and Barin, F. (2004). Evolutionary dynamics of the glycan shield of the human immunodeficiency virus envelope during natural infection and implications for exposure of the 2G12 epitope. *Journal of Virology*, 78(22):12625–37.

Deerwester, S., Dumais, S. T., Harshman, R., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Symantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Derdeyn, C. C. A., Decker, J. M. J., Bibollet-Ruche, F., Mokili, J. L. J., Muldoon, M., Denham, S. A. S., Heil, M. L. M., Kasolo, F., Musonda, R., Hahn, B. B. H., Shaw, G. M. G., Korber, B. T., Allen, S., and Hunter, E. (2004). Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*, 303(5666):2019–2022.

Diskin, R., Marcovecchio, P. M., and Bjorkman, P. J. (2010). Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. *Nature Structural & Molecular Biology*, 17(5):608–13.

Doria-Rose, N. A., Schramm, C. A., Gorman, J., Moore, P. L., Bhiman, J. N., DeKosky, B. J., Ernandes, M. J., Georgiev, I. S., Kim, H. J., Pancera, M., Staupe, R. P., Altae-Tran, H. R., Bailer, R. T., Crooks, E. T., Cupo, A., Druz, A., Garrett, N. J., Hoi, K. H., Kong, R., Louder, M. K., Longo, N. S., McKee, K., Nonyane, M., O'Dell, S., Roark, R. S., Rudicell, R. S., Schmidt, S. D., Sheward, D. J., Soto, C., Wibmer, C. K., Yang, Y., Zhang, Z., Mullikin, J. C., Binley, J. M., Sanders, R. W., Wilson, I. A., Moore, J. P., Ward, A. B., Georgiou, G., Williamson, C., Abdool Karim, S. S., Morris, L., Kwong, P. D., Shapiro, L., and Mascola, J. R. (2014). Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*, 509(7498):55–62.

Drenth, J. and Mesters, J. (2007). *Principles of Protein X-Ray*. Springer.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. (2002). Comparative protein structure modeling using modeller. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Farrell, D. W., Speranskiy, K., and Thorpe, M. F. (2010). Generating stereochemically acceptable protein pathways. *Proteins: Structure, Function and Bioinformatics*, 78(14):2908–2921.

Fauci, A. S. (2016). An HIV vaccine. *The Journal of the American Medical Association*, 316(2):143.

Fauci, A. S., Johnston, M. I., Dieffenbach, C. W., Burton, D. R., Hammer, S. M., Hoxie, J. A., Martin, M., Overbaugh, J., Watkins, D. I., Mahmoud, A., and Greene, W. C. (2008). HIV vaccine research: the way forward. *Science*, 321(5888):530–532.

Fischer, W., Perkins, S., Theiler, J., Bhattacharya, T., Yusim, K., Funkhouser, R., Kuiken, C., Haynes, B., Letvin, N. L., Walker, B. D., Hahn, B. H., and Korber, B. T. (2007). Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature medicine*, 13(1):100–6.

Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert Review of Proteomics*, 1(1):97–111.

Foley, B. T., Leitner, T. K., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., Wolinsky, S., and Korber, B. T. M. (2015). HIV Sequence Compendium 2015.

Frank, J. J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies : Visualization of Biological Molecules in Their Native State*. Oxford University Press.

Haaland, R. E., Hawkins, P. A., Salazar-Gonzalez, J., Johnson, A., Tichacek, A., Karita, E., Manigart, O., Mulenga, J., Keele, B. F., Shaw, G. M., Hahn, B. H., Allen, S. A., Derdeyn, C. A., and Hunter, E. (2009). Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *Public Library of Science Pathogens*, 5(1):e1000274.

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. (2012). Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1):17.

Haynes, B. F. and Mascola, J. R. (2017). The quest for an antibody-based HIV vaccine. *Immunological Reviews*, 275(1):5–10.

Haynes, B. F., Shaw, G. M., Korber, B., Kelsoe, G., Sodroski, J., Hahn, B. H., Borrow, P., and McMichael, A. J. (2016). HIV-Host interactions: Implications for vaccine design. *Cell Host & Microbe*, 19(3):292–303.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Howton, J. (2017). A computational electrostratic modeling pipeline for comparing pH-dependent gp120-CD4 interactions in founder and chronic HIV strains.

Howton, J. and Phillips, J. L. (2017). Computational modeling of pH-dependent gp120-CD4 interactions in founder and chronic HIV strains. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*, pages 644–649, Boston, MA, USA. ACM Press.

Huang, C.-c., Tang, M., Zhang, M.-Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I. A., Wyatt, R., and Kwong, P. D. (2005). Structure of a V3-containing HIV-1 gp120 core. *Science (New York, N.Y.)*, 310(5750):1025–8.

Huang, C.-c., Venturi, M., Majeed, S., Moore, M. J., Phogat, S., Zhang, M.-Y., Dimitrov, D. S., Hendrickson, W. A., Robinson, J., Sodroski, J., Wyatt, R., Choe, H., Farzan, M., and Kwong, P. D. (2004). Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proceedings of the National Academy of Sciences*, 101(9):2706–2711.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38.

ISO/IEC JTC 1/SC 22 (2017). ISO - ISO/IEC 21778:2017 - Information technology — The JSON data interchange syntax.

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G.-W., Holst, M. J., McCammon, J. A., and Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J. G., Fidler, S., Weber, J., Pybus, O. G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P. R., Walker, B. D., Takiguchi, M., and Goulder, P. (2009). Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–5.

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., Wei, X., Jiang, C., Kirchherr, J. L., Gao, F., Anderson, J. A., Ping, L.-H., Swanstrom, R., Tomaras, G. D., Blattner, W. A., Goepfert, P. A., Kilby, J. M., Saag, M. S., Delwart, E. L., Busch, M. P., Cohen, M. S., Montefiori, D. C., Haynes, B. F., Gaschen, B., Athreya, G. S., Lee, H. Y., Wood, N., Seoighe, C., Perelson, A. S., Bhattacharya, T., Korber, B. T., Hahn, B. H., and Shaw, G. M. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7552–7557.

Klatzmann, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., Gluckman, J. C., and Montagnier, L. (1984). T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature*, 312(5996):767–768.

Korber-Irrgang, B., Foley, B., Kuiken, C., Pillai, S., Sodroski, J., Sodroski, J., Foley, B., Leitner, T., Pillai, S., and Sodroski, J. (1998). Numbering positions in HIV relative to HXB2CG.

Kothe, D. L., Li, Y., Decker, J. M., Bibollet-Ruche, F., Zammit, K. P., Salazar, M. G., Chen, Y., Weng, Z., Weaver, E. A., Gao, F., Haynes, B. F., Shaw, G. M., Korber, B. T., and Hahn, B. H. (2006). Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology*, 352(2):438–449.

Kwon, Y. D., Chuang, G.-Y., Zhang, B., Bailer, R. T., Doria-Rose, N. A., Gindin, T. S., Lin, B., Louder, M. K., McKee, K., O'Dell, S., Pegu, A., Schmidt, S. D., Asokan, M., Chen, X., Choe, M., Georgiev, I. S., Jin, V., Pancera, M., Rawi, R., Wang, K., Chaudhuri, R., Kueltzo, L. A., Manceva, S. D., Todd, J.-P., Scorpio, D. G., Kim, M., Reinherz, E. L., Wagh, K., Korber, B. M., Connors, M., Shapiro, L., Mascola, J. R., and Kwong, P. D. (2018). Surface-matrix screening identifies semi-specific interactions that improve potency of a near pan-reactive HIV-1-neutralizing antibody. *Cell Reports*, 22(7):1798–1809.

Kwong, P. D., Wyatt, R., Majeed, S., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (2000). Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure (London, England : 1993)*, 8(12):1329–39.

LaBranche, C. C., Henderson, R., Hsu, A., Behrens, S., Chen, X., Zhou, T., Wiehe, K., Saunders, K. O., Alam, S. M., Bonsignori, M., Borgnia, M. J., Sattentau, Q. J., Eaton, A., Greene, K., Gao, H., Liao, H. X., Williams, W. B., Peacock, J., Tang, H., Perez, L. G., Edwards, R. J., Kepler, T. B., Korber, B. T., Kwong, P. D., Mascola, J. R., Acharya, P., Haynes, B. F., and Montefiori, D. C. (2019). Neutralization-guided design of HIV-1 envelope trimers with high affinity for the unmutated common ancestor of CH235 lineage CD4bs broadly neutralizing antibodies. *Public Library of Science Pathogens*, 15(9):e1008026.

LANL (2020). HIV sequence database main page.

Li, B., Decker, J. M., Johnson, R. W., Bibollet-Ruche, F., Wei, X., Mulenga, J., Allen, S., Hunter, E., Hahn, B. H., Shaw, G. M., Blackwell, J. L., and Derdeyn, C. A. (2006a). Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. *Journal of Virology*, 80(11):5211–8.

Li, B., Decker, J. M., Johnson, R. W., Bibollet-Ruche, F., Wei, X., Mulenga, J., Allen, S., Hunter, E., Hahn, B. H., Shaw, G. M., Blackwell, J. L., and Derdeyn, C. A. (2006b). Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. *Journal of Virology*, 80(11):5211–8.

Li, M., Gao, F., Mascola, J. R., Stamatatos, L., Polonis, V. R., Koutsoukos, M., Voss, G., Goepfert, P., Gilbert, P., Greene, K. M., Bilska, M., Kothe, D. L., Salazar-Gonzalez, J. F., Wei, X., Decker, J. M., Hahn, B. H., and Montefiori, D. C. (2005). Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *Journal of Virology*, 79(16):10108–25.

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*, 7(8):306–317.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):NA–NA.

Lindstrom, P. (2014). Fixed-Rate Compressed Floating-Point Arrays. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2674–2683.

Linux Containers (2008). Linux Containers.

Liu, M. K., Hawkins, N., Ritchie, A. J., Ganusov, V. V., Whale, V., Brackenridge, S., Li, H., Pavlicek, J. W., Cai, F., Rose-Abrahams, M., Treurnicht, F., Hraber, P., Riou, C., Gray, C., Ferrari, G., Tanner, R., Ping, L.-H., Anderson, J. A., Swanstrom, R., B, C. C., Cohen, M., Karim, S. S. A., Haynes, B., Borrow, P., Perelson, A. S., Shaw, G. M., Hahn, B. H., Williamson, C., Korber, B. T., Gao, F., Self, S., McMichael, A., and Goonetilleke, N. (2012). Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *Journal of Clinical Investigation*.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Šali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):291–325.

Mascola, J. R. and Haynes, B. F. (2013). HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunological Reviews*, 254(1):225–244.

Mason, A. C. and Jensen, J. H. (2008). Protein-protein binding is often associated with changes in protonation state. *Proteins: Structure, Function and Genetics*, 71(1):81–91.

McLellan, J. S., Pancera, M., Carrico, C., Gorman, J., Julien, J.-P., Khayat, R., Louder, R., Pejchal, R., Sastry, M., Dai, K., O'Dell, S., Patel, N., Shahzad-ul Hussan, S., Yang, Y., Zhang, B., Zhou, T., Zhu, J., Boyington, J. C., Chuang, G.-Y., Diwanji, D., Georgiev, I., Do Kwon, Y., Lee, D., Louder, M. K., Moquin, S., Schmidt, S. D., Yang, Z.-Y., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Burton, D. R., Koff, W. C., Walker, L. M., Phogat, S., Wyatt, R., Orwenyo, J., Wang, L.-X., Arthos, J., Bewley, C. A., Mascola, J. R., Nabel, G. J., Schief, W. R., Ward, A. B., Wilson, I. A., and Kwong, P. D. (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, 480(7377):336–343.

MCrooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). *WebLogo: A sequence logo generator*.

Mehrishi, J. N. and Bauer, J. (2002). Electrophoresis of cells and the biological relevance of surface charge. *ELECTROPHORESIS*, 23(13):1984.

Misura, K. M. S. and Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, 59(1):15–29.

Misura, K. M. S., Chivian, D., Rohl, C. A., Kim, D. E., and Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of the National Academy of Sciences of the United States of America*, 103(14):5361–6.

Montefiori, D. C. (2009). Measuring HIV neutralization in a luciferase reporter gene assay. In *HIV Protocols: Second Edition, Methods in Molecular Virology*, pages 395–405. Humana Press.

Morton, S. P., Barbosa, S., Butler, R., and Pettey, C. (2016). A JSON-based markup language for deploying virtual clusters via Docker. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications PDPTA'16*, pages 251–257. CSREA Press ©.

Morton, S. P., Howton, J., and Phillips, J. L. (2018). Sub-Class differences of pH-dependent HIV gp120-CD4 interactions. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, pages 663–668, New York, New York, USA. ACM Press.

Morton, S. P., Phillips, J. B., and Phillips, J. L. (2017). High-throughput structural modeling of the HIV transmission bottleneck. In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine - BIBM-HPCB '17*, volume 2017-Janua, Kansas City, MO, USA. IEEE Press.

Morton, S. P., Phillips, J. B., and Phillips, J. L. (2019). The molecular basis of pH-modulated HIV gp120 binding revealed. *Evolutionary Bioinformatics*, 15:117693431983130.

Mouquet, H., Scharf, L., Euler, Z., Liu, Y., Eden, C., Scheid, J. F., Halper-Stromberg, A., Gnanapragasam, P. N. P., Spencer, D. I. R., Seaman, M. S., Schuitemaker, H., Feizi, T., Nussenzweig, M. C., and Bjorkman, P. J. (2012). Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47):E3268–77.

Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I., and Kosakovsky Pond, S. L. (2007). HIV-specific probabilistic models of protein evolution. *Public Library of Science ONE*, 2(6).

Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation*, 7(2):525–37.

Pancera, M., Majeed, S., Ban, Y.-E. A., Chen, L., Huang, C.-c., Kong, L., Kwon, Y. D., Stuckey, J., Zhou, T., Robinson, J. E., Schief, W. R., Sodroski, J., Wyatt, R., and Kwong, P. D. (2010). Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):1166–71.

Pantophlet, R. and Burton, D. R. (2006). GP120: Target for neutralizing HIV-1 antibodies. *Annual Review of Immunology*, 24(1):739–769.

Paradis, E., Claude, J., and Strimmer, K. (2004). {APE}: Analyses of phylogenetics and evolution in {R} language. *Bioinformatics*, 20(2):289–290.

Parrish, N. F., Gao, F., Li, H., Giorgi, E. E., Barbian, H. J., Parrish, E. H., Zajic, L., Iyer, S. S., Decker, J. M., Kumar, A., Hora, B., Berg, A., Cai, F., Hopper, J., Denny, T. N., Ding, H., Ochsenbauer, C., Kappes, J. C., Galimidi, R. P., West, A. P., Bjorkman, P. J., Wilen, C. B., Doms, R. W., O'Brien, M., Bhardwaj, N., Borrow, P., Haynes, B. F., Muldoon, M., Theiler, J. P., Korber, B., Shaw, G. M., Hahn, B. H., and Hahn, B. H. (2013). Phenotypic properties of transmitted founder HIV-1. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6626–33.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–8.

Petrey, D. and Honig, B. (2005). Protein Structure Prediction: Inroads to Biology. *Molecular Cell*, 20(6):811–819.

Richmond, D. and Fisher, D. (1973). The Electrophoretic Mobility of Micro-Organisms. *Advances in Microbial Physiology*, 9:1–29.

Rodrigues, R. and Druschel, P. (2010). Docker: lightweight Linux containers for consistent development and deployment. *Communications of the ACM*, 53(10):72–82.

Rong, R., Li, B., Lynch, R. M., Haaland, R. E., Murphy, M. K., Mulenga, J., Allen, S. A., Pinter, A., Shaw, G. M., Hunter, E., Robinson, J. E., Gnanakaran, S., and Derdeyn, C. A. (2009). Escape from autologous neutralizing antibodies in acute/early subtype C HIV-1 infection requires multiple pathways. *Public Library of Science Pathogens*, 5(9):e1000594.

Salazar-Gonzalez, J. F., Salazar, M. G., Keele, B. F., Learn, G. H., Giorgi, E. E., Li, H., Decker, J. M., Wang, S., Baalwa, J., Kraus, M. H., Parrish, N. F., Shaw, K. S., Guffey, M. B., Bar, K. J., Davis, K. L., Ochsenbauer-Jambor, C., Kappes, J. C., Saag, M. S., Cohen, M. S., Mulenga, J., Derdeyn, C. A., Allen, S., Hunter, E., Markowitz, M., Hraber, P., Perelson, A. S., Bhattacharya, T., Haynes, B. F., Korber, B. T., Hahn, B. H., and Shaw, G. M. (2009). Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *The Journal of Experimental Medicine*, 206(6):1273–1289.

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815.

Scheid, J. F., Horwitz, J. A., Bar-On, Y., Kreider, E. F., Lu, C.-L., Lorenzi, J. C. C., Feldmann, A., Braunschweig, M., Nogueira, L., Oliveira, T., Shimeliovich, I., Patel, R., Burke, L., Cohen, Y. Z., Hadrigan, S., Settler, A., Witmer-Pack, M., West, A. P., Juelg, B., Keler, T., Hawthorne, T., Zingman, B., Gulick, R. M., Pfeifer, N., Learn, G. H., Seaman, M. S., Bjorkman, P. J., Klein, F., Schlesinger, S. J., Walker, B. D., Hahn, B. H., Nussenzweig, M. C., and Caskey, M. (2016). HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature*, 535(7613):556–560.

Schoofs, T., Klein, F., Braunschweig, M., Kreider, E. F., Feldmann, A., Nogueira, L., Oliveira, T., Lorenzi, J. C. C., Parrish, E. H., Learn, G. H., West, A. P., Bjorkman, P. J., Schlesinger, S. J., Seaman, M. S., Czartoski, J., McElrath, M. J., Pfeifer, N., Hahn, B. H., Caskey, M., and Nussenzweig, M. C. (2016). HIV-1 therapy with monoclonal antibody 3BNC117 elicits host immune responses against HIV-1. *Science*, 352(6288):997–1001.

Schultz, A. M. and Bradac, J. A. (2001). The HIV vaccine pipeline, from preclinical to phase III. *AIDS*, 15.

Shen, M.-y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524.

Stamatakis, A. (2014). {RAxML} version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stieh, D. J., Phillips, J. L., Rogers, P. M., King, D. F., Cianci, G. C., Jeffs, S. A., Gnanakaran, S., and Shattock, R. J. (2013). Dynamic electrophoretic fingerprinting of the HIV-1 envelope glycoprotein. *Retrovirology*, 10(1):33.

Sullivan, N., Sun, Y., Sattentau, Q., Thali, M., Wu, D., Denisova, G., Gershoni, J., Robinson, J., Moore, J., and Sodroski, J. (1998). CD4-Induced conformational changes in the human immunodeficiency virus type 1 gp120 glycoprotein: consequences for virus entry and neutralization. *Journal of Virology*, 72(6):4694–703.

Trask, S. A., Derdeyn, C. A., Fideli, U., Chen, Y., Meleth, S., Kasolo, F., Musonda, R., Hunter, E., Gao, F., Allen, S., and Hahn, B. H. (2002). Molecular epidemiology of human immunodeficiency virus Type 1 transmission in a heterosexual cohort of discordant couples in Zambia. *Journal of Virology*, 76(1):397–405.

Turnbull, E. L., Wong, M., Wang, S., Wei, X., Jones, N. A., Conrod, K. E., Aldam, D., Turner, J., Pellegrino, P., Keele, B. F., Williams, I., Shaw, G. M., and Borrow, P. (2009). Kinetics of expansion of epitope-specific T cell responses during primary HIV-1 infection. *Journal of Immunology (Baltimore, Md. : 1950)*, 182(11):7131–7145.

Van Rossum, G. and Drake Jr, F. L. (1995). Python reference manual.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, A., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., Salazar-Gonzalez, J. F., Salazar, M. G., Kilby, J. M., Saag, M. S., Komarova, N. L., Nowak, M. A., Hahn, B. H., Kwong, P. D., and Shaw, G. M. (2003). Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312.

Wyatt, R. (1998). The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, 280(5371):1884–1888.

Wyatt, R., Moore, J., Accola, M., Desjardin, E., Robinson, J., and Sodroski, J. (1995). Involvement of the V1/V2 variable loop structure in the exposure of human immunodeficiency virus type 1 gp120 epitopes induced by receptor binding. *Journal of Virology*, 69(9):5723–33.

Wyatt, R., Sullivan, N., Thali, M., Repke, H., Ho, D., Robinson, J., Posner, M., and Sodroski, J. (1993). Functional and immunologic characterization of human immunodeficiency virus type 1 envelope glycoproteins containing deletions of the major variable regions. *Journal of Virology*, 67(8):4557–65.

Zhou, T., Xu, L., Dey, B., Hessell, A. J., Van Ryk, D., Xiang, S.-H., Yang, X., Zhang, M.-Y., Zwick, M. B., Arthos, J., Burton, D. R., Dimitrov, D. S., Sodroski, J., Wyatt, R., Nabel, G. J., and Kwong, P. D. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445(7129):732–737.

APPENDICES

APPENDIX A : BESI

Boeras et al.

Table A.1: Couple Z242 details sequence name, HIV Clade, donor/recipient classification and BESI scores. This set contains the control gp120 variant with a score of 1.

Sequence	Clade	Donor/Recipient	Score
Z242FPL25jan038_plasmid	C	R	0.7165
Z242FPL25JAN03PCR23ENV1.1	C	R	0.8074
Z242FPL25JAN03PCR8ENV1.1	C	R	0.7678
Z242MPL25jan0323_plasmid	C	D	0.6160
Z242MPL25jan0326_plasmid	C	D	0.2545
Z242MPL25JAN0326	C	D	0.4055
Z242MPL25JAN0327-1	C	D	0.4842
Z242MPL25JAN0327-2	C	D	0.6960
Z242MPL25JAN0327-3	C	D	0.6077
Z242MPL25jan0328_plasmid_8-1	C	D	0.7817
Z242MPL25jan0328_plasmid_8-2	C	D	0.3643
Z242MPL25jan0328_plasmid_8-3	C	D	0.3562
Z242MPL25jan0333_plasmid	C	D	0.6345
Z242MPL25JAN03PCR23ENV1.1-DT	C	D	1
Z242MPL25JAN03PCR33ENV1.1-DNT	C	D	0.7255
Z242MPL26_plasmid	C	D	0.0567
Z242MPL28_plasmid_8-1	C	D	0.7180
Z242MPL28_plasmid_8-2	C	D	0.5666
Z242MPL28_plasmid_8-3	C	D	0.5867

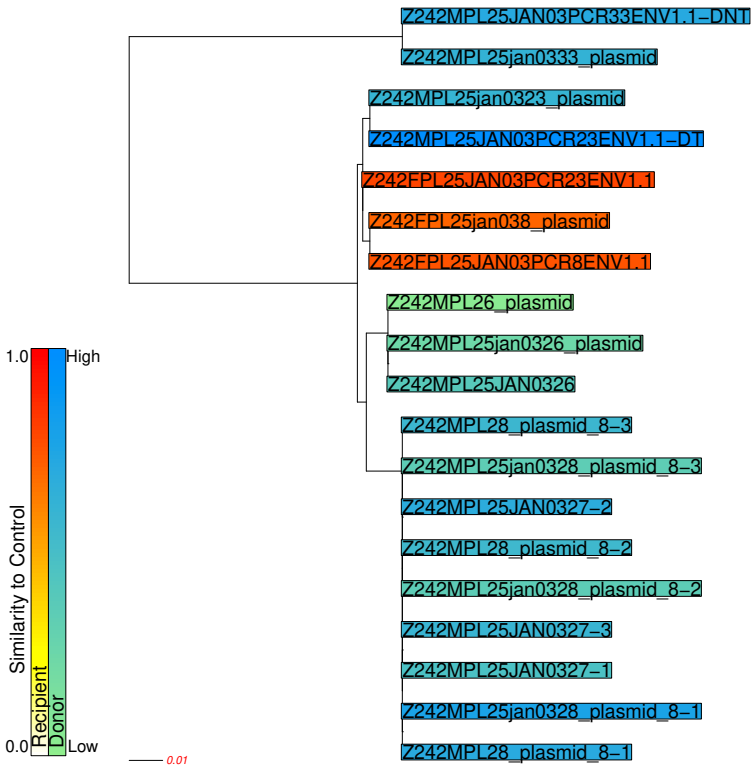

Figure A.1: BESI versus phylogenetic tree for couple Z242. This couple contains the control variant gp120 Z242MPL25JAN03PCR23ENV1.1-DT. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

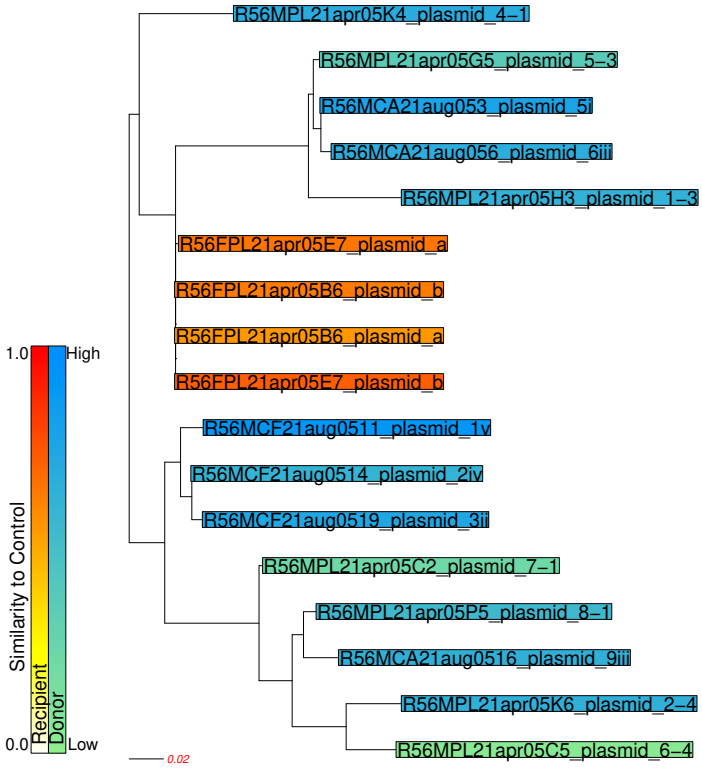

Figure A.2: BESI versus phylogenetic tree for couple R56. This tree requires the scores in Table A.2 and the knowledge that BESI typically includes the candidate gp120 in the top 3 scores. This information indicates R56MCA21aug053_plasmid_5i is the transmitted founder for this couple and is the second highest score returned for this donor. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.2: Couple R56 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
R56FPL21apr05B6_plasmid_a	A1	R	0.5639
R56FPL21apr05B6_plasmid_b	A1	R	0.6491
R56FPL21apr05E7_plasmid_a	A1	R	0.6681
R56FPL21apr05E7_plasmid_b	A1	R	0.7265
R56MCA21aug0516_plasmid_9iii	A1	D	0.6510
R56MCA21aug053_plasmid_5i	A1	D	0.7849
R56MCA21aug056_plasmid_6iii	A1	D	0.6874
R56MCF21aug0511_plasmid_1v	A1	D	0.9149
R56MCF21aug0514_plasmid_2iv	A1	D	0.6241
R56MCF21aug0519_plasmid_3ii	A1	D	0.7520
R56MPL21apr05C2_plasmid_7-1	A1	D	0.2402
R56MPL21apr05C5_plasmid_6-4	A1	D	0.0690
R56MPL21apr05G5_plasmid_5-3	A1	D	0.3853
R56MPL21apr05H3_plasmid_1-3	A1	D	0.6418
R56MPL21apr05K4_plasmid_4-1	A1	D	0.7099
R56MPL21apr05K6_plasmid_2-4	A1	D	0.6610
R56MPL21apr05P5_plasmid_8-1	A1	D	0.5628

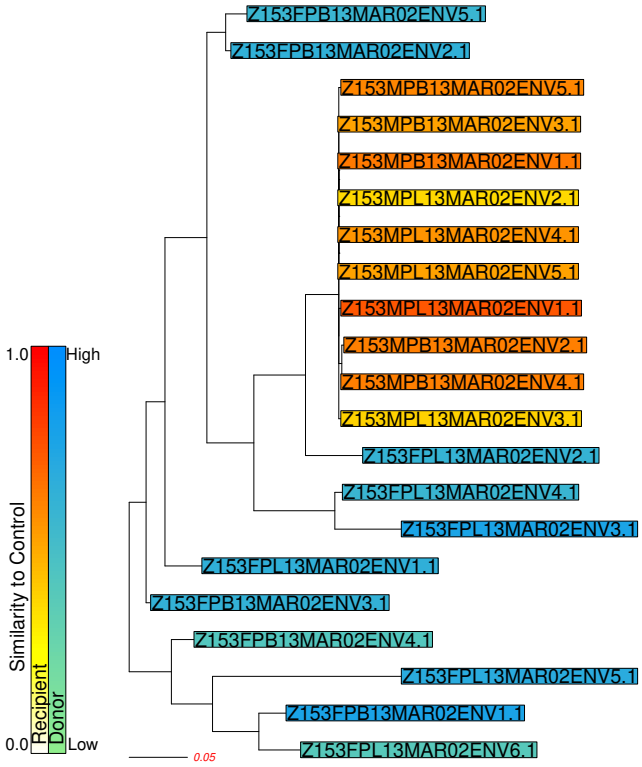

Figure A.3: Couple Z153 with the second highest donor BESI score (see Table A.3), Z153FPL13MAR02ENV3.1, being in the correct sub-tree as a candidate to cross the transmission barrier. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.3: Couple Z153 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z153FPB13MAR02ENV1.1	C	D	0.7805
Z153FPB13MAR02ENV2.1	C	D	0.6534
Z153FPB13MAR02ENV3.1	C	D	0.6418
Z153FPB13MAR02ENV4.1	C	D	0.4343
Z153FPB13MAR02ENV5.1	C	D	0.5929
Z153FPL13MAR02ENV1.1	C	D	0.6646
Z153FPL13MAR02ENV2.1	C	D	0.6235
Z153FPL13MAR02ENV3.1	C	D	0.7729
Z153FPL13MAR02ENV4.1	C	D	0.5975
Z153FPL13MAR02ENV5.1	C	D	0.7057
Z153FPL13MAR02ENV6.1	C	D	0.4000
Z153MPB13MAR02ENV1.1	C	R	0.6535
Z153MPB13MAR02ENV2.1	C	R	0.6395
Z153MPB13MAR02ENV3.1	C	R	0.5342
Z153MPB13MAR02ENV4.1	C	R	0.6366
Z153MPB13MAR02ENV5.1	C	R	0.6169
Z153MPL13MAR02ENV1.1	C	R	0.7532
Z153MPL13MAR02ENV2.1	C	R	0.3788
Z153MPL13MAR02ENV3.1	C	R	0.3973
Z153MPL13MAR02ENV4.1	C	R	0.5760
Z153MPL13MAR02ENV5.1	C	R	0.5313

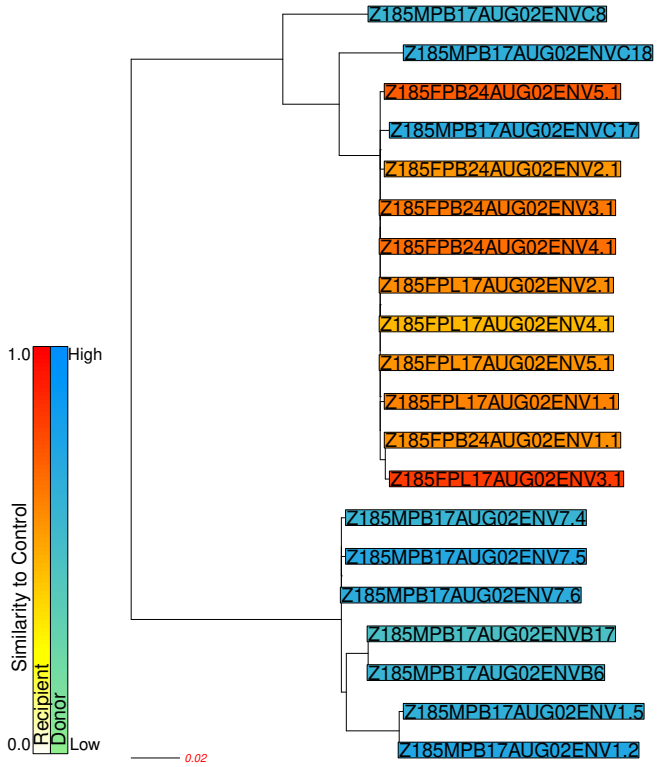

Figure A.4: BESI versus phylogenetics for couple Z185. This tree implies that the transmitting sequence is not included in this set or BESI fails at an undetermined level. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.4: Couple Z185 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z185FPB24AUG02ENV1.1	C	R	0.5886
Z185FPB24AUG02ENV2.1	C	R	0.5685
Z185FPB24AUG02ENV3.1	C	R	0.6572
Z185FPB24AUG02ENV4.1	C	R	0.6877
Z185FPB24AUG02ENV5.1	C	R	0.7378
Z185FPL17AUG02ENV1.1	C	R	0.6261
Z185FPL17AUG02ENV2.1	C	R	0.5967
Z185FPL17AUG02ENV3.1	C	R	0.8361
Z185FPL17AUG02ENV4.1	C	R	0.4797
Z185FPL17AUG02ENV5.1	C	R	0.5713
Z185MPB17AUG02ENV1.2	C	D	0.7578
Z185MPB17AUG02ENV1.5	C	D	0.6644
Z185MPB17AUG02ENV7.4	C	D	0.6073
Z185MPB17AUG02ENV7.5	C	D	0.7367
Z185MPB17AUG02ENV7.6	C	D	0.6901
Z185MPB17AUG02ENVB17	C	D	0.4992
Z185MPB17AUG02ENVB6	C	D	0.6004
Z185MPB17AUG02ENVC17	C	D	0.7019
Z185MPB17AUG02ENVC18	C	D	0.6798
Z185MPB17AUG02ENVC8	C	D	0.6053

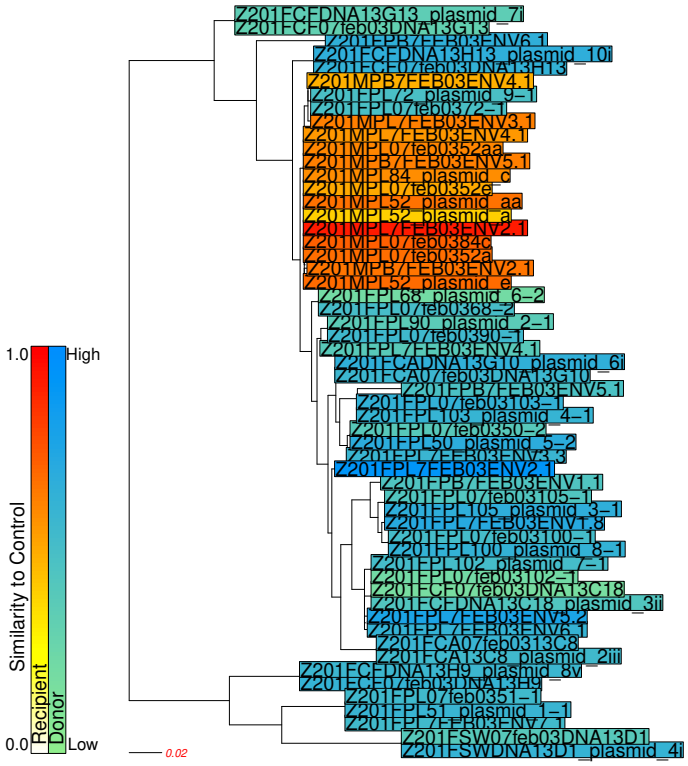

Figure A.5: BESI versus phylogenetic tree for couple Z201. BESI scores in this set indicates the actual donor sequence is not in the selected set from the 123 variations available as determined through (LANL, 2020). Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.5: Couple Z201 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score	Sequence	Clade	Donor/Recipient	Score
Z201FCA07feb0313C8	C	D	0.6002	Z201FPL50_plasmid_5-2	C	D	0.6983
Z201FCA07feb03DNA13G10	C	D	0.5850	Z201FPL51_plasmid_1-1	C	D	0.6061
Z201FCA13C8_plasmid_2iii	C	D	0.5584	Z201FPL68_plasmid_6-2	C	D	0.2184
Z201FCADNA13G10_plasmid_6i	C	D	0.6675	Z201FPL72_plasmid_9-1	C	D	0.4966
Z201FCF07feb03DNA13C18	C	D	0.1855	Z201FPL7FEB03ENV1.8	C	D	0.7299
Z201FCF07feb03DNA13G13	C	D	0.3276	Z201FPL7FEB03ENV2.1	C	D	0.9377
Z201FCF07feb03DNA13H13	C	D	0.6400	Z201FPL7FEB03ENV3.3	C	D	0.5970
Z201FCF07feb03DNA13H9	C	D	0.6218	Z201FPL7FEB03ENV4.1	C	D	0.3923
Z201FCFDNA13C18_plasmid_3ii	C	D	0.4423	Z201FPL7FEB03ENV5.2	C	D	0.7720
Z201FCFDNA13G13_plasmid_7i	C	D	0.3565	Z201FPL7FEB03ENV6.1	C	D	0.6239
Z201FCFDNA13H13_plasmid_10i	C	D	0.6728	Z201FPL7FEB03ENV7.1	C	D	0.5498
Z201FCFDNA13H9_plasmid_8v	C	D	0.5832	Z201FPL90_plasmid_2-1	C	D	0.3754
Z201FPB7FEB03ENV1.1	C	D	0.4512	Z201FSW07feb03DNA13D1	C	D	0.4027
Z201FPB7FEB03ENV5.1	C	D	0.4884	Z201FSWDNA13D1_plasmid_4i	C	D	0.6227
Z201FPB7FEB03ENV6.1	C	D	0.6664	Z201MPB7FEB03ENV2.1	C	R	0.6653
Z201FPL07feb03100-1	C	D	0.5254	Z201MPB7FEB03ENV4.1	C	R	0.4916
Z201FPL07feb03102-1	C	D	0.2237	Z201MPB7FEB03ENV5.1	C	R	0.6396
Z201FPL07feb03103-1	C	D	0.6404	Z201MPL07feb0352a	C	R	0.6864
Z201FPL07feb03105-1	C	D	0.4429	Z201MPL07feb0352aa	C	R	0.6169
Z201FPL07feb0350-2	C	D	0.4097	Z201MPL07feb0352e	C	R	0.5116
Z201FPL07feb0351-1	C	D	0.5520	Z201MPL07feb0384c	C	R	0.7266
Z201FPL07feb0368-2	C	D	0.5132	Z201MPL52_plasmid_a	C	R	0.4055
Z201FPL07feb0372-1	C	D	0.4909	Z201MPL52_plasmid_aa	C	R	0.6714
Z201FPL07feb0390-1	C	D	0.5614	Z201MPL52_plasmid_e	C	R	0.6998
Z201FPL100_plasmid_8-1	C	D	0.6385	Z201MPL7FEB03ENV2.1	C	R	0.9211
Z201FPL102_plasmid_7-1	C	D	0.5214	Z201MPL7FEB03ENV3.1	C	R	0.6378
Z201FPL103_plasmid_4-1	C	D	0.6180	Z201MPL7FEB03ENV4.1	C	R	0.5640
Z201FPL105_plasmid_3-1	C	D	0.6777	Z201MPL84_plasmid_c	C	R	0.5964

Table A.6: Couple Z205 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z205FPB27MAR03ENV1.1	C	R	0.6282
Z205FPB27MAR03ENV4.2	C	R	0.4622
Z205FPL27MAR03ENV4.1	C	R	0.7423
Z205FPL27MAR03ENV5.2	C	R	0.6494
Z205FPL27MAR03ENV6.3	C	R	0.6827
Z205MPB27MAR03ENV4.1	C	D	0.6576
Z205MPB27MAR03ENV6.1	C	D	0.5764
Z205MPB27MAR03ENV9.1	C	D	0.7500
Z205MPL27MAR03ENV1.1NF	C	D	0.6472
Z205MPL27MAR03ENV2.3	C	D	0.7407
Z205MPL27MAR03ENV3.1NF	C	D	0.7357
Z205MPL27MAR03ENV6.3	C	D	0.6919

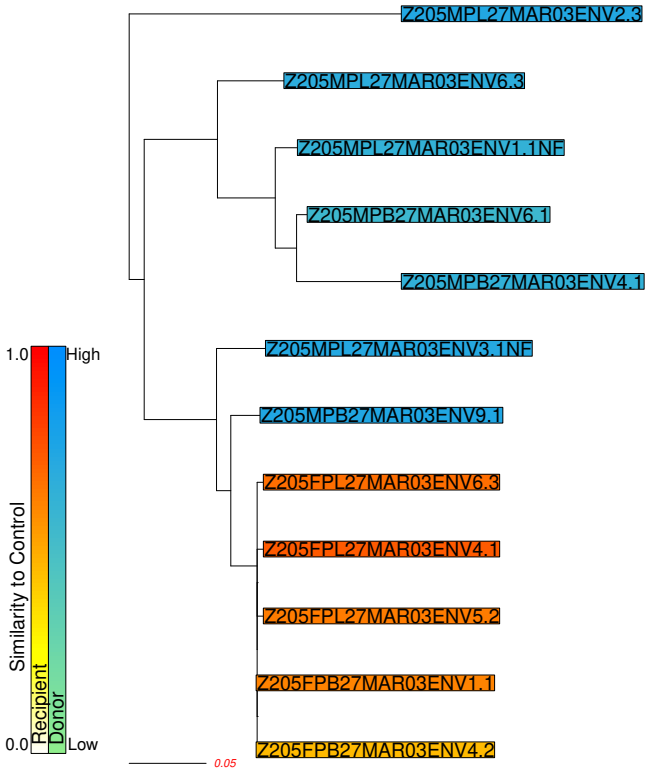

Figure A.6: BESI versus phylogenetic tree for couple Z205. BESI scores in this set indicate a potential match for the transmitted variant Z205MPB27MAR03ENV9.1 in that all recipient variations are shown as descendants. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

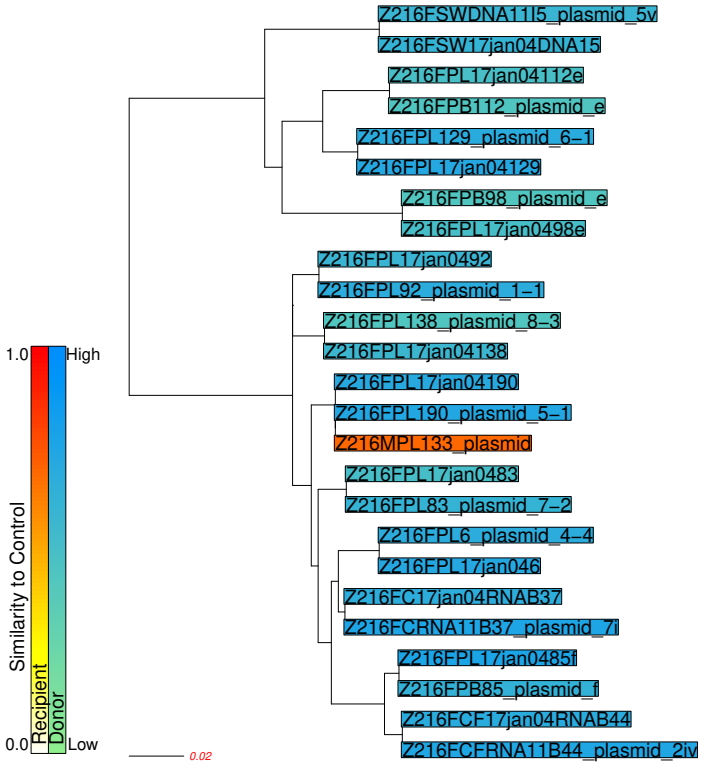

Figure A.7: BESI versus phylogenetic tree for couple Z216. BESI scores in this set indicate a potential that the transmitting variant is not included in the list of studied variants out of the 78 extracted from the patient (LANL, 2020). Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.7: Couple Z216 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z216FC17jan04RNAB37	C	D	0.6857
Z216FCF17jan04RNAB44	C	D	0.7025
Z216FCFRNA11B44_plasmid_2iv	C	D	0.7594
Z216FCRNA11B37_plasmid_7i	C	D	0.7649
Z216FPB112_plasmid_e	C	D	0.4585
Z216FPB85_plasmid_f	C	D	0.6496
Z216FPB98_plasmid_e	C	D	0.4431
Z216FPL129_plasmid_6-1	C	D	0.7031
Z216FPL138_plasmid_8-3	C	D	0.4495
Z216FPL17jan04112e	C	D	0.5090
Z216FPL17jan04129	C	D	0.7305
Z216FPL17jan04138	C	D	0.5819
Z216FPL17jan04190	C	D	0.7299
Z216FPL17jan046	C	D	0.7307
Z216FPL17jan0483	C	D	0.5189
Z216FPL17jan0485f	C	D	0.7766
Z216FPL17jan0492	C	D	0.6053
Z216FPL17jan0498e	C	D	0.5118
Z216FPL190_plasmid_5-1	C	D	0.7420
Z216FPL6_plasmid_4-4	C	D	0.6970
Z216FPL83_plasmid_7-2	C	D	0.6244
Z216FPL92_plasmid_1-1	C	D	0.6998
Z216FSW17jan04DNA15	C	D	0.6213
Z216FSWDNA11I5_plasmid_5v	C	D	0.6041
Z216MPL133_plasmid	C	R	0.7030

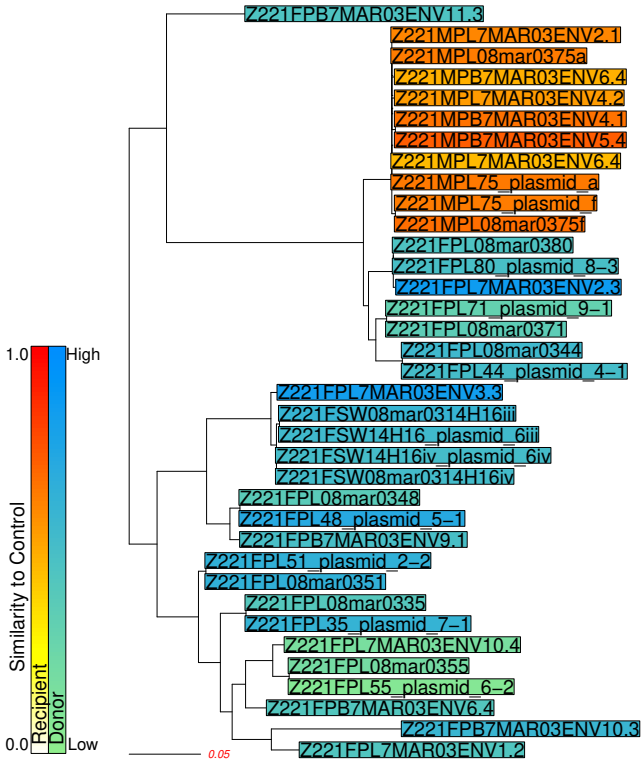

Figure A.8: BESI versus phylogenetic tree for couple Z221. BESI scores in this set indicate BESI performing poorly or potentially that the transmitting variant is not included in the list of studied variants out of the 75 extracted from the patient (LANL, 2020). Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.8: Couple Z221 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z221FPB7MAR03ENV10.3	C	D	0.5970
Z221FPB7MAR03ENV11.3	C	D	0.4638
Z221FPB7MAR03ENV6.4	C	D	0.4345
Z221FPB7MAR03ENV9.1	C	D	0.5074
Z221FPL08mar0335	C	D	0.4239
Z221FPL08mar0344	C	D	0.5674
Z221FPL08mar0348	C	D	0.3801
Z221FPL08mar0351	C	D	0.6545
Z221FPL08mar0355	C	D	0.2047
Z221FPL08mar0371	C	D	0.3647
Z221FPL08mar0380	C	D	0.4614
Z221FPL35_plasmid_7-1	C	D	0.6257
Z221FPL44_plasmid_4-1	C	D	0.5208
Z221FPL48_plasmid_5-1	C	D	0.7250
Z221FPL51_plasmid_2-2	C	D	0.6522
Z221FPL55_plasmid_6-2	C	D	0.0882
Z221FPL71_plasmid_9-1	C	D	0.3137
Z221FPL7MAR03ENV1.2	C	D	0.4447
Z221FPL7MAR03ENV10.4	C	D	0.1714
Z221FPL7MAR03ENV2.3	C	D	0.8690
Z221FPL7MAR03ENV3.3	C	D	0.8440
Z221FPL80_plasmid_8-3	C	D	0.5000
Z221FSW08mar0314H16iii	C	D	0.6183
Z221FSW08mar0314H16iv	C	D	0.5307
Z221FSW14H16_plasmid_6iii	C	D	0.5581
Z221FSW14H16iv_plasmid_6iv	C	D	0.5209
Z221MPB7MAR03ENV4.1	C	R	0.6787
Z221MPB7MAR03ENV5.4	C	R	0.7317
Z221MPB7MAR03ENV6.4	C	R	0.4948
Z221MPL08mar0375a	C	R	0.6477
Z221MPL08mar0375f	C	R	0.6811
Z221MPL75_plasmid_a	C	R	0.6570
Z221MPL75_plasmid_f	C	R	0.6478
Z221MPL7MAR03ENV2.1	C	R	0.6140
Z221MPL7MAR03ENV4.2	C	R	0.5595
Z221MPL7MAR03ENV6.4	C	R	0.4602

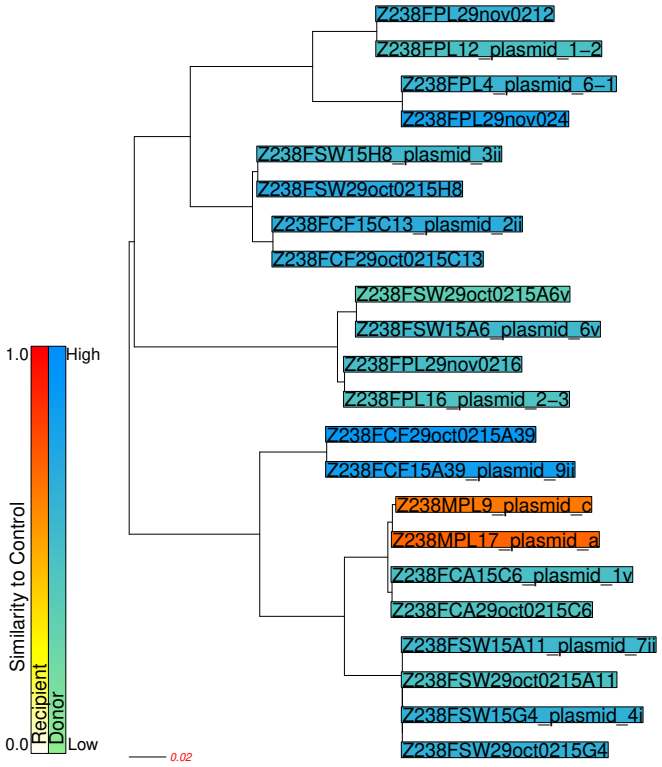

Figure A.9: BESI versus phylogenetic tree for couple Z238. BESI scores in this set indicate a solid hit from BESI where the two top scores are at the clade top containing recipient variations. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.9: Couple Z238 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z238FCA15C6_plasmid_1v	C	D	0.4906
Z238FCA29oct0215C6	C	D	0.4516
Z238FCF15A39_plasmid_9ii	C	D	0.8425
Z238FCF15C13_plasmid_2ii	C	D	0.6939
Z238FCF29oct0215A39	C	D	0.8924
Z238FCF29oct0215C13	C	D	0.6836
Z238FPL12_plasmid_1-2	C	D	0.4856
Z238FPL16_plasmid_2-3	C	D	0.4638
Z238FPL29nov0212	C	D	0.6379
Z238FPL29nov0216	C	D	0.5126
Z238FPL29nov024	C	D	0.8266
Z238FPL4_plasmid_6-1	C	D	0.6018
Z238FSW15A11_plasmid_7ii	C	D	0.5691
Z238FSW15A6_plasmid_6v	C	D	0.5517
Z238FSW15G4_plasmid_4i	C	D	0.6534
Z238FSW15H8_plasmid_3ii	C	D	0.5672
Z238FSW29oct0215A11	C	D	0.4837
Z238FSW29oct0215A6v	C	D	0.3516
Z238FSW29oct0215G4	C	D	0.6531
Z238FSW29oct0215H8	C	D	0.7268
Z238MPL17_plasmid_a	C	R	0.7134
Z238MPL9_plasmid_c	C	R	0.6540

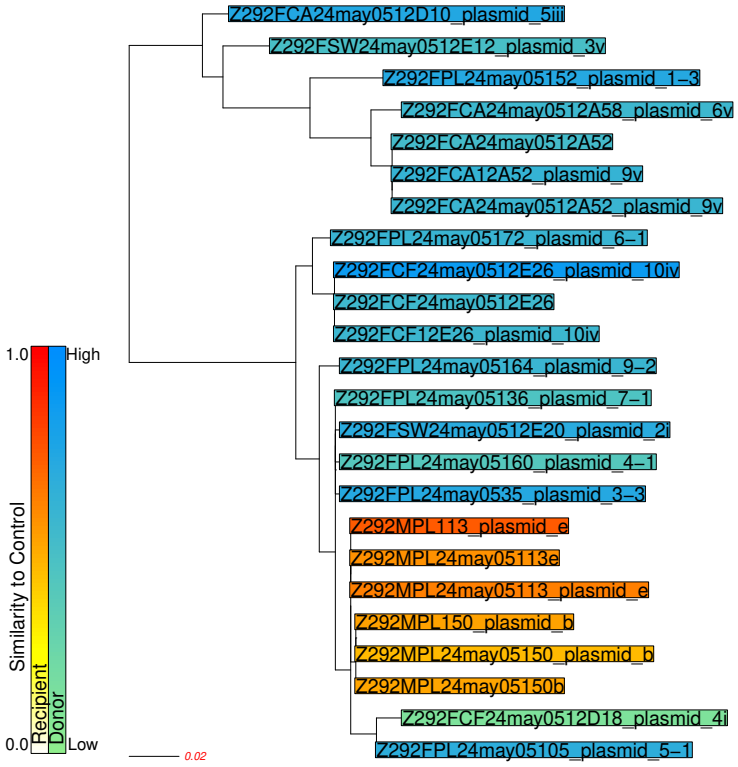

Figure A.10: BESI versus phylogenetic tree for couple Z292. BESI scores in this set indicate a solid hit from BESI where the top score is near the clade top containing recipient variations. Donor sequences are shaded from light green to blue and recipient sequences are shaded from white to red, lowest to highest similarity respectively in comparison to the control sequence.

Table A.10: Couple Z292 details sequence name, HIV Clade, donor/recipient classification and BESI scores.

Sequence	Clade	Donor/Recipient	Score
Z292FCA12A52_plasmid_9v	A1	D	0.6247
Z292FCA24may0512A52_plasmid_9v	A1	D	0.5981
Z292FCA24may0512A52	A1	D	0.5379
Z292FCA24may0512A58_plasmid_6v	A1	D	0.5721
Z292FCA24may0512D10_plasmid_5iii	A1	D	0.7436
Z292FCF12E26_plasmid_10iv	A1	D	0.5866
Z292FCF24may0512D18_plasmid_4i	A1	D	0.1378
Z292FCF24may0512E26_plasmid_10iv	A1	D	0.8701
Z292FCF24may0512E26	A1	D	0.5617
Z292FPL24may05105_plasmid_5-1	A1	D	0.6860
Z292FPL24may05136_plasmid_7-1	A1	D	0.4621
Z292FPL24may05152_plasmid_1-3	A1	D	0.7396
Z292FPL24may05160_plasmid_4-1	A1	D	0.4234
Z292FPL24may05164_plasmid_9-2	A1	D	0.5957
Z292FPL24may05172_plasmid_6-1	A1	D	0.6003
Z292FPL24may0535_plasmid_3-3	A1	D	0.7354
Z292FSW24may0512E12_plasmid_3v	A1	D	0.5076
Z292FSW24may0512E20_plasmid_2i	A1	D	0.7083
Z292MPL113_plasmid_e	A1	R	0.7443
Z292MPL150_plasmid_b	A1	R	0.5396
Z292MPL24may05113_plasmid_e	A1	R	0.6382
Z292MPL24may05113e	A1	R	0.5833
Z292MPL24may05150_plasmid_b	A1	R	0.4663
Z292MPL24may05150b	A1	R	0.5250

Complete PCA Reconstructions for Boeras et al.

Figure A.11: Comparison of original (left) and PCA reconstruction (right) for sequence R56FPL21apr05B6_plasmid_a.

Figure A.12: Comparison of original (left) and PCA reconstruction (right) for sequence R56FPL21apr05B6_plasmid_b.

Figure A.13: Comparison of original (left) and PCA reconstruction (right) for sequence R56FPL21apr05E7_plasmid_a.

Figure A.14: Comparison of original (left) and PCA reconstruction (right) for sequence R56FPL21apr05E7_plasmid_b.

Figure A.15: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCA21aug0516_plasmid_9iii.

Figure A.16: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCA21aug053_plasmid_5i.

Figure A.17: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCA21aug056_plasmid_6iii.

Figure A.18: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCF21aug0511_plasmid_1v.

Figure A.19: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCF21aug0514_plasmid_2iv.

Figure A.20: Comparison of original (left) and PCA reconstruction (right) for sequence R56MCF21aug0519_plasmid_3ii.

Figure A.21: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05C2_plasmid_7-1.

Figure A.22: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05C5_plasmid_6-4.

Figure A.23: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05G5_plasmid_5-3.

Figure A.24: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05H3_plasmid_1-3.

Figure A.25: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05K4_plasmid_4-1.

Figure A.26: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05K6_plasmid_2-4.

Figure A.27: Comparison of original (left) and PCA reconstruction (right) for sequence R56MPL21apr05P5_plasmid_8-1.

Figure A.28: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPB13MAR02ENV1.1.

Figure A.29: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPB13MAR02ENV2.1.

Figure A.30: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPB13MAR02ENV3.1.

Figure A.31: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPB13MAR02ENV4.1.

Figure A.32: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPB13MAR02ENV5.1.

Figure A.33: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV1.1.

Figure A.34: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV2.1.

Figure A.35: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV3.1.

Figure A.36: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV4.1.

Figure A.37: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV5.1.

Figure A.38: Comparison of original (left) and PCA reconstruction (right) for sequence Z153FPL13MAR02ENV6.1.

Figure A.39: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPB13MAR02ENV1.1.

Figure A.40: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPB13MAR02ENV2.1.

Figure A.41: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPB13MAR02ENV3.1.

Figure A.42: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPB13MAR02ENV4.1.

Figure A.43: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPB13MAR02ENV5.1.

Figure A.44: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPL13MAR02ENV1.1.

Figure A.45: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPL13MAR02ENV2.1.

Figure A.46: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPL13MAR02ENV3.1.

Figure A.47: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPL13MAR02ENV4.1.

Figure A.48: Comparison of original (left) and PCA reconstruction (right) for sequence Z153MPL13MAR02ENV5.1.

Figure A.49: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPB24AUG02ENV1.1.

Figure A.50: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPB24AUG02ENV2.1.

Figure A.51: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPB24AUG02ENV3.1.

Figure A.52: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPB24AUG02ENV4.1.

Figure A.53: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPB24AUG02ENV5.1.

Figure A.54: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPL17AUG02ENV1.1.

Figure A.55: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPL17AUG02ENV2.1.

Figure A.56: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPL17AUG02ENV3.1.

Figure A.57: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPL17AUG02ENV4.1.

Figure A.58: Comparison of original (left) and PCA reconstruction (right) for sequence Z185FPL17AUG02ENV5.1.

Figure A.59: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENV1.2.

Figure A.60: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENV1.5.

Figure A.61: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENV7.4.

Figure A.62: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENV7.5.

Figure A.63: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENV7.6.

Figure A.64: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENVB17.

Figure A.65: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENVB6.

Figure A.66: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENVC17.

Figure A.67: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENVC18.

Figure A.68: Comparison of original (left) and PCA reconstruction (right) for sequence Z185MPB17AUG02ENVC8.

Figure A.69: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCA07feb0313C8.

Figure A.70: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCA07feb03DNA13G10.

Figure A.71: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCA13C8_plasmid_2iii.

Figure A.72: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCADNA13G10_plasmid_6i.

Figure A.73: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCF07feb03DNA13C18.

Figure A.74: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCF07feb03DNA13G13.

Figure A.75: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCF07feb03DNA13H13.

Figure A.76: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCF07feb03DNA13H9.

Figure A.77: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCFDNA13C18_plasmid_3ii.

Figure A.78: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCFDNA13G13_plasmid_7i.

Figure A.79: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCFDNA13H13_plasmid_10i.

Figure A.80: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FCFDNA13H9_plasmid_8v.

Figure A.81: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPB7FEB03ENV1.1.

Figure A.82: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPB7FEB03ENV5.1.

Figure A.83: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPB7FEB03ENV6.1.

Figure A.84: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb03100-1.

Figure A.85: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb03102-1.

Figure A.86: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb03103-1.

Figure A.87: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb03105-1.

Figure A.88: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb0350-2.

Figure A.89: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb0351-1.

Figure A.90: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb0368-2.

Figure A.91: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb0372-1.

Figure A.92: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL07feb0390-1.

Figure A.93: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL100_plasmid_8-1.

Figure A.94: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL102_plasmid_7-1.

Figure A.95: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL103_plasmid_4-1.

Figure A.96: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL105_plasmid_3-1.

Figure A.97: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL50_plasmid_5-2.

Figure A.98: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL51_plasmid_1-1.

Figure A.99: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL68_plasmid_6-2.

Figure A.100: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL72_plasmid_9-1.

Figure A.101: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV1.8.

Figure A.102: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV2.1.

Figure A.103: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV3.3.

Figure A.104: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV4.1.

Figure A.105: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV5.2.

Figure A.106: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV6.1.

Figure A.107: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL7FEB03ENV7.1.

Figure A.108: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FPL90_plasmid_2-1.

Figure A.109: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FSW07feb03DNA13D1.

Figure A.110: Comparison of original (left) and PCA reconstruction (right) for sequence Z201FSWDNA13D1_plasmid_4i.

Figure A.111: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPB7FEB03ENV2.1.

Figure A.112: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPB7FEB03ENV4.1.

Figure A.113: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPB7FEB03ENV5.1.

Figure A.114: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL07feb0352a.

Figure A.115: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL07feb0352aa.

Figure A.116: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL07feb0352e.

Figure A.117: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL07feb0384c.

Figure A.118: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL52_plasmid_a.

Figure A.119: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL52_plasmid_aa.

Figure A.120: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL52_plasmid_e.

Figure A.121: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL7FEB03ENV2.1.

Figure A.122: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL7FEB03ENV3.1.

Figure A.123: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL7FEB03ENV4.1.

Figure A.124: Comparison of original (left) and PCA reconstruction (right) for sequence Z201MPL84_plasmid_c.

Figure A.125: Comparison of original (left) and PCA reconstruction (right) for sequence Z205FPB27MAR03ENV1.1.

Figure A.126: Comparison of original (left) and PCA reconstruction (right) for sequence Z205FPB27MAR03ENV4.2.

Figure A.127: Comparison of original (left) and PCA reconstruction (right) for sequence Z205FPL27MAR03ENV4.1.

Figure A.128: Comparison of original (left) and PCA reconstruction (right) for sequence Z205FPL27MAR03ENV5.2.

Figure A.129: Comparison of original (left) and PCA reconstruction (right) for sequence Z205FPL27MAR03ENV6.3.

Figure A.130: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPB27MAR03ENV4.1.

Figure A.131: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPB27MAR03ENV6.1.

Figure A.132: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPB27MAR03ENV9.1.

Figure A.133: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPL27MAR03ENV1.1NF.

Figure A.134: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPL27MAR03ENV2.3.

Figure A.135: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPL27MAR03ENV3.1NF.

Figure A.136: Comparison of original (left) and PCA reconstruction (right) for sequence Z205MPL27MAR03ENV6.3.

Figure A.137: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FC17jan04RNAB37.

Figure A.138: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FCF17jan04RNAB44.

Figure A.139: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FCFRNA11B44_plasmid_2iv.

Figure A.140: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FCRNA11B37_plasmid_7i.

Figure A.141: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPB112_plasmid_e.

Figure A.142: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPB85_plasmid_f.

Figure A.143: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPB98_plasmid_e.

Figure A.144: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL129_plasmid_6-1.

Figure A.145: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL138_plasmid_8-3.

Figure A.146: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan04112e.

Figure A.147: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan04129.

Figure A.148: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan04138.

Figure A.149: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan04190.

Figure A.150: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan046.

Figure A.151: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan0483.

Figure A.152: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan0485f.

Figure A.153: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan0492.

Figure A.154: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL17jan0498e.

Figure A.155: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL190_plasmid_5-1.

Figure A.156: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL6_plasmid_4-4.

Figure A.157: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL83_plasmid_7-2.

Figure A.158: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FPL92_plasmid_1-1.

Figure A.159: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FSW17jan04DNA15.

Figure A.160: Comparison of original (left) and PCA reconstruction (right) for sequence Z216FSWDNA11I5_plasmid_5v.

Figure A.161: Comparison of original (left) and PCA reconstruction (right) for sequence Z216MPL133_plasmid.

Figure A.162: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPB7MAR03ENV10.3.

Figure A.163: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPB7MAR03ENV11.3.

Figure A.164: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPB7MAR03ENV6.4.

Figure A.165: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPB7MAR03ENV9.1.

Figure A.166: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0335.

Figure A.167: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0344.

Figure A.168: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0348.

Figure A.169: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0351.

Figure A.170: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0355.

Figure A.171: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0371.

Figure A.172: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL08mar0380.

Figure A.173: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL35_plasmid_7-1.

Figure A.174: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL44_plasmid_4-1.

Figure A.175: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL48_plasmid_5-1.

Figure A.176: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL51_plasmid_2-2.

Figure A.177: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL55_plasmid_6-2.

Figure A.178: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL71_plasmid_9-1.

Figure A.179: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL7MAR03ENV1.2.

Figure A.180: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL7MAR03ENV10.4.

Figure A.181: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL7MAR03ENV2.3.

Figure A.182: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL7MAR03ENV3.3.

Figure A.183: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FPL80_plasmid_8-3.

Figure A.184: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FSW08mar0314H16iii.

Figure A.185: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FSW08mar0314H16iv.

Figure A.186: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FSW14H16_plasmid_6iii.

Figure A.187: Comparison of original (left) and PCA reconstruction (right) for sequence Z221FSW14H16iv_plasmid_6iv.

Figure A.188: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPB7MAR03ENV4.1.

Figure A.189: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPB7MAR03ENV5.4.

Figure A.190: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPB7MAR03ENV6.4.

Figure A.191: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL08mar0375a.

Figure A.192: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL08mar0375f.

Figure A.193: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL75_plasmid_a.

Figure A.194: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL75_plasmid_f.

Figure A.195: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL7MAR03ENV2.1.

Figure A.196: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL7MAR03ENV4.2.

Figure A.197: Comparison of original (left) and PCA reconstruction (right) for sequence Z221MPL7MAR03ENV6.4.

Figure A.198: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCA15C6_plasmid_1v.

Figure A.199: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCA29oct0215C6.

Figure A.200: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCF15A39_plasmid_9ii.

Figure A.201: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCF15C13_plasmid_2ii.

Figure A.202: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCF29oct0215A39.

Figure A.203: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FCF29oct0215C13.

Figure A.204: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL12_plasmid_1-2.

Figure A.205: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL16_plasmid_2-3.

Figure A.206: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL29nov0212.

Figure A.207: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL29nov0216.

Figure A.208: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL29nov024.

Figure A.209: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FPL4_plasmid_6-1.

Figure A.210: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW15A11_plasmid_7ii.

Figure A.211: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW15A6_plasmid_6v.

Figure A.212: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW15G4_plasmid_4i.

Figure A.213: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW15H8_plasmid_3ii.

Figure A.214: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW29oct0215A11.

Figure A.215: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW29oct0215A6v.

Figure A.216: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW29oct0215G4.

Figure A.217: Comparison of original (left) and PCA reconstruction (right) for sequence Z238FSW29oct0215H8.

Figure A.218: Comparison of original (left) and PCA reconstruction (right) for sequence Z238MPL17_plasmid_a.

Figure A.219: Comparison of original (left) and PCA reconstruction (right) for sequence Z238MPL9_plasmid_c.

Figure A.220: Comparison of original (left) and PCA reconstruction (right) for sequence Z242FPL25JAN03PCR23ENV1.1.

Figure A.221: Comparison of original (left) and PCA reconstruction (right) for sequence Z242FPL25JAN03PCR8ENV1.1.

Figure A.222: Comparison of original (left) and PCA reconstruction (right) for sequence Z242FPL25jan038_plasmid.

Figure A.223: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN0326.

Figure A.224: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN0327-1.

Figure A.225: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN0327-2.

Figure A.226: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN0327-3.

Figure A.227: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN03PCR23ENV1.1-DT.

Figure A.228: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25JAN03PCR33ENV1.1-DNT.

Figure A.229: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0323_plasmid.

Figure A.230: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0326_plasmid.

Figure A.231: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0328_plasmid_8-1.

Figure A.232: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0328_plasmid_8-2.

Figure A.233: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0328_plasmid_8-3.

Figure A.234: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL25jan0333_plasmid.

Figure A.235: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL26_plasmid.

Figure A.236: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL28_plasmid_8-1.

Figure A.237: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL28_plasmid_8-2.

Figure A.238: Comparison of original (left) and PCA reconstruction (right) for sequence Z242MPL28_plasmid_8-3.

Figure A.239: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCA12A52_plasmid_9v.

Figure A.240: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCA24may0512A52.

Figure A.241: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCA24may0512A52_plasmid_9v.

Figure A.242: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCA24may0512A58_plasmid_6v.

Figure A.243: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCA24may0512D10_plasmid_5iii.

Figure A.244: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCF12E26_plasmid_10iv.

Figure A.245: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCF24may0512D18_plasmid_4i.

Figure A.246: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCF24may0512E26.

Figure A.247: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FCF24may0512E26_plasmid_10iv.

Figure A.248: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05105_plasmid_5-1.

Figure A.249: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05136_plasmid_7-1.

Figure A.250: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05152_plasmid_1-3.

Figure A.251: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05160_plasmid_4-1.

Figure A.252: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05164_plasmid_9-2.

Figure A.253: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may05172_plasmid_6-1.

Figure A.254: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FPL24may0535_plasmid_3-3.

Figure A.255: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FSW24may0512E12_plasmid_3v.

Figure A.256: Comparison of original (left) and PCA reconstruction (right) for sequence Z292FSW24may0512E20_plasmid_2i.

Figure A.257: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL113_plasmid_e.

Figure A.258: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL150_plasmid_b.

Figure A.259: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL24may05113_plasmid_e.

Figure A.260: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL24may05113e.

Figure A.261: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL24may05150_plasmid_b.

Figure A.262: Comparison of original (left) and PCA reconstruction (right) for sequence Z292MPL24may05150b.

APPENDIX B : EVM

Complete EVM Results for Morton et al. (2018)

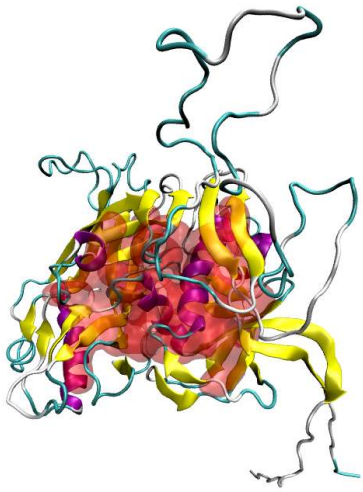

Figure B.1: EVM imagery displaying the selected residues for sequence 03_CH40TF in red.

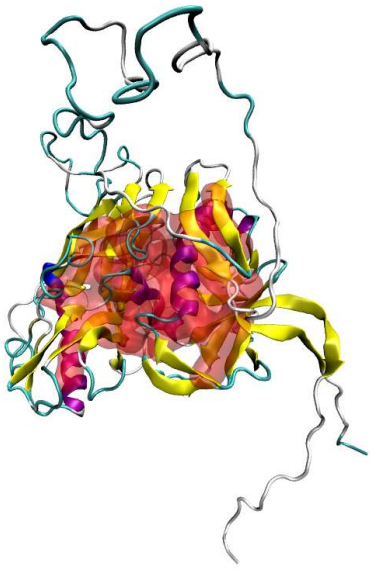

Figure B.2: EVM imagery displaying the selected residues for sequence 46_CH40M6 in red.

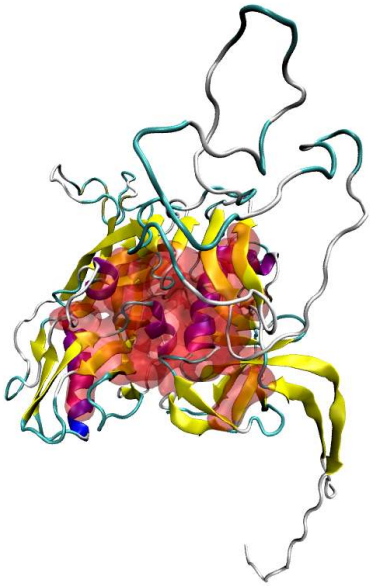

Figure B.3: EVM imagery displaying the selected residues for sequence 47_CH58TF in red.

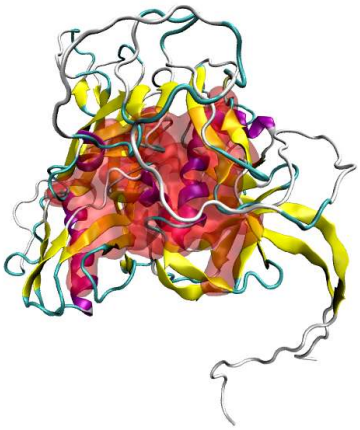

Figure B.4: EVM imagery displaying the selected residues for sequence 48_CH58M6 in red.

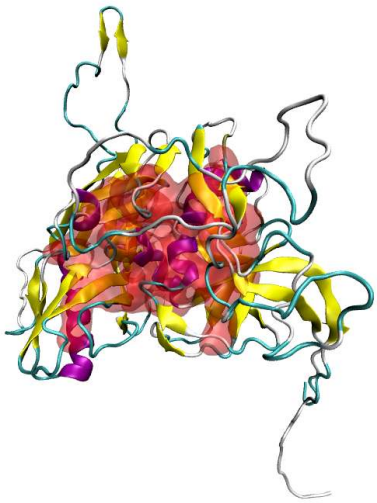

Figure B.5: EVM imagery displaying the selected residues for sequence 49_CH77TF in red.

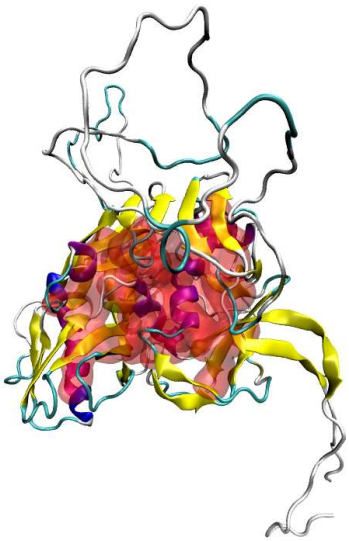

Figure B.6: EVM imagery displaying the selected residues for sequence 50_CH77M6 in red.

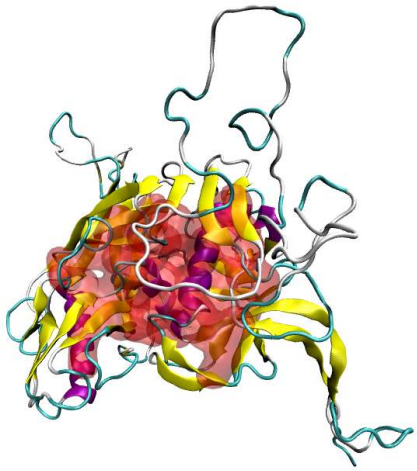

Figure B.7: EVM imagery displaying the selected residues for sequence 51_CH470TF in red.

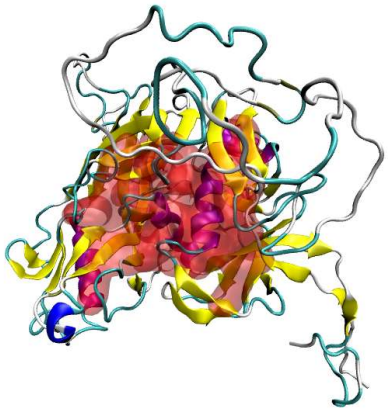

Figure B.8: EVM imagery displaying the selected residues for sequence 52_CH470M6 in red.

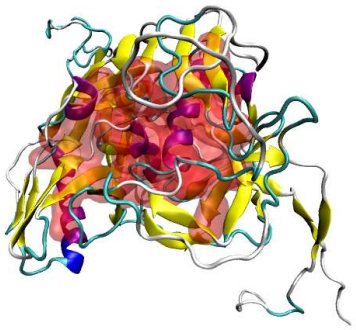

Figure B.9: EVM imagery displaying the selected residues for sequence 53_CH569TF in red.

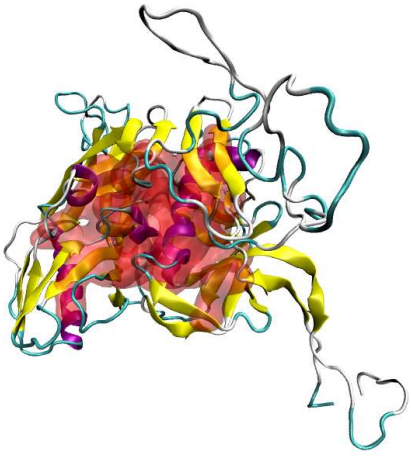

Figure B.10: EVM imagery displaying the selected residues for sequence 54_CH569M6 in red.

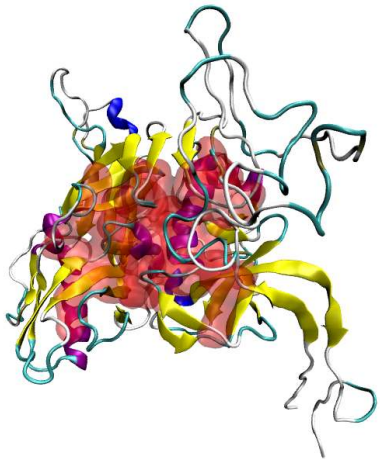

Figure B.11: EVM imagery displaying the selected residues for sequence 55_CH42TF in red.

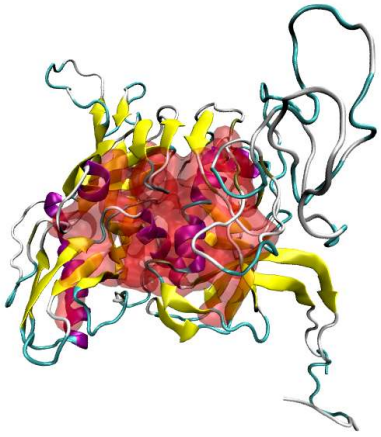

Figure B.12: EVM imagery displaying the selected residues for sequence 56_CH42M6 in red.

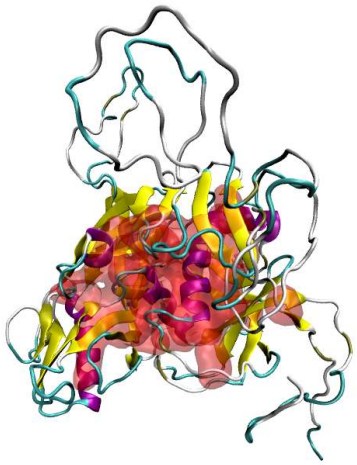

Figure B.13: EVM imagery displaying the selected residues for sequence 57_CH236TF in red.

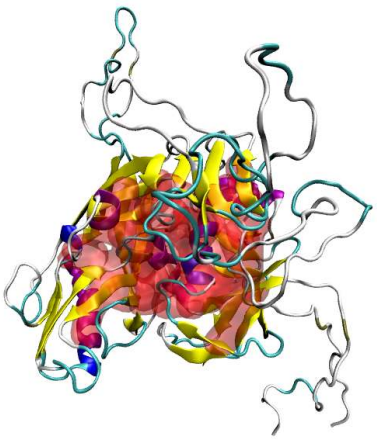

Figure B.14: EVM imagery displaying the selected residues for sequence 58_CH236M6 in red.

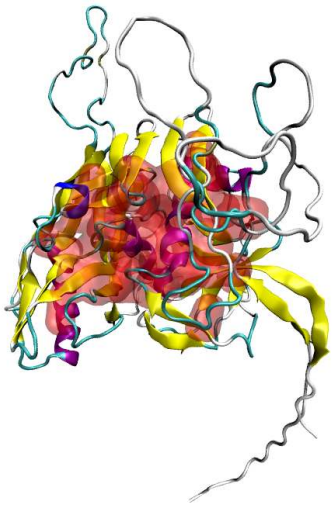

Figure B.15: EVM imagery displaying the selected residues for sequence 59_CH850TF in red.

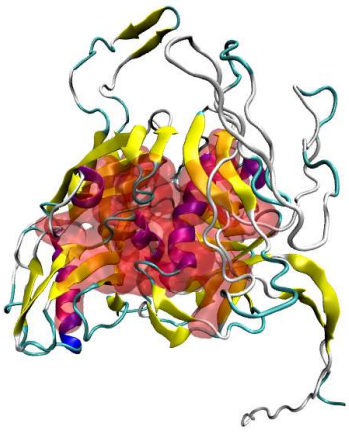

Figure B.16: EVM imagery displaying the selected residues for sequence 60_CH850M6 in red.

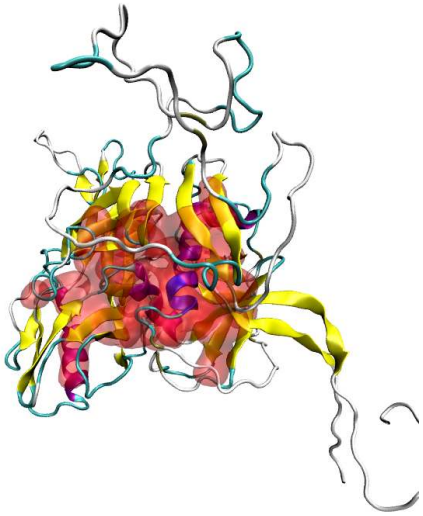

Figure B.17: EVM imagery displaying the selected residues for sequence 61_CH264TF in red.

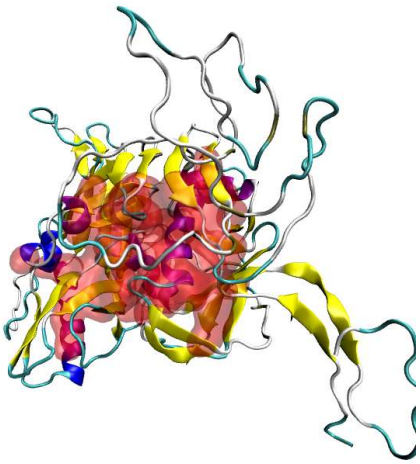

Figure B.18: EVM imagery displaying the selected residues for sequence 62_CH264M6 in red.

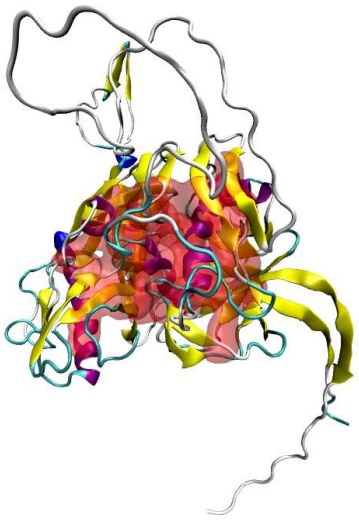

Figure B.19: EVM imagery displaying the selected residues for sequence 63_CH164M6 in red.

Figure B.20: EVM imagery displaying the selected residues for sequence 64_CH164TF in red.

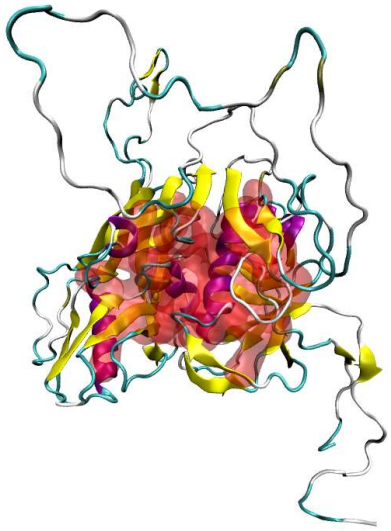

Figure B.21: EVM imagery displaying the selected residues for sequence 3w.21dps in red.

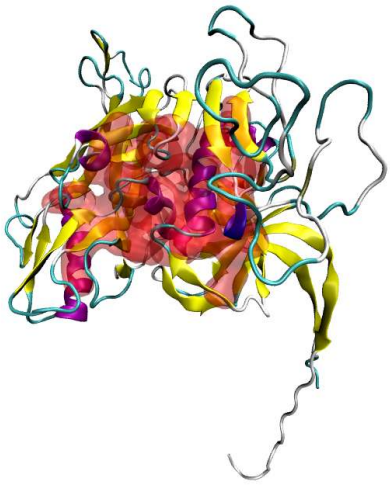

Figure B.22: EVM imagery displaying the selected residues for sequence 1992.133-7 in red.

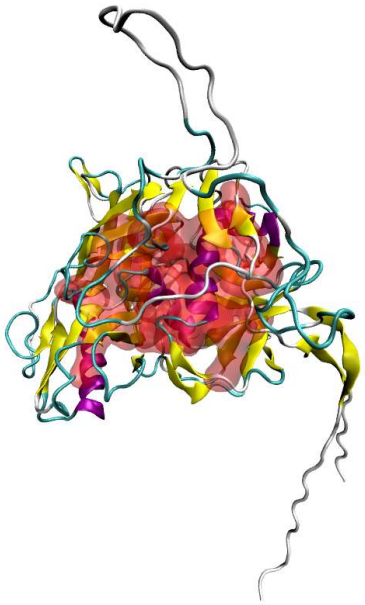

Figure B.23: EVM imagery displaying the selected residues for sequence 1993.153-10 in red.

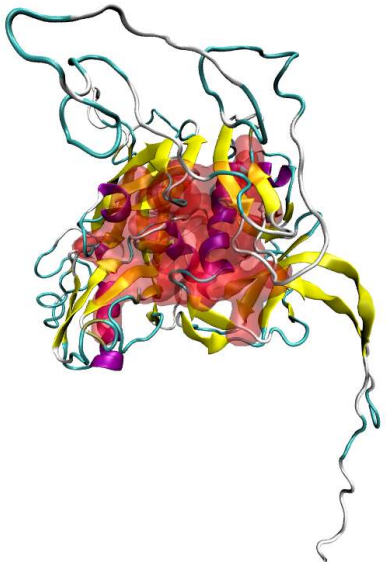

Figure B.24: EVM imagery displaying the selected residues for sequence 1993.159-4 in red.

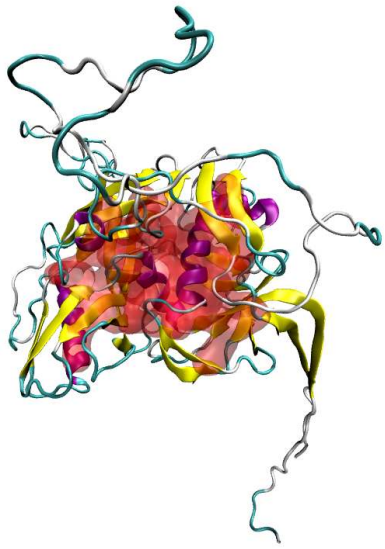

Figure B.25: EVM imagery displaying the selected residues for sequence 1994.309-2 in red.

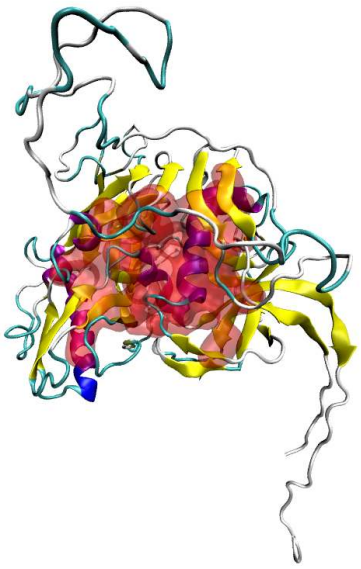

Figure B.26: EVM imagery displaying the selected residues for sequence 1997.133-L-10 in red.

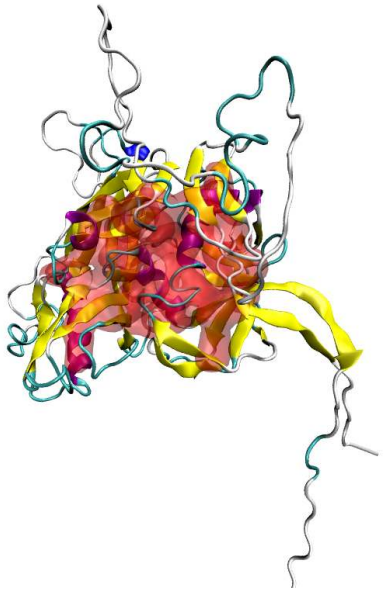

Figure B.27: EVM imagery displaying the selected residues for sequence 1997.159-L-1 in red.

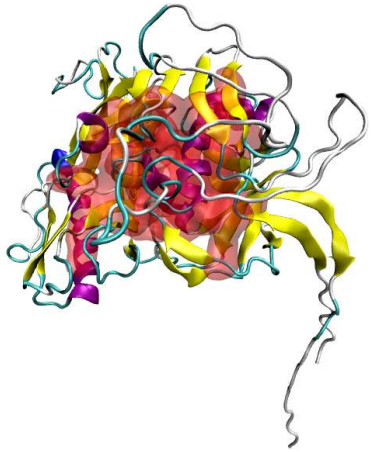

Figure B.28: EVM imagery displaying the selected residues for sequence 1999.153-L-7 in red.

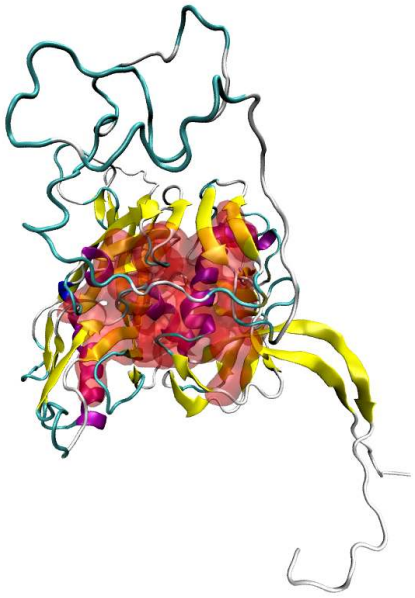

Figure B.29: EVM imagery displaying the selected residues for sequence 2000.309-L-7 in red.

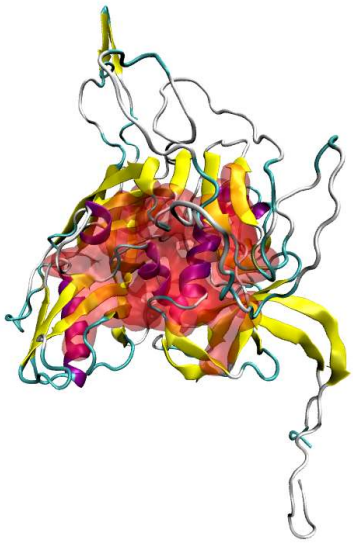

Figure B.30: EVM imagery displaying the selected residues for sequence 2004.MM42-d22_GN1 in red.

Figure B.31: EVM imagery displaying the selected residues for sequence 2005.MM42-d324_GN1 in red.

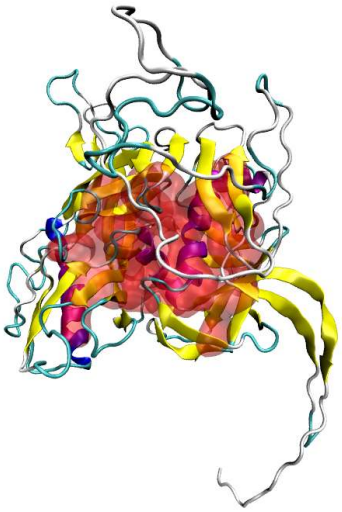

Figure B.32: EVM imagery displaying the selected residues for sequence 1985.H2_5_12E3 in red.

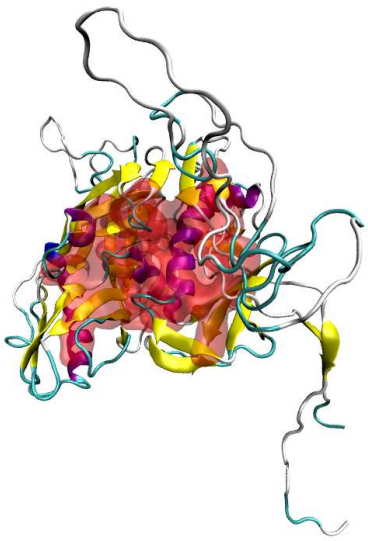

Figure B.33: EVM imagery displaying the selected residues for sequence 1985.H5_4 in red.

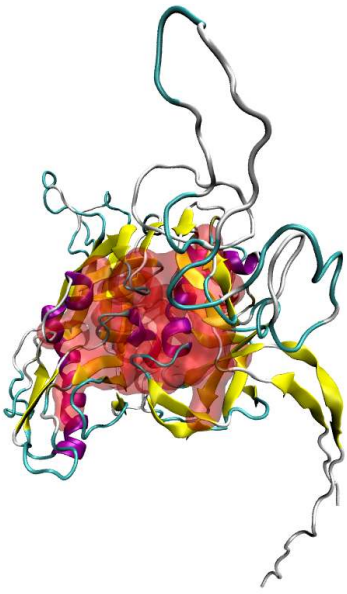

Figure B.34: EVM imagery displaying the selected residues for sequence 1986.H1_7_2D5 in red.

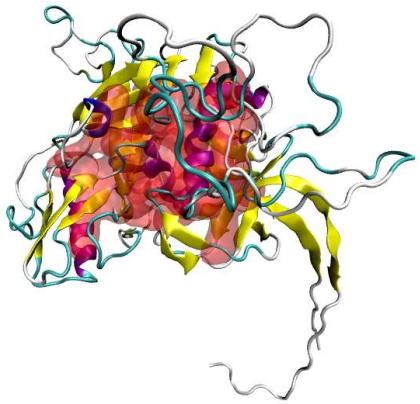

Figure B.35: EVM imagery displaying the selected residues for sequence 1986.H4_007_-1C11 in red.

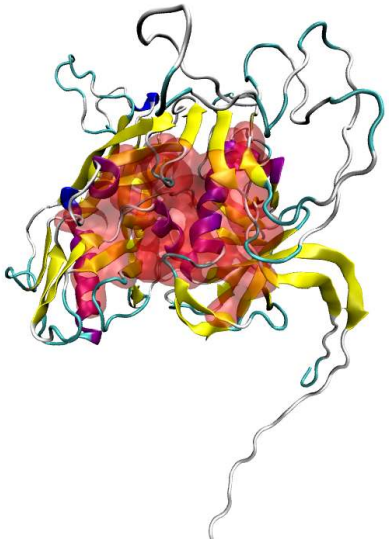

Figure B.36: EVM imagery displaying the selected residues for sequence 1987.H3_12_7D5 in red.

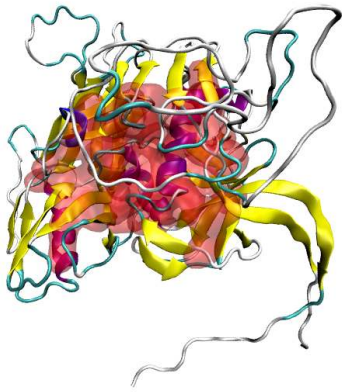

Figure B.37: EVM imagery displaying the selected residues for sequence 1995.H2_114_-8F6 in red.

Figure B.38: EVM imagery displaying the selected residues for sequence 1996.H1_62_1A8 in red.

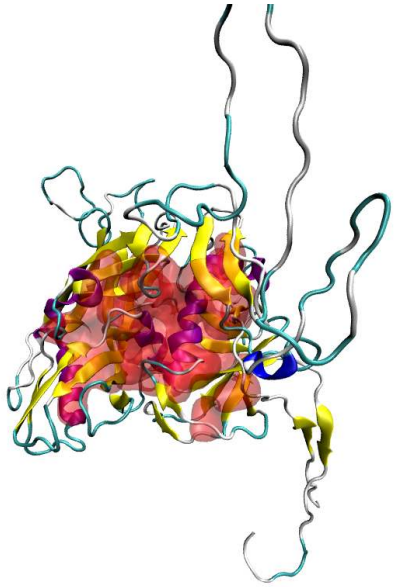

Figure B.39: EVM imagery displaying the selected residues for sequence 1996.H5_75_-7G12 in red.

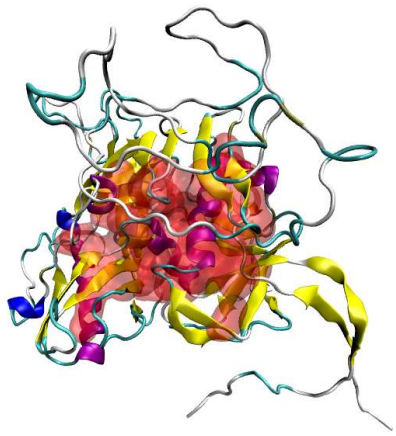

Figure B.40: EVM imagery displaying the selected residues for sequence 1997.H3_110_-8G7 in red.

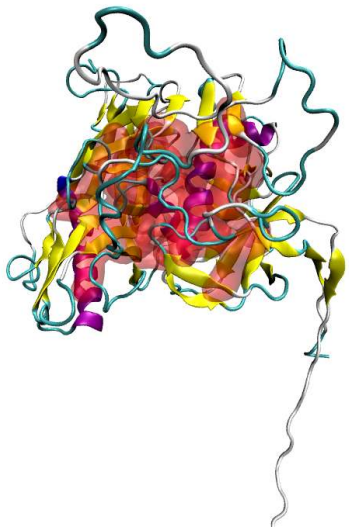

Figure B.41: EVM imagery displaying the selected residues for sequence 1998.H4_146_-2H10 in red.

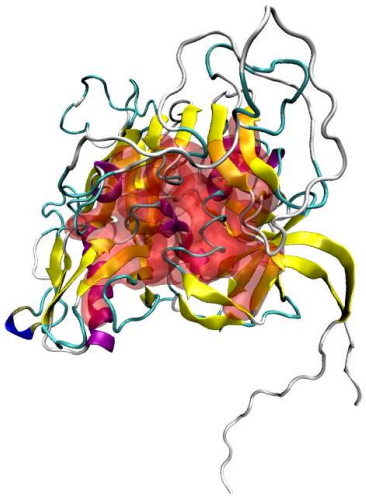

Figure B.42: EVM imagery displaying the selected residues for sequence BORI556_49 in red.

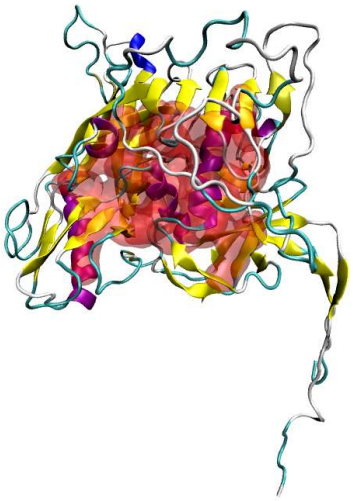

Figure B.43: EVM imagery displaying the selected residues for sequence HOBRd16_20 in red.

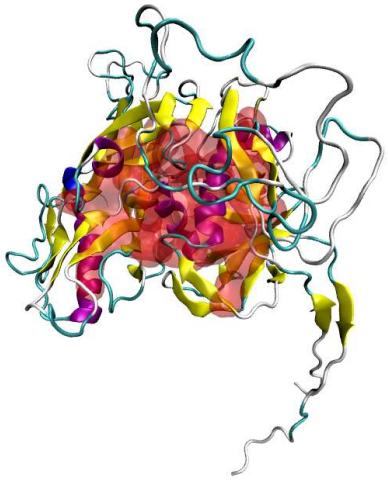

Figure B.44: EVM imagery displaying the selected residues for sequence SUMA736_59 in red.

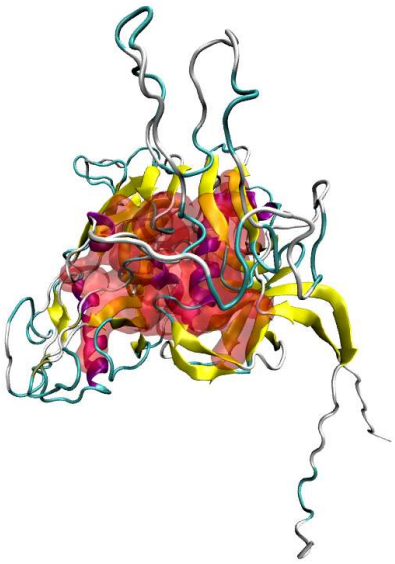

Figure B.45: EVM imagery displaying the selected residues for sequence 1990.BORId9_-3F12 in red.

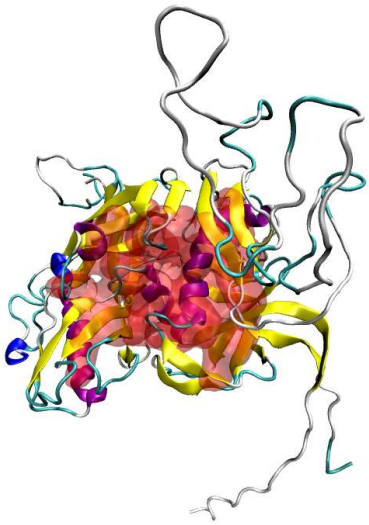

Figure B.46: EVM imagery displaying the selected residues for sequence 1990.WEAU-d15_B2 in red.

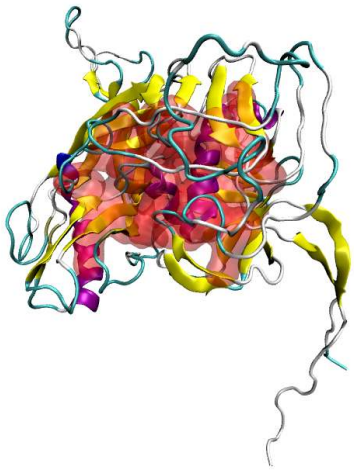

Figure B.47: EVM imagery displaying the selected residues for sequence 1991.HOBR-0961_A21 in red.

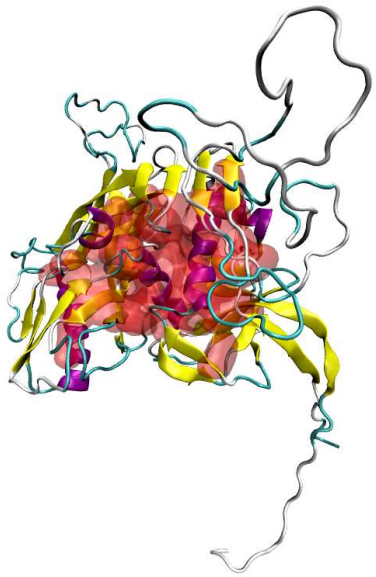

Figure B.48: EVM imagery displaying the selected residues for sequence 1991.SUMAd4_-A32 in red.

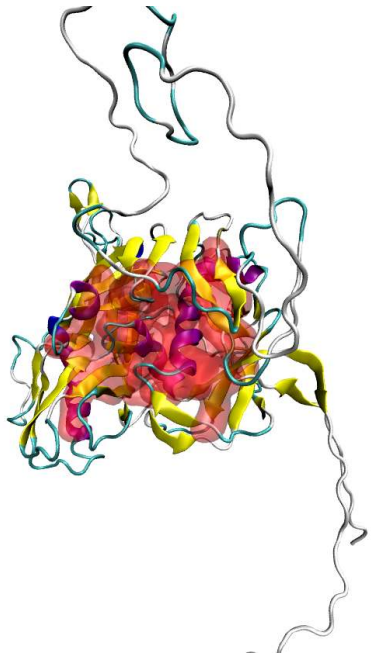

Figure B.49: EVM imagery displaying the selected residues for sequence 1993.WEAU-1166_39 in red.

EVM Selections

- 03_CH40TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 164 166 178 179 181 210 213 214 215 216 218 219 225 226 228 248 252 259 261 326 332 333 338 340 353 384 385 387 388 389 401 403 405 408 413 414 429 430 431 432 433 435 437 438 439 440 441 443 445

- 46_CH40M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 164 166 178 179 181 210 213 214 215 216 218 219 225 226 228 248 252 259 261 326 332 333 338 340 353 384 385 387 388 389 401 403 405 408 413 414 429 430 431 432 433 435 437 438 439 440 441 443 445

- 47_CH58TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 168 170 182 183 185 214 217 218

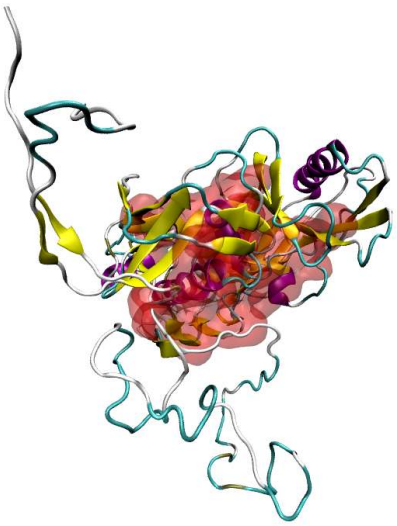

Figure B.50: EVM imagery displaying the selected residues for sequence Z242MPL-25JAN03PCR23ENV1.1-DT in red.

219 220 222 223 229 230 232 252 256 263 265 330 336 337 342 344 357 387 388 390 391 392 404 406 408 411 416 417 430 431 432 433 434 436 438 439 440 441 442 444 446

- 48_CH58M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 168 170 182 183 185 214 217 218 219 220 222 223 229 230 232 252 256 263 265 330 336 337 342 344 357 387 388 390 391 392 404 406 408 411 416 417 430 431 432 433 434 436 438 439 440 441 442 444 446

- 49_CH77TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 167 169 181 182 184 213 216 217 218 219 221 222 228 229 231 251 255 262 264 330 336 337 342 344 357 386 387 389 390 391 403 405 407 410 415 416 430 431 432 433 434 436 438 439 440 441 442 444 446

- 50_CH77M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 167 169 181 182 184 213 216 217 218 219 221 222 228 229 231 251 255 262 264 330 336 337 342 344 357 386 387 389 390 391 403 405 407 410 415 416 430 431 432 433 434 436 438 439 440 441 442 444 446

- 51_CH470TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 339 345 346 351 353 366 398 399 401 402 403 415 417 419 422 427 428 444 445 446 447 448 450 452 453 454 455 456 458 460

- 52_CH470M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 339 345 346 351 353 366 398 399 401 402 403 415 417 419 422 427 428 444 445 446 447 448 450 452 453 454 455 456 458 460

- 53_CH569TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 161 163 175 176 178 207 210 211 212 213 215 216 222 223 225 245 249 256 258 323 329 330 335 337 350 377 378 380 381 382 394 396 398 401 406 407 421 422 423 424 425 427 429 430 431 432 433 435 437

- 54_CH569M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 161 163 175 176 178 207 210 211 212 213 215 216 222 223 225 245 249 256 258 323 329 330 335 337 350 377 378 380 381 382 394 396 398 401 406 407 421 422 423 424 425 427 429 430 431 432 433 435 437

- 55_CH42TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 171 173 185 186 188 217 220 221 222 223 225 226 232 233 235 255 259 266 268 334 340 341 346 348 361 395 396 398 399 400 412 414 416 419 424 425 441 442 443 444 445 447 449 450 451 452 453 455 457

- 56_CH42M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 171 173 185 186 188 217 220 221 222 223 225 226 232 233 235 255 259 266 268 334 340 341 346 348 361 395 396 398 399 400 412 414 416 419 424 425 441 442 443 444 445 447 449 450 451 452 453 455 457

- 57_CH236TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 181 183 195 196 198 227 230 231 232 233 235 236 242 243 245 265 269 276 278 344 350 351 356 358 371 403 404 406 407 408 420 422 424 427 432 433 450 451 452 453 454 456 458 459 460 461 462 464 466

- 58_CH236M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 181 183 195 196 198 227 230 231 232 233 235 236 242 243 245 265 269 276 278 344 350 351 356 358 371 403 404 406 407 408 420 422 424 427 432 433 450 451 452 453 454 456 458 459 460 461 462 464 466

- 59_CH850TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 171 173 185 186 188 217 220 221 222 223 225 226 232 233 235 255 259 266 268 334 340 341 346 348 361 388 389 391 392 393 405 407 409 412 417 418 431 432 433 434 435 437 439 440 441 442 443 445 447

- 60_CH850M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 171 173 185 186 188 217 220 221 222 223 225 226 232 233 235 255 259 266 268 334 340 341 346 348 361 388 389 391 392 393 405 407 409 412 417 418 432 433 434 435 436 438 440 441 442 443 444 446 448

- 61_CH264TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 337 343 344 349 351 364 396 397 399 400 401 413 415 417 420 425 426 438 439 440 441 442 444 446 447 448 449 450 452 454

- 62_CH264M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 337 343 344 349 351 364 396 397 399 400 401 413 415 417 420 425 426 438 439 440 441 442 444 446 447 448 449 450 452 454

- 63_CH164M6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 165 167 179 180 182 211 214 215 216 217 219 220 226 227 229 249 253 260 262 328 334 335 340 342 355 388 389 391 392 393 405 407 409 412 417 418 432 433 434 435 436 438 440 441 442 443 444 446 448

- 64_CH164TF:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 165 167 179 180 182 211 214 215 216 217 219 220 226 227 229 249 253 260 262 328 334 335 340 342 355 388 389 391 392 393 405 407 409 412 417 418 432 433 434 435 436 438 440 441 442 443 444 446 448

- 3w.21dps:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 173 175 187 188 190 219 222 223 224 225 227 228 234 235 237 257 261 268 270 336 342 343 348 350 363 395 396 398 399 400 412 414 416 419 424 425 440 441 442 443 444 446 448 449 450 451 452 454 456

- 1992.133-7:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 161 163 175 176 178 207 210 211 212 213 215 216 222 223 225 245 249 256 258 324 330 331 336 338 351 381 382 384 385 386 398 400 402 405 410 411 426 427 428 429 430 432 434 435 436 437 438 440 442

- 1993.153-10:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 162 164 176 177 179 208 211 212 213 214 216 217 223 224 226 246 250 257 259 326 332 333 338 340 353 389 390 392 393 394 406 408 410 413 418 419 434 435 436 437 438 440 442 443 444 445 446 448 450

- 1993.159-4:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 159 161 173 174 176 205 208 209 210 211 213 214 220 221 223 243 247 254 256 322 328 329 334 336 349 387 388 390 391 392 404 406 408 411 416 417 434 435 436 437 438 440 442 443 444 445 446 448 450

- 1994.309-2:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 169 171 183 184 186 215 218 219 220 221 223 224 230 231 233 253 257 264 266 332 338 339 344 346 359 391 392 394 395 396 408 410 412 415 420 421 436 437 438 439 440 442 444 445 446 447 448 450 452

- 1997.133-L-10:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 255 259 266 268 334 340 341 346 348 361 397 398 400 401 402 414 416 418 421 426 427 441 442 443 444 445 447 449 450 451 452 453 455 457

- 1997.159-L-1:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 159 161 173 174 176 205 208 209 210 211 213 214 220 221 223 243 247 254 256 322 328 329 334 336 349 387 388 390 391 392 404 406 408 411 416 417 434 435 436 437 438 440 442 443 444 445 446 448 450

- 1999.153-L-7:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 162 164 176 177 179 208 211 212 213 214 216 217 223 224 226 246 250 257 259 325 331 332 337 339 352 383 384 386 387 388 400 402 404 407 412 413 428 429 430 431 432 434 436 437 438 439 440 442 444

- 2000.309-L-7:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 175 177 189 190 192 221 224 225 226 227 229 230 236 237 239 259 263 270 272 339 345 346 351 353 366 394 395 397 398 399 411 413 415 418 423 424 443 444 445 446 447 449 451 452 453 454 455 457 459

- 2004.MM42d22_GN1:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 166 168 180 181 183 212 215 216 217 218 220 221 227 228 230 250 254 261 263 328 334 335 340 342 355 381 382 384 385 386 398 400 402 405 410 411 427 428 429 430 431 433 435 436 437 438 439 441 443

- 2005.MM42d324_GN1:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 166 168 180 181 183 212 215 216 217 218 220 221 227 228 230 250 254 261 263 328 334 335 340 342 355 381 382 384 385 386 398 400 402 405 410 411 427 428 429 430 431 433 435 436 437 438 439 441 443

- 1985.H2_5_12E3:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 169 171 183 184 186 215 218 219 220 221 223 224 230 231 233 253 257 264 266 332 338 339 344 346 359 395 396 398 399 400 412 414 416 419 424 425 441 442 443 444 445 447 449 450 451 452 453 455 457

- 1985.H5_4:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 166 168 180 181 183 212 215 216 217 218 220 221 227 228 230 250 254 261 263 329 335 336 341 343 356 386 387 389 390 391 403 405 407 410 415 416 431 432 433 434 435 437 439 440 441 442 443 445 447

- 1986.H1_7_2D5:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 164 166 178 179 181 210 213 214 215 216 218 219 225 226 228 248 252 259 261 327 333 334 339 341 354 386 387 389 390 391 403 405 407 410 415 416 434 435 436 437 438 440 442 443 444 445 446 448 450

- 1986.H4_007_1C11:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 166 168 180 181 183 212 215 216 217 218 220 221 227 228 230 250 254 261 263 329 335 336 341 343 356 388 389 391 392 393 405 407 409 412 417 418 431 432 433 434 435 437 439 440 441 442 443 445 447

- 1987.H3_12_7D5:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 160 162 174 175 177 206 209 210 211 212 214 215 221 222 224 244 248 255 257 323 329 330 335 337 350 381 382 384 385 386 398 400 402 405 410 411 425 426 427 428 429 431 433 434 435 436 437 439 441

- 1995.H2_114_8F6:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 175 177 189 190 192 221 224 225 226 227 229 230 236 237 239 259 263 270 272 338 344 345 350 352 365 397 398 400 401 402 414 416 418 421 426 427 441 442 443 444 445 447 449 450 451 452 453 455 457

- 1996.H1_62_1A8:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 177 179 191 192 194 223 226 227 228 229 231 232 238 239 241 261 265 272 274 340 346 347 352 354 367 415 416 418 419 420 432 434 436 439 444 445 462 463 464 465 466 468 470 471 472 473 474 476 478

- 1996.H5_75_7G12:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 167 169 181 182 184 213 216 217 218 219 221 222 228 229 231 251 255 262 264 330 336 337 342 344 357 384 385 387 388 389 401 403 405 408 413 414 428 429 430 431 432 434 436 437 438 439 440 442 444

- 1997.H3_110_8G7:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 179 181 193 194 196 225 228 229 230 231 233 234 240 241 243 263 267 274 276 342 348 349 354 356 369 404 405 407 408 409 421 423 425 428 433 434 450 451 452 453 454 456 458 459 460 461 462 464 466

- 1998.H4_146_2H10:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 170 172 184 185 187 216 219 220 221 222 224 225 231 232 234 254 258 265 267 334 340 341 346 348 361 393 394 396 397 398 410 412 414 417 422 423 438 439 440 441 442 444 446 447 448 449 450 452 454

- BORI556_49:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 167 169 181 182 184 213 216 217 218 219 221 222 228 229 231 251 255 262 264 331 337 338 343 345 358 398 399 401 402 403 415 417 419 422 427 428 443 444 445 446 447 449 451 452 453 454 455 457 459

- HOBRd16_20:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 168 170 182 183 185 214 217 218 219 220 222 223 229 230 232 252 256 263 265 332 338 339 344 346 359 392 393 395 396 397 409 411 413 416 421 422 437 438 439 440 441 443 445 446 447 448 449 451 453

- SUMA736_59:

  15 19 32 58 63 65 66 69 73 81 90 91 92 93 94 168 170 182 183 185 214 217 218 219 220 222 223 229 230 232 252 256 263 265 331 337 338 343 345 358 392 393 395 396 397 409 411 413 416 421 422 439 440 441 442 443 445 447 448 449 450 451 453 455

- 1990.BORId9_3F12:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 175 177 189 190 192 221 224 225 226 227 229 230 236 237 239 259 263 270 272 339 345 346 351 353 366 405 406 408 409 410 422 424 426 429 434 435 450 451 452 453 454 456 458 459 460 461 462 464 466

- 1990.WEAUd15_B2:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 337 343 344 349 351 364 396 397 399 400 401 413 415 417 420 425 426 441 442 443 444 445 447 449 450 451 452 453 455 457

- 1991.HOBR0961_A21:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 168 170 182 183 185 214 217 218 219 220 222 223 229 230 232 252 256 263 265 332 338 339 344 346 359 392 393 395 396 397 409 411 413 416 421 422 437 438 439 440 441 443 445 446 447 448 449 451 453

- 1991.SUMAd4_A32:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 169 171 183 184 186 215 218 219 220 221 223 224 230 231 233 253 257 264 266 332 338 339 344 346 359 393 394 396 397 398 410 412 414 417 422 423 440 441 442 443 444 446 448 449 450 451 452 454 456

- 1993.WEAU1166_39:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 172 174 186 187 189 218 221 222 223 224 226 227 233 234 236 256 260 267 269 337 343 344 349 351 364 396 397 399 400 401 413 415 417 420 425 426 440 441 442 443 444 446 448 449 450 451 452 454 456

- Z242MPL25JAN03PCR23ENV1.1-DT:

  15 19 32 59 64 66 67 70 74 82 91 92 93 94 95 166 168 180 181 183 212 215 216 217 218 220 221 227 228 230 249 253 260 262 327 333 334 339 341 354 378 379 381 382 383 395 397 399 402 407 408 421 422 423 424 425 427 429 430 431 432 433 435 437

Figure B.51: EVM imagery displaying the selected residues for sequence R56MCF21aug-0511_plasmid_1v in red.

Complete EVM Results for Morton et al. (2019)

Figure B.52: EVM imagery displaying the selected residues for sequence R56MPL21apr-05C5_plasmid_6-4 in red.

Figure B.53: EVM imagery displaying the selected residues for sequence Z153FPB-13MAR02ENV1.1 in red.

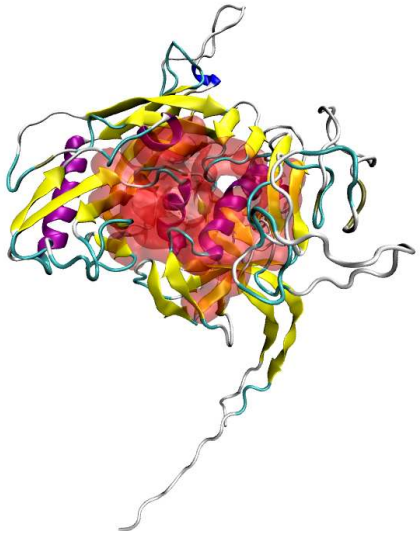

Figure B.54: EVM imagery displaying the selected residues for sequence Z153FPL-13MAR02ENV6.1 in red.

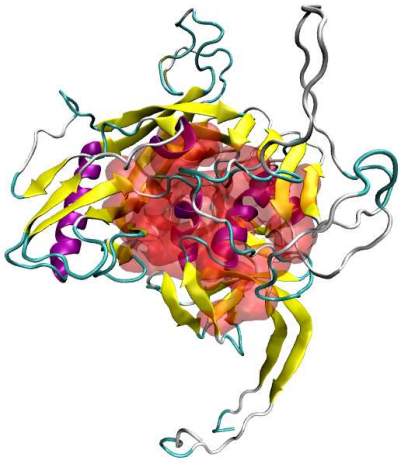

Figure B.55: EVM imagery displaying the selected residues for sequence Z185MPB-17AUG02ENV1.2 in red.

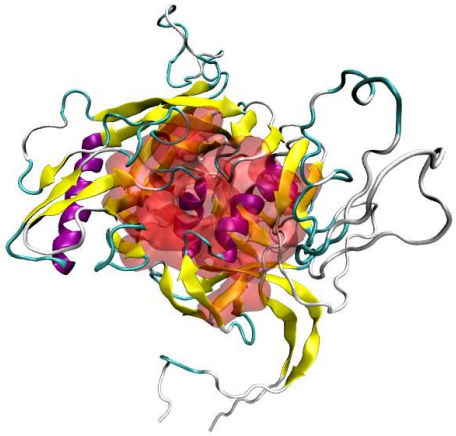

Figure B.56: EVM imagery displaying the selected residues for sequence Z185MPB-17AUG02ENVB17 in red.

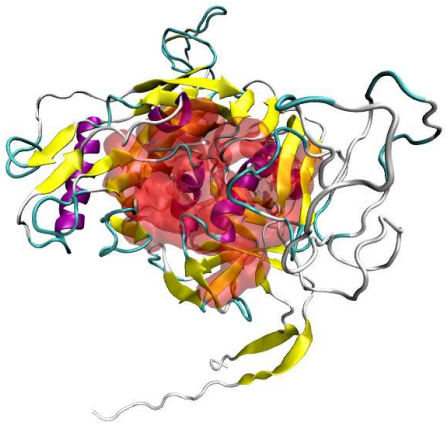

Figure B.57: EVM imagery displaying the selected residues for sequence Z201FCF-07feb03DNA13C18 in red.

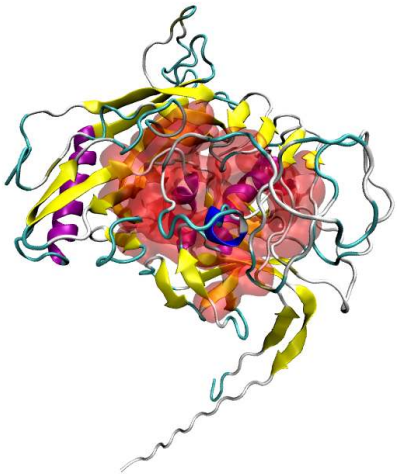

Figure B.58: EVM imagery displaying the selected residues for sequence Z201FPL-7FEB03ENV2.1 in red.

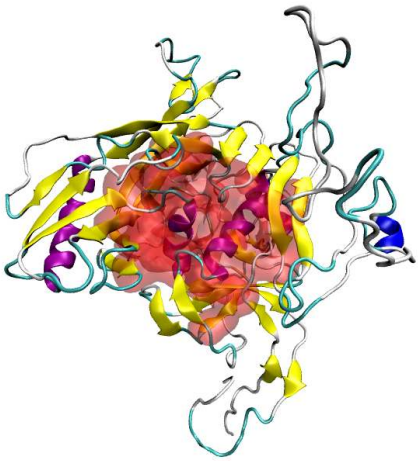

Figure B.59: EVM imagery displaying the selected residues for sequence Z205MPB-27MAR03ENV6.1 in red.

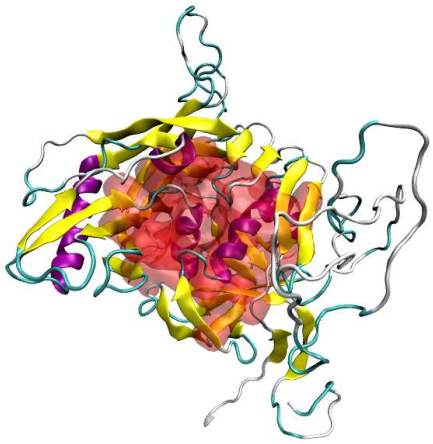

Figure B.60: EVM imagery displaying the selected residues for sequence Z205MPB-27MAR03ENV9.1 in red.

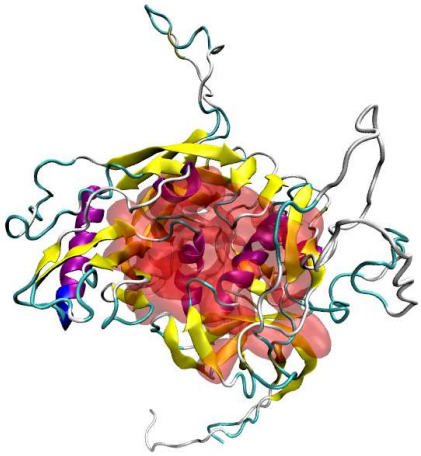

Figure B.61: EVM imagery displaying the selected residues for sequence Z216FPB98_-plasmid_e in red.

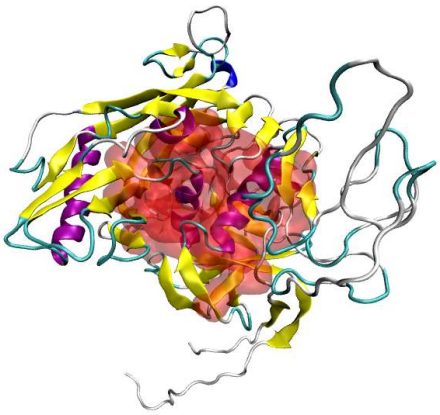

Figure B.62: EVM imagery displaying the selected residues for sequence Z216FPL-17jan0485f in red.

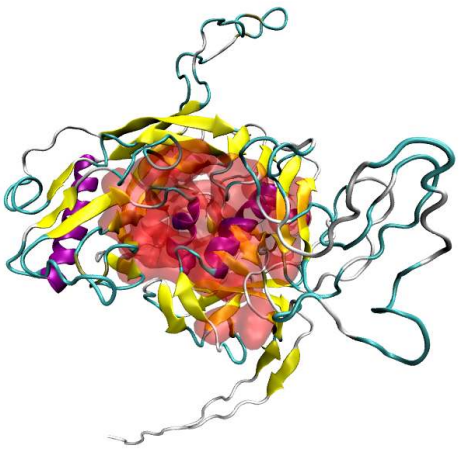

Figure B.63: EVM imagery displaying the selected residues for sequence Z221FPL55_-plasmid_6-2 in red.

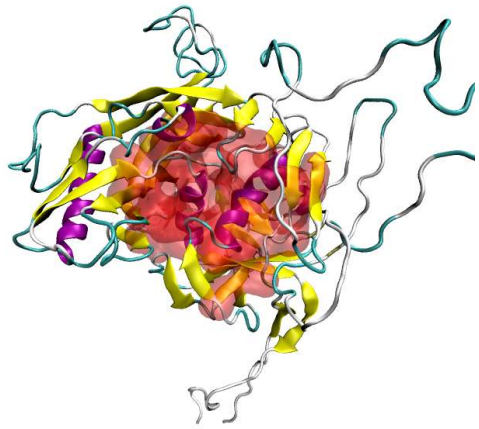

Figure B.64: EVM imagery displaying the selected residues for sequence Z221FPL-7MAR03ENV2.3 in red.

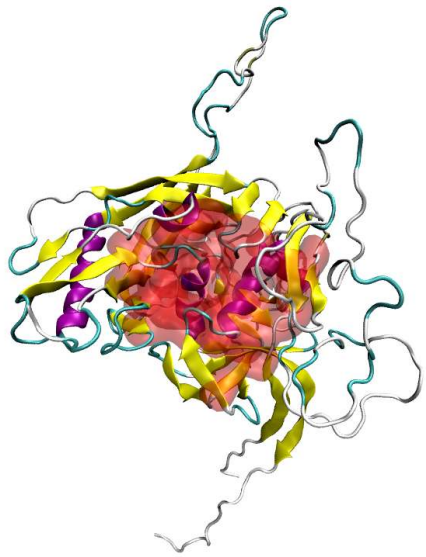

Figure B.65: EVM imagery displaying the selected residues for sequence Z238FCF-29oct0215A39 in red.

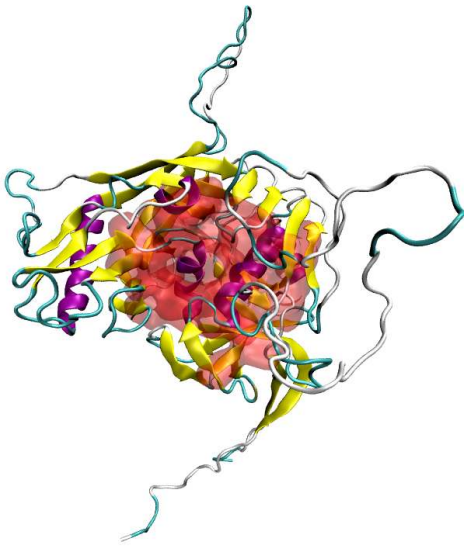

Figure B.66: EVM imagery displaying the selected residues for sequence Z238FSW-29oct0215A6v in red.

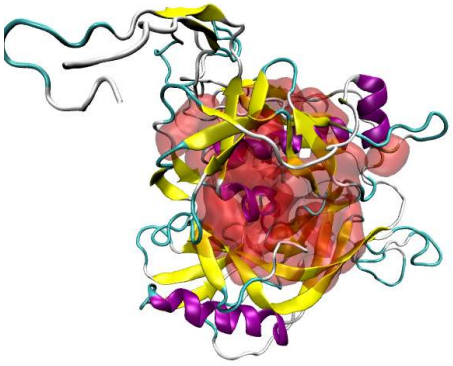

Figure B.67: EVM imagery displaying the selected residues for sequence Z242MPL-25JAN03PCR23ENV1.1-DT in red.

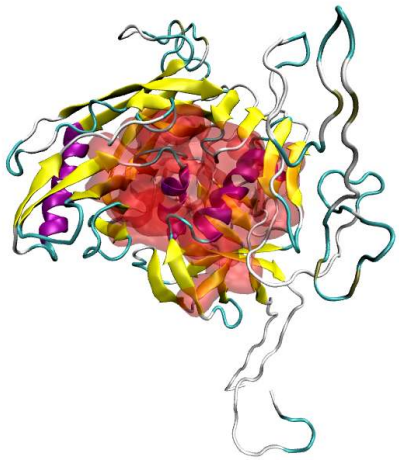

Figure B.68: EVM imagery displaying the selected residues for sequence Z242MPL26_-plasmid in red.

EVM Selections

- R56MCF21aug0511_plasmid_1v:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 161 163 176 178 207 211 212 213 215 216 222 223 225 235 245 249 256 258 330 331 336 338 378 381 383 397 399 402 407 423 424 425 426 427 429 431 432 434 435 437 439

- R56MPL21apr05C5_plasmid_6-4:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 166 168 181 183 212 216 217 218 220 221 227 228 230 240 250 254 261 263 335 336 341 343 378 381 383 397 399 402 407 422 423 424 425 426 428 430 431 433 434 436 438

- Z153FPB13MAR02ENV1.1:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 161 163 176 178 207 211 212 213 215 216 222 223 225 235 245 249 256 258 331 332 337 339 376 379 381 395 397 400 405 418 419 420 421 422 424 426 427 429 430 432 434

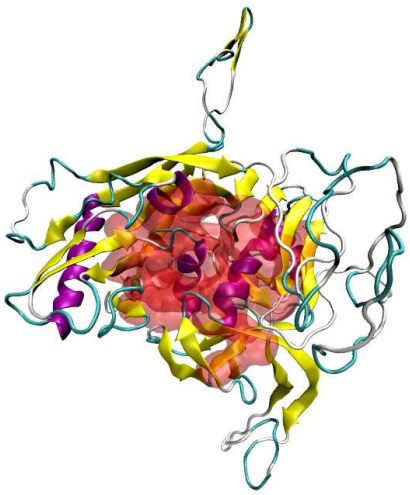

Figure B.69: EVM imagery displaying the selected residues for sequence Z292FCF-24may0512D18_plasmid_4i in red.

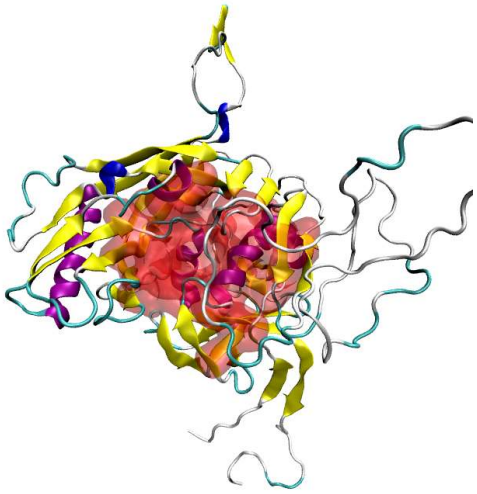

Figure B.70: EVM imagery displaying the selected residues for sequence Z292FCF-24may0512E26_plasmid_10iv in red.

- Z153FPL13MAR02ENV6.1:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 161 163 176 178 207 211 212 213 215 216 222 223 225 235 245 249 256 258 329 330 335 337 374 377 379 393 395 398 403 421 422 423 424 425 427 429 430 432 433 435 437

- Z185MPB17AUG02ENV1.2:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 165 167 180 182 211 215 216 217 219 220 226 227 229 239 249 253 260 262 334 335 340 342 386 389 391 405 407 410 415 431 432 433 434 435 437 439 440 442 443 445 447

- Z185MPB17AUG02ENVB17:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 165 167 180 182 211 215 216 217 219 220 226 227 229 239 249 253 260 262 334 335 340 342 384 387 389 403 405 408 413 426 427 428 429 430 432 434 435 437 438 440 442

- Z201FCF07feb03DNA13C18:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 174 176 189 191 220 224 225 226 228 229 235 236 238 248 258 262 269 271 343 344 349 351 395 398 400 414 416 419 424 437 438 439 440 441 443 445 446 448 449 451 453

- Z201FPL7FEB03ENV2.1:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 174 176 189 191 220 224 225 226 228 229 235 236 238 248 258 262 269 271 343 344 349 351 395 398 400 414 416 419 424 437 438 439 440 441 443 445 446 448 449 451 453

- Z205MPB27MAR03ENV6.1:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 172 174 187 189 218 222 223 224 226 227 233 234 236 246 256 260 267 269 341 342 347 349 389 392 394 408 410 413 418 433 434 435 436 437 439 441 442 444 445 447 449

- Z205MPB27MAR03ENV9.1:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 172 174 187 189 219 223 224 225 227 228 234 235 237 247 257 261 268 270 342 343 348 350 387 390 392 406 408 411 416 431 432 433 434 435 437 439 440 442 443 445 447

- Z216FPB98_plasmid_e:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 169 171 184 186 215 219 220 221 223 224 230 231 233 243 253 257 264 266 339 340 345 347 393 396 398 412 414 417 422 439 440 441 442 443 445 447 448 450 451 453 455

- Z216FPL17jan0485f:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 164 166 179 181 210 214 215 216 218 219 225 226 228 238 248 252 259 261 333 334 339 341 383 386 388 402 404 407 412 426 427 428 429 430 432 434 435 437 438 440 442

- Z221FPL55_plasmid_6-2:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 186 188 201 203 232 236 237 238 240 241 247 248 250 260 270 274 281 283 354 355 360 362 406 409 411 425 427 430 435 453 454 455 456 457 459 461 462 464 465 467 469

- Z221FPL7MAR03ENV2.3:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 173 175 188 190 219 223 224 225 227 228 234 235 237 247 257 261 268 270 342 343 348 350 394 397 399 413 415 418 423 439 440 441 442 443 445 447 448 450 451 453 455

- Z238FCF29oct0215A39:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 163 165 178 180 209 213 214 215 217 218 224 225 227 237 247 251 258 260 332 333 338 340 379 382 384 398 400 403 408 425 426 427 428 429 431 433 434 436 437 439 441

- Z238FSW29oct0215A6v:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 163 165 178 180 209 213 214 215 217 218 224 225 227 237 247 251 258 260 332 333 338 340 388 391 393 407 409 412 417 436 437 438 439 440 442 444 445 447 448 450 452

- Z242MPL25JAN03PCR23ENV1.1-DT:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 166 168 181 183 212 216 217 218 220 221 227 228 230 239 249 253 260 262 333 334 339 341 378 381 383 397 399 402 407 421 422 423 424 425 427 429 430 432 433 435 437

- Z242MPL26_plasmid:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 166 168 181 183 212 216 217 218 220 221 227 228 230 239 249 253 260 262 333 334 339 341 378 381 383 397 399 402 407 421 422 423 424 425 427 429 430 432 433 435 437

- Z292FCF24may0512D18_plasmid_4i:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 172 174 187 189 218 222 223 224 226 227 233 234 236 246 256 260 267 269 342 343 348 350 394 397 399 413 415 418 423 438 439 440 441 442 444 446 447 449 450 452 454

- Z292FCF24may0512E26_plasmid_10iv:

  15 17 19 32 59 64 66 67 70 74 82 91 92 93 94 170 172 185 187 216 220 221 222 224 225 231 232 234 244 254 258 265 267 340 341 346 348 393 396 398 412 414 417 422 437 438 439 440 441 443 445 446 448 449 451 453

APPENDIX C : BINDING ENERGIES

Experiment 2 of Unpublished Works

Table C.1: Experiment 2 is a duplication of the procedures used to produce results for Table 4.12 (top) to validate methods.

		WITO.33 ID#4164		TRO.11 ID#4654		CAAN.A2 ID#1839		6535.3 ID#5021	
Mab	Specificity	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**	**pH5.5**	**pH7.4**
VRC01	CD4bs	0.38	0.41	1.09	1.74	1.41	3.95	4.37	6.55
3BNC117	CD4bs	N/A	N/A	0.08	0.15	0.74	2.42	N/A	>25
CH31	CD4bs	0.23	0.19	0.14	0.44	11.31	>25	>25	>25
CH01	V2/V3/Glycan	0.17	0.14	>25	>25	>25	>25	3.46	3.7
PG9	V2/V3/Glycan	0.05	0.03	>5	>5	>5	>5	1.35	0.95
PG16	V2/V3/Glycan	0.02	0.01	>5	3.96	>5	>5	>5	>5
PGT121	N332 Glycan	>5	4.19	0.06	0.09	0.05	0.08	0.06	0.04
PGT128	N332 Glycan	>5	>5	0.10	0.06	0.87	0.63	0.07	0.04
IgG1b12	CD4bs	>25	>25	>25	>25	>25	>25	>25	15.11
2G12	Glycan	2.94	3.04	0.38	0.41	>25	>25	8.27	13.25
2F5	MPER	2.13	6.35	>25	>25	>25	>25	17.68	>25
4E10	MPER	5.92	9.28	11.58	7.68	>25	>25	>25	12.83
10E8		0.67	0.71	0.83	0.26	>5	>5	>5	2.64
HIVIG-C	Polyclonal	>625	541.19	110.98	111.15	173.68	310.23	74.99	81.44
	Vir. dilution	undiluted		diluted 1:8		diluted 1:10		undiluted	
	VCTRL	95K	150K	30K	71K	33K	55K	60K	100K