Using Molecular Modeling to Determine Protein Stabilities and Energies

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Physics

in the

College of Graduate Studies

University of Idaho

by

Kyle P. Martin

Major Professor: F. Marty Ytreberg, Ph.D.

Committee Members: Jason W. Barnes, Ph.D.; Celeste J. Brown, Ph.D.; Jagdish S. Patel, Ph.D.

Department Administrator: John R. Hiller, Ph.D.

May 2020

Authorization to Submit Dissertation

This Dissertation of Kyle P. Martin, submitted for the degree of Doctor of Philosophy with a Major in Physics and titled "Using Molecular Modeling to Determine Protein Stabilities and Energies," has been reviewed in final form. Permission, as indicated by the signatures and dates below is now granted to submit final copies for the College of Graduate Studies for approval.

Advisor:	F. Marty Ytreberg, Ph.D.	Date
Committee Members:	Jason W. Barnes, Ph.D.	Date
	Celeste J. Brown, Ph.D.	Date
	Jagdish S. Patel, Ph.D.	Date
Department Chair:	John R. Hiller, Ph.D.	Date

Abstract

Proteins are the molecular machines that perform the functions necessary for life. Interactions between proteins and other biomolecules are at the heart of all biological processes in a cell. This thesis explores how molecular modeling can be used to understand both proteins and their interactions. Examples include antibody-antigen interactions in Ebola, how proteins might behave in the subsurface oceans of Titan, and the ability of different software to accurately predict protein interactions. We predict mutations in Ebola that could lead to antibody escape. We explore aspects of possible life on exoplanets by modeling how Earth-based proteins would behave in the environment thought to exist in subsurface oceans on Titan. We analyze a suite of different software to find those that have better predictive capabilities, depending on the location and type of mutation. In short, we show that molecular modeling can be used to make predictions about protein behavior and interactions.

Acknowledgements

First and foremost I would like to thank the members of my committee. Dr. F. Marty Ytreberg has been much more than my major professor the last seven years. He has been my invaluable mentor who has supported and encouraged me through every step of the way. I thank Dr. Jason W. Barnes for his time, expertise, and help co-authoring my first paper. I thank Dr. Celeste J. Brown for her breadth and depth of experience she brought to the committee. I thank Dr. Jagdish S. Patel for his patience and driving force of inspiration. The final product was greatly improved as a result of their participation. I would also like to thank the faculty and staff of the Department of Physics and the Department of Biological Sciences at the University of Idaho for a friendly, supportive and intellectually stimulating graduate experience.

This study was supported by numerous funds. The Ebola Paper was supported by funds provided by the National Science Foundation (DEB1521049) and the Center for Modeling Complex Interactions sponsored by the National Institutes of Health (P20 GM104420). Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (P30 GM103324). The Titan Paper was supported by the Center for Modeling Complex Interactions sponsored by the National Institutes of Health (P20 GM104420), the National Science Foundation EPSCoR program (OIA-1736253), and the National Aeronautics and Space Administration Earth and Space Science Fellowship Program (NNX14AO30H). Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (P30 GM103324). The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The Methods Paper was supported by the National Science Foundation (OIA1736253) and the Institute for Modeling Collaboration and Innovation (P20 GM104420).

I would also like to thank Chris Mirabzadeh, Caleb Quates, Jonathan Barnes, Casey Beard, Tawny Gonzalez, and Dharmesh Patel.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 An Introduction to Molecular Modeling

Molecular modeling is the science of representing molecular structures numerically and simulating their behavior with the equations of quantum and classical physics [45]. There is an inherent trade-off between accuracy (highest with quantum compared to classical) and simulation time (shortest with classical compared to quantum). Pure classical simulations are thus typically used for larger molecular systems such as proteins, but mixing quantum mechanics with molecular mechanics (QM/MM) is also possible via the Hartree-Fock or the Kohn-Sham model. These mixed methods can be more accurate, but much more computationally expensive than pure classical simulations.

Molecular modeling can also be used to study protein structure, energy, and stability in different environments (see chapter 3), even those that are inaccessible to experiment. Molecular modeling can be used to calculate binding free energies between proteins and potential therapeutics for cancer or disease treatments[159, 119]. The process of drug design typically includes molecular modeling to streamline the drug discovery process[124, 97]. This is done by virtual screening where small molecules are identified that are most likely to bind to a protein target, such as amyloids. Molecular modeling can be used to understand escape mutants and identify vaccines that are more robust to evolution (see chapter 2). The ability of molecular modeling to estimate protein stabilities can be used to understand potential life on other planets[94].

In order to simulate proteins, it is important to have a 3-D representation. Many methods exist to extract 3-D structural data such as cryo-electron microscopy (EM)[19], X-ray crystallography[153], and nuclear magnetic resonance (NMR) spectroscopy[7]. These efforts together have given rise to a dramatic increase in available 3-D structural data over the past few decades, with more than 160,000 structures currently in the protein data bank[10]. Once a 3-D protein structure is obtained, it can be simulated and its energies and stabilities can be analyzed.

1.1.1 Structure Analysis

Molecular modeling can be used to help understand many properties of proteins. For example, simulations of protein complexes can enable us to model the forces driving their assembly, and their stability, which in turn may help us to understand these processes better.

In the absence of a 3-D structure for a specific protein, homology modeling can be used. Homology modeling looks at the sequence of the protein with unknown structure and compares it against similar sequences with known structures. The assumption is that similar sequences lead to similar structures.

Provided the sequences are similar enough, this process can provide an accurate estimate of the unknown structure.

After a simulation is complete the results can be analyzed statistically. Examples of such analyses include the radius of gyration, root-mean-square fluctuation, or Ramachandran plots to predict protein structural stability (see chapter 3).

We can also look at the effects of mutations or environmental changes on a protein. Amino acid changes (called mutations) can destabilize a protein if they are too different from the wildtype amino acid. Some factors that can influence this destabilization are amino acid size, polarity, and charge of the mutation compared to the wildtype. Changes in environment can also influence protein stability, such as the solvent differences (water-ammonia vs water, see chapter 3) or temperature (temperature dependent mutations, see chapter 5).

1.1.2 Free Energy

The free energy is a statistical mechanical quantity that can be thought of as a measure of the probability of finding a system in a given state. In biophysics, free energy is used to determine the folding and binding energies of proteins. The free energy involves both entropy and enthalpy, and will favor conformations with low enthalpy and high entropy.

Protein folding is the physical process by which the amino acid chain acquires its native 3-D structure. The resulting 3-D structure (or lack of 3-D structure) is determined by the amino acid sequence or primary structure[3]. The configuration space of a protein during folding can be visualized as an energy landscape (see fig. 1.1).

Protein-protein interactions are the physical contacts between two or more proteins. This includes the case when two proteins bind to each other. The binding free energy (often called binding affinity, ΔG_{Bind}, see fig. 1.2) represents the strength of this binding interaction.

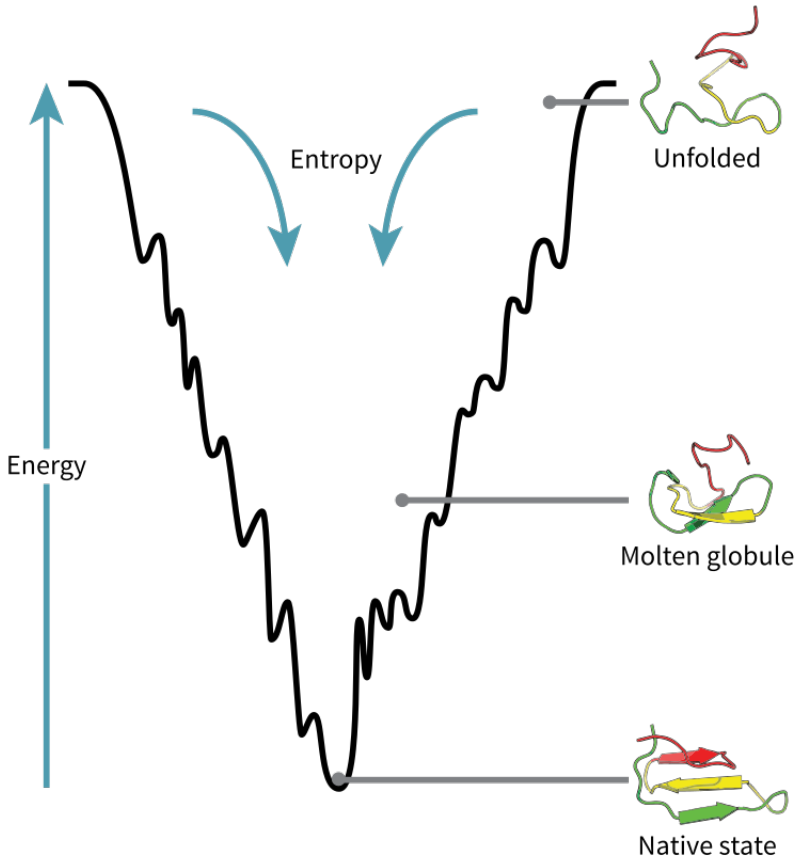


Figure 1.1: The unfolded protein (chain of amino acids) begins at the top of the funnel where it may assume the largest number of unfolded variations and is in its highest free energy state. Energy landscapes such as these indicate that there are a large number of possible configurations, but only a single native (lowest free energy) state. [140].

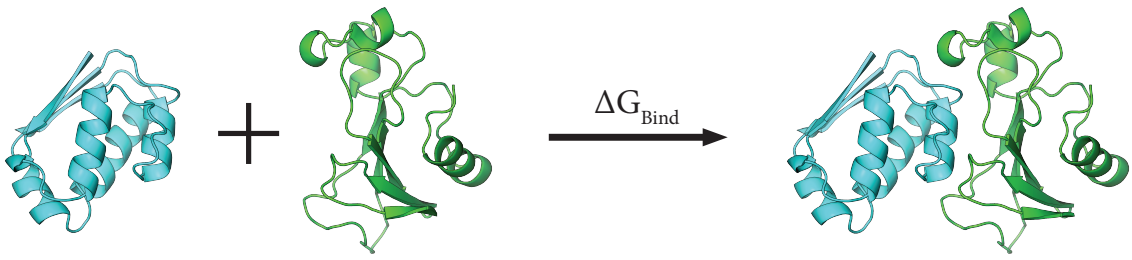


Figure 1.2: Graphical depiction of protein-protein binding. δG shows the binding energy required to bind the two proteins into a single protein complex.

1.2 Overview of Dissertation

Below is a brief summary of the thesis chapters. These show how molecular modeling can be used to understand proteins.

1.2.1 Initiating a watch list for Ebola virus antibody escape mutations

The 2014 Ebola virus (EBOV) outbreak in West Africa was the largest in recorded history and resulted in over 11,000 deaths. It is essential that strategies for treatment and containment be developed to avoid future epidemics of this magnitude. In this study we have initiated a watch list of potential antibody escape mutations of EBOV by modeling interactions between GP and the antibody KZ52. The watch list was generated using molecular modeling to estimate stability changes due to mutation. Every possible mutation of GP was considered and the list was generated from those that are predicted to disrupt GP-KZ52 binding but not to disrupt the ability of GP to fold and to form trimers. The resulting watch list contains 34 mutations (one of which has already been seen in humans) at six sites in the GP2 subunit.

1.2.2 Protein Stability in Titan's Subsurface Water Ocean

Models of Titan predict that there is a subsurface ocean of water and ammonia under a layer of ice. Such an ocean would be important in the search for extraterrestrial life since it provides a potentially habitable environment. To evaluate how Earth-based proteins would behave in Titan's subsurface ocean environment, we used molecular dynamics simulations to calculate the properties of proteins with the most common secondary structure types (alpha helix and beta sheet) in both Earth and Titan-like conditions. We analyzed protein compactness, flexibility, and backbone dihedral distributions to identify differences between the two environments. Secondary structures in the Titan environment were found to be less long-lasting, less flexible, and had small differences in backbone dihedral preferences (e.g., in one instance a pi helix formed). These environment-driven differences could lead to changes in how these proteins interact with other biomolecules and therefore changes in how evolution would potentially shape proteins to function in subsurface ocean environments.

1.2.3 Analysis of Software Methods for Estimation of Protein-Protein Relative Binding Affinity

Here, eight computational tools were assessed on their ability to accurately predict relative binding affinities due to single mutations ($\Delta\Delta G$) for eight antibody-antigen and eight non-antibody-antigen complexes. All methods for predicting $\Delta\Delta G$ values performed worse when applied to antibody-antigen complexes compared to non-antibody-antigen complexes, with a few exceptions. Rosetta-based JayZ and

EasyE were able to classify mutations as destabilizing with a 83-98% accuracy. Combining molecular dynamics with FoldX provided some of the better results for non-antibody-antigen binding affinities with a correlation coefficient of 0.46. Overall, our results suggest that non-rigorous methods can be used to quickly approximate destabilizing mutations, but are less accurate with approximating binding affinities.

1.2.4 DmORC

DmORC was an unfinished project, but led to a better understanding of proteins and molecular dynamics. DmORC was an ongoing project being done in conjunction with experimentalists from the University of Vermont. The DmORC structure was partially rebuilt in Schrodinger and simulated to look at stability.

Chapter 2: Initiating a watch list for Ebola virus antibody escape mutations

Craig R. Miller,[1,2,3] Erin L. Johnson,[3] Aran Z. Burke,[3] Kyle P. Martin,[3,4], Tanya A. Miura,[1,3] Holly A. Wichman,[1,3] Celeste J. Brown,[1,3], F. Marty Ytreberg[3,4]

[1]Department of Biological Sciences, University of Idaho, [2]Department of Mathematics, University of Idaho, [3]Center for Modeling Complex Interactions, University of Idaho, [4]Department of Physics, University of Idaho

Published in PeerJ. On this project I contributed by writing the manuscript, analyzing simulations, and performing some background reading. Writing and editing was done on Overleaf (LaTeXwebsite) in collaboration with all the co-authors. Results of analysis were shared and discussed in group meetings. The background reading was done to understand previously written related research. This paper was previously published in PeerJ and falls under the NIH public access policy (see https://publicaccess.nih.gov/) and can be used freely in this thesis. Final publication is available from PeerJ https://dx.doi.org/10.7717/peerj.1674

.

2.1 Introduction

With nearly 30,000 confirmed cases and over 11,000 deaths, the recent Ebola virus (EBOV) epidemic in West Africa has dwarfed all recorded outbreaks of the disease [26]. Now that the 2014 outbreak appears to be waning it is critical to develop strategies for treatment and containment to avoid future epidemics of this magnitude. One important strategy is the development of vaccines. Two vaccines that express the EBOV envelope glycoprotein (GP) from the 1976 Mayinga strain are in phase III clinical trials: rVSV-ZEBOV and ChAd3-ZEBOV [39, 141, 96]. Early reports suggest that rVSV-ZEBOV is highly effective at preventing EBOV infection [49]. A related strategy is antibody-based therapeutics. For example, ZMapp has been shown to be effective in treating non-human primates and has been used to treat small numbers of humans with Ebola [123, 14]. The monoclonal antibodies in ZMapp were generated by vaccination of mice with GP from the 1976 Mayinga strain [161, 122, 123].

A key course of action to prepare for future EBOV outbreaks is to anticipate how the evolution of antibody escape mutants in the virus might compromise treatment efforts. Antibody escape mutants have arisen in the laboratory when recombinant vesicular stomatitis viruses expressing the GP protein of

EBOV or Marburg virus were grown in the presence of anti-GP antibodies [69]. In that study a single amino acid substitution conferred viral resistance to the antibodies. Similarily, a single amino acid change in GP of the EBOV Kikwit 95 strain in a macaque treated with monoclonal antibodies resulted in fatal infection [121]. Mutational changes in GP have also been found to impact immune responses to the virus; substitutions at N-linked glycosylation sites can alter antigenicity and immunogenicity, in some cases preventing binding to the KZ52 antibody [33, 84]. Antibody escape mutants are also known in influenza A, HIV 1, measles and respiratory syncytial virus infections [134, 40, 128, 168].

Sequencing studies have shown that there is a high level of genetic variation in EBOV and that GP has the largest variation among EBOV proteins [42, 146, 115]. As of August 2015, sequences from the 2014 outbreak show that 106 of the 676 sites in GP experienced a mutation and the strains differ from the 1976 Mayinga strain used in developing interventions by an average of 20.2 nucleotide changes. Thus, there is a very real possibility of antibody escape mutants arising in EBOV GP. A recent study found that none of the genetic changes have altered the function of the virus [110]. However, they did not consider interactions with antibodies or implications of unobserved mutations.

The purpose of this study is to initiate a watch list of potential antibody escape mutants for the EBOV GP. We focus on the KZ52 antibody as it is one of the few with an available structure bound to EBOV GP. KZ52 has virus neutralization activity in vitro and protects guinea pigs from EBOV disease [95, 117]. Although KZ52 does not protect non-human primates from EBOV disease [112], it was originally isolated from the blood of human EBOV survivors (Maruyama et al., 1999). Using the experimental structure of the Zaire EBOV GP bound to antibody KZ52 (Fig. 1) [81], we performed molecular modeling to estimate the folding and binding stabilities for every possible amino acid mutation of GP. Our approach is general and could be applied to other EBOV epitopes, or other viruses, as experimental structures become available. We emphasize from the outset that this is an in silico study aimed at identifying mutations with an increased risk of escaping immune response; our intention is to provoke experimental research on evolutionary escape in both Ebola and other viral pathogens.

2.2 Methods

2.2.1 Overview

To initiate a watch list for the Ebola virus (EBOV) glycoprotein (GP) it is necessary to determine how amino acid mutations modify stabilities for GP folding, forming a trimer and binding to the KZ52 antibody. That is, we need to calculate $\Delta\Delta G$ values for binding and folding. Ideally, these calculations would be performed using a statistical-mechanics-based method such as we have done previously [82, 167].

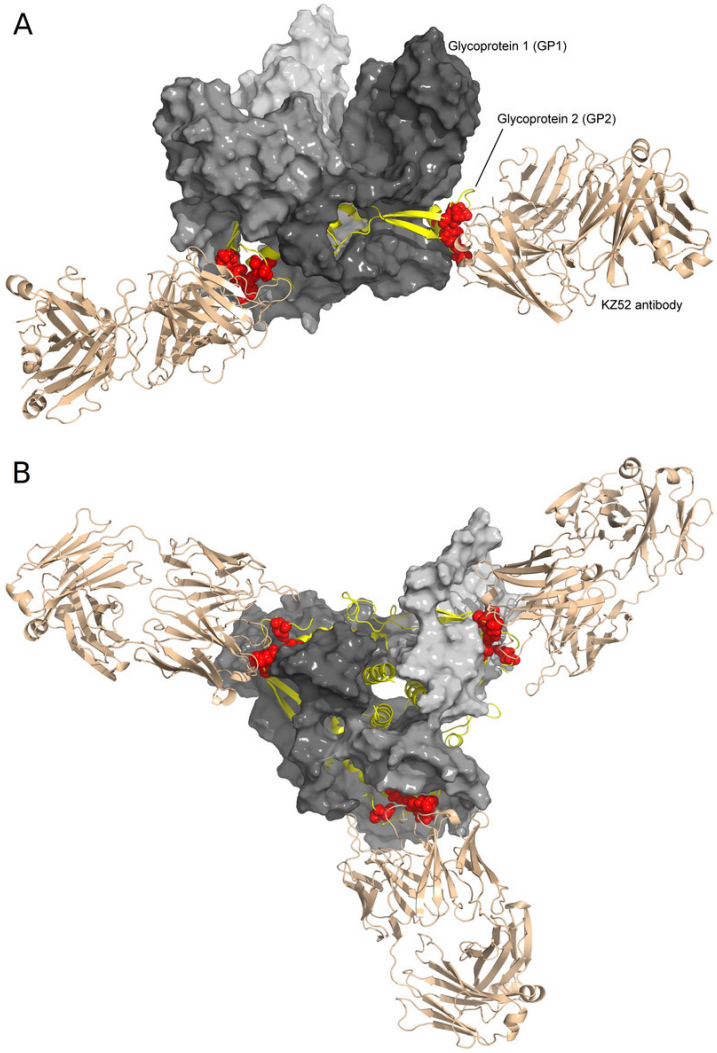


Figure 2.1: Structure of Ebola glycoprotein trimer (GP1, gray; GP2, yellow) in complex with the KZ52 antibody as viewed from the side (A) and the bottom (B). GP1 is in gray, GP2 is in yellow and the structure is after 10 ns of MD simulation. The six watch list sites that are predicted to contain antibody escape mutants are shown as red spheres and are all located in GP2 (Table 2.2, Fig. 2.2.)

However, such methods are computationally expensive and are not feasible for the current study where it was necessary to calculate 25,840 values of $\Delta\Delta G$ (340 residues $\times$ 19 possible mutations to other residues $\times$ 4 types of stability calculations). Instead, we decided to use a semi-empirical method for calculating $\Delta\Delta G$ values. Because online-only software was not practical given the large number of mutations, we chose to use the software FoldX [129, 46]. FoldX can be run in parallel on a computer cluster since the binary is available.

We hypothesized that because protein structures are not static, improvements in $\Delta\Delta G$ estimation might be achieved by using molecular dynamics simulation to sample the configurational space for the

proteins and then analyze snapshots from these simulations in FoldX. We selected 20 test systems (10 folding and 10 binding) to assess whether this strategy improves estimation of experimental stability data. In the Supplemental Information, we describe our criteria for selecting test systems and then show that using 100 molecular dynamics snapshots and averaging the FoldX results provides more accurate estimates of $\Delta\Delta G$ as compared to using FoldX on a single experimental structure. The molecular dynamics plus FoldX methodology we used on the test systems was identically applied to the Ebola system. After explaining how structures were prepared and arranged, we describe this methodology in the subsections below.

2.2.2 Stability preparation

Preparation of the test system structures is described in the Supplemental Information. For EBOV GP, the amino acid sequence was based on the 1976 Mayinga strain obtained from GenBank accession number AF086833. We downloaded PDB accession number 3CSY as our template structure. The file 3csy.pdb was modified to remove all but one copy each of GP1, GP2, antibody light chain and antibody heavy chain (one third of the GP-KZ52 trimeric complex). SWISS-MODEL was then used to generate structures for each of the four chains using 3csy.pdb as a template [4]. The experimental structure 3csy has missing residues 190-213 that are predicted to be intrinsically disordered but SWISS-MODEL incorrectly generated helical structures for these residues. Thus, we removed residues 190-213 from the SWISS-MODEL structure and used MODELLER to rebuild the coordinates of the missing residues [125]. The resulting structure had no secondary structure content in residues 190-213. The full trimeric complex was then created using the symexp command in PyMOL. The final trimer structure (see Fig. 1) contains three copies each of residues 32-276 for GP1, residues 503-597 for GP2, residues 1-225 for KZ52 heavy chain and residues 1-216 for KZ52 light chain.

2.2.3 System configuration

Arrangement of the test systems is described in the Supplemental Information. EBOV GP was configured as four systems: (i) unbound GP1, (ii) unbound GP2, (iii) trimer consisting of three copies of GP1 and GP2 and (iv) antibody complex consisting of three copies each of GP1, GP2 and the KZ52 antibody. Snapshots from systems (i) and (ii) were used to estimate mutational effects on folding stability of the unbound proteins GP1 and GP2, respectively. Snapshots from (iii) were used to estimate the affinity of GP1–GP2 (dimer bind). This was done by calculating the affinity for all three copies of GP1 binding to GP2 and then dividing this value by three. Snapshots from (iii) were also used to estimate the affinity for GP1–GP2 dimers binding to one another (trimer bind). This was done by calculating the affinity for

one GP1–GP2 dimer binding to the other two dimers. Finally, snapshots from (iv) were used to estimate the GP-KZ52 affinity by calculating the affinity of all of the GP1–GP2 dimers to their corresponding KZ52 antibodies and dividing this value by three.

2.2.3.1 Molecular dynamics simulations

The software package GROMACS 5.0.3 was used for all MD simulations with the Charmm22* force-field [52]. The system was placed in a dodecahedral box of TIP3P water and given neutral charge by adding Na+ and Cl- ions at a concentration of 0.15 mol/L. Each system was then minimized using steepest decent for 1,000 steps. To allow for some equilibration of the water around the proteins, each system was then simulated for 1 ns with the positions of all heavy atoms in the complex harmonically restrained, and then simulated for another 1 ns with no restraints. During the restrained simulations the temperature of the system was increased linearly from 100 K to 300 K for the test systems and to 310 K for the EBOV GP systems and the pressure was maintained at 1 atm using the Berendsen algorithm. Production simulations for each system were then carried out for 100 ns with pressure maintained using Parrinello-Rahman coupling. For all simulations, the LINCS algorithm was used to constrain all bonds to their ideal lengths and virtual sites were used allowing the use of a 5 fs timestep. The temperature was controlled using the v-rescale option. Particle mesh Ewald was used for electrostatics with a real-space cutoff of 1.2 nm. Van der Waals interactions were cut off at 1.2 nm with the Potential-shift-Verlet method for smoothing interactions. During the 100 ns production simulation snapshots were saved every 100 ps giving 100 snapshots for each system.

2.2.3.2 FoldX

Each of the 100 snapshots captured during MD simulations was then analyzed using FoldX [129, 46]. We initially minimized structures six times in succession using the RepairPDB command to obtain convergence of the potential energy. All single amino acid mutations were then generated using BuildModel. Finally, protein folding stabilities were estimated using Stability on the monomer structures and binding stabilities were estimated using AnalyseComplex on the protein complexes. For each mutation we then estimated $\Delta\Delta G$ by averaging across all 100 individual snapshots estimates.

2.2.3.3 Thresholds for functionality and antibody disruption

o define the range of stability change where the GP protein is likely to remain functional, we began by noting that in previous work on the bacteriophage ϕX174 [102], 77 of 79 (97.5%) of observed functional mutations have estimated stability effects on both folding and binding in the range $-2.5 < \Delta\Delta G < 2.5$ kcal/mol. The large amount of available Ebola sequences allows us to survey a set of presumably

functional mutations in Ebola and ask how many of these are categorized as functional vs non-functional using this preliminary criteria. We downloaded 922 sequences from the NCBI Virus Variation Ebolavirus Database on August 20, 2015 [18, 104] (Species = Zaire ebolavirus, Host = Any, Region = Any, Genome Region = Spike glycoprotein). To this set we appended 39 sequences from Leroy et al. [85] and Wittmann et al. [162] along with the two escape mutations described in Qiu et al. [121]. We compared all 963 sequences to our reference sequence, GenBank Accession AF086833, and thereby identified 41 mutational differences (Table 1) within the structured regions modeled here. Four of the 41 mutations (9.8%) have a functional stability effect (i.e., $\Delta\Delta G$ for monomer folding, dimer binding or trimer binding) that falls outside the ± 2.5 zone. Because our objective is to limit the rate of false exclusions to $\leq 5\%$, we expanded the functional zone to ± 3.0. This shifts two of the mutations back into the functional zone, leaving 2 of 41 (4.9%) predicted to be non-functional.

It is worth noting that the observed incidence of two false exclusions in a sample of 41 is consistent with our method having predictive power to distinguish functional from non-functional mutations. Of the 6,460 possible mutations for GP, our method categorizes 5,303 (82.1%) as functional and 1,157 (17.9%) as non-functional. If our method lacked predictive power we would expect a random sample of 41 mutations to contain 33.7 functional and 7.3 non-functional mutants. The binomial probability that such a random sample would contain ≤ 2 non-functional proteins by chance is 0.018. Unfortunately, because we lack a list of known non-functional mutations, we cannot perform the converse test and ask what proportion of non-functional mutations does our method correctly identify as such.

How sensitive is the size of the watch list to the rate of false exclusions? The following argument suggests that even if the false exclusion rate could be reduced to zero, it would have a very small effect on the watch list. The application of a functional zone between ± 3.0 kcal/mol along with an antibody disruption criteria of $\Delta\Delta G > 2.0$ kcal/mol leads to a watch list of 34 mutations. Of the 6,460 possible mutations, our method categorizes 1,157 as non-functional. If 5% of these are actually functional, it suggests that we have omitted approximately $6{,}460(0.05) = 58$ mutations from the set of functional mutations. However, very few of these would likely disrupt antibody binding. Among all 6,460 mutations, 66 (or $\approx 1\%$) are identified as disrupting antibody binding. Assuming false exclusion is independent of antibody disruption, we would expect that $58(0.01) = 0.6$, or less than one mutation being falsely omitted from the watch list.

2.3 Results and Discussion

We identified potential antibody escape mutations for the watch list by considering every possible GP mutation and finding those that disrupt binding between GP and KZ52 but do not disrupt the ability of

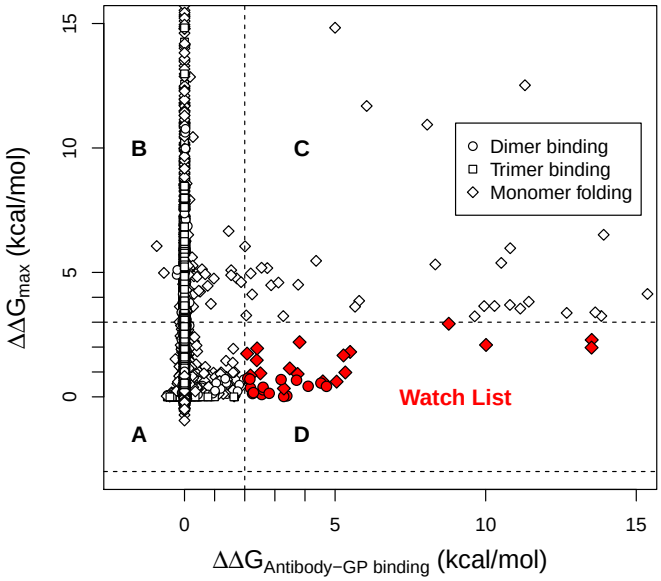


Figure 2.2: Watch list mutations are those that disrupt KZ52 antibody binding but not GP folding and trimer formation. For each possible GP mutation, only the maximum of folding stability, dimer binding stability (interaction of GP1 and GP2) or trimer binding stability (interaction of a GP1-GP2 dimer with other dimers) is plotted on the y-axis. Symbols in the inset legend indicate which of the three is plotted. The GP-KZ52 binding affinity is plotted on the x-axis. Mutations with x-axis values $-3 < \Delta\Delta G < 3$ kcal/mol are considered functional since they are likely to retain the ability to fold and form trimers (regions A and D). Mutations with y-axis values $\Delta\Delta G > 2$ kcal/mol have the potential to disrupt antibody binding (regions C and D). The watch list mutations (region D) are those that are likely to be both functional and disrupt antibody binding. The reasoning behind using a different cutoff for functional as compared to antibody binding is described in the main text.

GP to fold and form a complex. The GP protein is cleaved into two subunits, GP1 and GP2, and the final structure is a trimer consisting of three GP1–GP2 dimers (Fig 2.1). We used a combination of molecular dynamics and FoldX [129, 46] because preliminary analysis of 20 test systems showed that combining these methods improved our ability to predict experimental results (see Supplemental Information). To our knowledge, this method has not been used in previous studies.

Our conceptual approach to creating a watch list is to identify mutations that are both functional and disrupt antibody binding. We therefore sought to remove mutations that are non-functional and, from those that remain, identify the ones that disrupt antibody binding. The function of GP is to mediate viral entry into the cell. There are multiple ways mutation can disrupt this function. For example, studies have shown that mutations in GP can reduce infectivity [61, 157, 30], transduction and host cell binding [34, 17]. Another way to be non-functional is for a mutation to render GP unable to fold and

bind together to form a stable complex. Here we focus on this stability aspect of functionality and remove those mutations our model predicts will not fold or form a complex. It is important to appreciate that our approach is conservatively inclusive: if we could remove all non-functional mutations instead of the subset identified as unstable, the watch list would be reduced in size.

Identifying mutations that disrupt antibody binding but not the ability to fold and bind into a functional complex requires defining thresholds on changes in stability ($\Delta\Delta G$) for both criteria. These criteria should be conservative to reduce exclusion of mutations that could compromise treatment efficacy from the watch list. For functionality, previous work on a coat protein in a different virus [102] indicated that the stability effect of virtually all observed mutations is in the range of $-2.5 < \Delta\Delta G < 2.5$ kcal/mol. To determine if those criteria also hold for EBOV GP, we compared 963 available sequences of GP, identified 41 mutations in the structured regions that have arisen in natural or lab populations, and found that four of the 41 (9.8%) were classified as non-functional. To be conservative, we expanded the functional zone to $-3.0 < \Delta\Delta G < 3.0$ kcal/mol. This functional threshold is more inclusive and reduces our error rate to below 5%: two of the 41 mutations (4.9%) are falsely classified as non-functional (Table 1). As we reason in the Methods, even if the false exclusion rate could be driven to zero, we expect it would change our watch list very little. For disruption of antibody binding, we used a threshold of $\Delta\Delta G > 2.0$ kcal/mol. This was based on refining our preliminary threshold by 0.5 kcal/mol, but in the opposite direction so as to be more inclusive. The implications of this threshold choice and alternatives to it will be discussed below.

Figure 2.2 provides a graphical illustration of how mutations were selected to be on the watch list. The maximum functional stability for all mutations is plotted against the corresponding change in the antibody binding affinity. The 34 mutations in the lower right quadrant are those that belong on the watch list since they are classified as both functional and disruptive to antibody binding. The specific mutations on the watch list are given in Table 2.2. If any of the mutations in this table appear in a real population, it indicates an increased risk of escaping the normal immune response. One of these mutations (N550K) has already appeared in humans thought to have been infected by gorillas in Central Africa between 2001 and 2003 [85]. This mutation is present in all sequenced isolates from that outbreak.

In contrast to constructing a simple list of all possible mutations near an epitope, the watch list in Table 2.2 is quite specific. The 34 watch list mutations are concentrated at just six residues and all of these lie at the interface between GP2 and KZ52, as one might intuitively expect from the structure (Fig. 2.1). Yet, most mutations of GP sites that are within four angstroms of the KZ52 antibody are not predicted to disrupt antibody binding. Only six of the 23 (26%) interface sites and 34 of 437 (7.8%) of the possible mutations at these sites are on the watch list.

In order to facilitate use of other possible criteria and/or thresholds, the Supplemental Information includes a sortable and searchable spreadsheet with all 6,460 mutations of GP. We provide this spreadsheet because we recognize that the relationship between antibody binding affinity and the ability of the antibody to neutralize EBOV is not well understood. Work in influenza suggests that as affinity decreases, the ability of an antibody to neutralize a virus decreases rapidly and in a non-linear fashion [76]). This makes intuitive sense because the relationship between the change in affinity and the change in the ratio of bound to unbound antibody is nonlinear; $\Delta\Delta G$ values of 1.0, 1.5 and 2.0 kcal/mol correspond to changing the ratio from 19 to 8.1 to 3.5% of its original value. The size of the watch list depends on how we define the threshold for antibody binding (Fig. 2.2). If the threshold is lowered from 2.0 to 1.5 to 1.0, the watch list grows from 34 to 49 to 73 mutations. This highlights the need for more experimental studies that assess how disrupting antibody binding influences immune response.

Davidson et al. [30] recently conducted an alanine-scanning mutagenesis study on GP that can be qualitatively compared to our work. Specifically, they individually mutated each residue of the GP protein to alanine and measured changes in GP-KZ52 binding affinities relative to the unmutated form. They identified five residues that are critical for KZ52 antibody binding: C511, N550, D552, G553, and C556. Three of these sites are found on our watch list in Table 2.2 (N550, D552, and G553) and 25 of the 34 (74%) watch list mutations are found at these three sites. For the two other critical residues identified by Davidson et al. [30] (C511 and C556), our results agree that antibody binding is disrupted by mutations at these sites, but we estimate that folding is also disrupted, and hence the exclusion from the watch list. If we ignore our criteria that mutations do not disrupt folding stability or the formation of dimers and trimers, we identify eight residues where at least one mutation will disrupt KZ52 antibody binding: N506, C511, P513, N550, D552, G553, C556, G557 (all individual mutations can be obtained from the spreadsheet in the Supplemental Information). Overall, we conclude that our results are generally consistent with the findings of Davidson et al. [30].

The watch list remains incomplete and putative for several reasons. First, although our list was generated for one EBOV epitope and its interactions with the KZ52 antibody, it is known that there are multiple epitopes (Fig. 2.3). Indeed, a recent study found mutations of a conserved threonine in the EBOV mucin-like domain that is required for protection by the 14G7 antibody [115]. This highlights the need for more experimental structures of antibodies interacting with viral proteins. With more experimental structures, it would be possible to expand the watch list to incorporate more epitopes. Second, the watch list only includes substitutions that are predicted to individually disrupt antibody binding while remaining functional. It is alternatively possible that immune escape could arise by the accumulation of several changes, each of modest stability effect but with a large cumulative effect on

antibody binding. How multiple substitutions interact to produce cumulative effects on stability is not well understood and is an important consideration for future studies. Third, the watch list has not been experimental validated (except in its general consistency with the work of Davidson et al. [30]) either in terms of mutational effects on GP folding and binding affinities, nor on the downstream immune system consequences. Our hope is that this work will motivate such research.

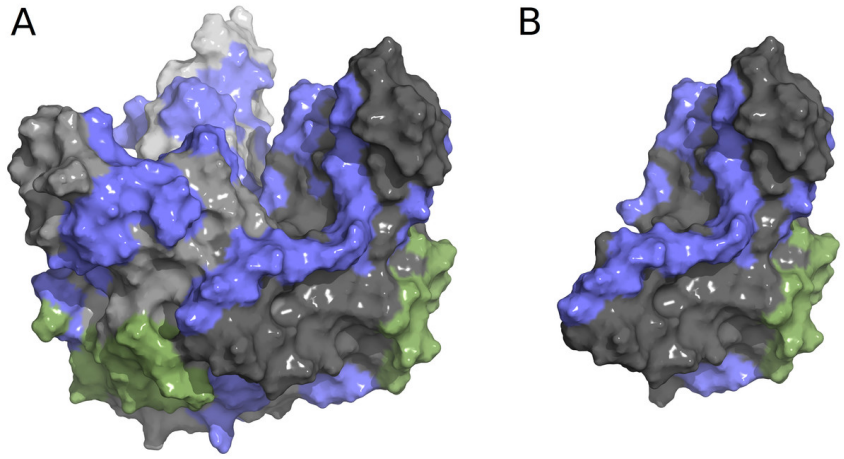


Figure 2.3: Structure of Ebola GP1-GP2 trimer complex (A) and individual GP1-GP2 dimer (B) with structural epitopes from KZ52 and other known linear epitopes. KZ52 is in green, other known linear epitopes are in blue [8]. The watch list generated for the current study is for the green region only, since structures are required for the method used, highlighting the need for more experimental structures of Ebola with antibodies.

In summary, we have initiated a watch list of potential antibody escape mutations of EBOV by considering the interactions between GP and antibody KZ52. This initial watch list contains 34 mutations in six sites in GP2, and one of these mutations (N550K) was seen in humans in a previous outbreak. We believe initiating a watch list is an important first step to predicting how the evolution of EBOV could undermine treatment efforts. Our intention is that the watch list motives experimental research testing the strategy we have employed. This study further emphasizes the need for more experimental structures of antibodies interacting with EBOV in order to produce a comprehensive watch list. We highlight the need for ongoing monitoring of EBOV sequences in human outbreaks. If mutations on the watch list appear in human populations infected by EBOV, treatment with vaccines or antibody therapies may be compromised. Furthermore, if mutations from the watch list arise and increase in frequency within an immunized population, it would suggest that the virus is responding to selective pressure exerted by the vaccine. Monitoring will be much more powerful as the watch list is expanded and experimentally validated. Finally, we suggest that the approach used here is general and could be applied to other viruses for which experimental structures are available.

2.4 Acknowledgments

Grant support for this research was provided by the National Science Foundation (DEB1521049) and the Center for Modeling Complex Interactions sponsored by the National Institutes of Health (P20 GM104420). Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (P30 GM103324). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Mutation	Antibody bind[a]	Dimer bind[b]	Trimer bind[c]	Monomer fold[d]
N107D	0.00	-0.09	0.00	1.31
L111F	0.00	0.00	0.00	2.31
I129V	0.00	0.05	0.02	0.77
D150A	0.00	0.00	0.01	0.6
D163N	0.00	1.12	0.57	0.39
I170L	0.00	0.01	0.02	2.31
I170F	0.00	0.03	0.05	16.48
V181I	0.00	0.05	0.00	-0.73
T206M	-0.30	-0.33	0.01	-0.14
G212D	0.00	0.19	-0.04	-0.11
Y213H	0.00	0.41	-0.01	1.27
Y214H	0.00	-0.01	0.00	0.18
T216P	0.00	-0.01	0.00	2.27
R219K	0.00	0.00	0.00	0.00
A222V	0.00	0.00	0.00	-0.11
E229K	0.00	0.00	0.00	-0.17
T230A	0.00	0.00	0.00	0.62
T240N	0.00	0.00	0.00	0.86
S246P	0.00	0.00	0.00	-1.11
L254I	0.00	0.00	0.00	0.81
L254V	0.00	0.00	0.00	1.39
Q255R	0.00	0.00	0.00	0.1
I260R	0.00	0.00	0.00	1.78
T262A	0.00	0.00	0.00	-0.08
W275L	0.00	0.00	0.00	0.09
A503V	-0.17	0.09	0.00	0.1
Q508R	0.16	-0.03	0.00	0.54
Y517C	0.01	0.26	0.01	1.38
G524D	-0.01	0.14	2.31	2.12
A526T	0.00	0.02	0.56	0.80
I527T	0.00	0.18	0.15	1.04
P537L	0.00	0.23	0.42	0.53
I544T	0.00	0.36	-0.01	0.47
E545D	0.00	0.5	0.00	0.46
N550K	4.59	0.01	0.00	0.62
D552N	1.76	0.23	0.00	0.13
A562D	-0.06	2.98	0.02	0.67
L571R	0.00	0.05	2.34	0.27
L573R	0.00	2.78	-0.25	1.30
W597F	0.00	0.07	0.48	-0.11
W597C	0.00	0.35	3.51	0.26

Table 2.1: Model predicted effects on stability of 41 observed mutations in EBOV GP. The one observed mutation that is also on the watch list is indicated in red. The two mutations that our methods falsely excludes as non-functional are indicated in blue. All numerical entries are $\Delta\Delta G$ values in units of kcal/mol.

GP2 mutation[a]	Antibody bind[b]	Dimer bind[c]	Monomer fold[d]
N506W	3.40	0.04	-0.41
N506Y	2.56	0.09	-0.55
P513H	2.52	0.01	0.95
P513W	2.19	0.01	0.86
N550Q	3.76	0.01	0.92
N550K	4.59	0.01	0.62
N550P	3.82	0.14	2.20
N550F	10.01	0.03	2.09
N550H	5.50	0.03	1.81
N550I	5.28	0.02	1.66
N550E	3.49	-0.04	1.14
N550R	5.34	-0.03	0.98
N550W	13.52	0.02	2.29
N550V	2.08	0.02	1.74
N550Y	13.52	0.04	1.98
N550M	3.29	0.02	-0.15
D552S	2.10	0.75	0.33
D552Q	2.19	0.36	0.29
D552K	2.61	0.28	0.16
D552T	2.40	1.01	1.47
D552F	4.11	0.43	0.14
D552A	2.17	0.71	0.48
D552H	4.53	0.55	0.29
D552G	2.61	0.39	0.01
D552R	3.30	0.26	0.34
D552W	5.05	0.41	0.61
D552V	2.41	0.75	1.95
D552Y	4.71	0.42	0.13
G553M	8.77	-0.01	2.94
G557F	2.26	0.13	-1.34
G557H	3.72	0.67	-0.05
G557R	2.29	0.17	-0.62
G557W	3.21	0.69	-1.32
G557Y	2.81	0.14	-1.19

Table 2.2: Watch list mutations and their effects on stability. All numerical entries are $\Delta\Delta G$ values in units of kcal/mol. a) The 34 mutations are distributed among six sites in GP2. b) Binding affinity between GP and the KZ52 antibody. c) Binding affinity between GP1 and GP2. d) Folding stability for GP2. Binding affinity results for forming the GP trimer are all zero and are not shown. Note: [a]Binding affinity between GP and the KZ52 antibody. [b]Binding affinity between GP1 and GP2. [c]Binding affinity between three GP1-GP2 dimers. [d]Folding stability for GP2.

Chapter 3: Protein Stability in Titan's Subsurface Water Ocean

Kyle P. Martin,[1,2] Shannon M. MacKenzie,[1] Jason W. Barnes,[1] F. Marty Ytreberg,[1,2]

[1]Department of Physics, University of Idaho, [2]Institute for Modeling Complex Interactions, University of Idaho

Published in Astrobiology. As first author, I performed all molecular dynamics simulations and analyses, and was the main writer and editor on the manuscript. Molecular dynamics simulations involved setting up the model systems, writing scripts, and utilizing the IBEST computer cluster. Analyses were performed on the molecular modeling results and plots were generated via Python scripts. Writing was done in collaboration primarily on Google Docs. I was also in charge of uploading all materials to the journal.

3.1 Introduction

Despite the lack of sunlight, extreme pressures, and high temperatures, chemotrophic bacteria take advantage of the products of water-rock interactions to survive at Earth's deep-sea vents [108]. Earth may not be the only body to host such habitable niches. Subsurface oceans are considered confirmed [50] on the jovian moons Ganymede, Callisto, and Europa [74, 24, 73, 59, 60, 21] and the saturnian moons Enceladus and Titan [59, 60]. Hidden under ice crusts tens to hundreds of kilometers thick, the liquid water environments of these ocean worlds interact with rocky cores, albeit to different extents (Sohl et al.,2010), making them excellent targets in the search for life elsewhere in the Solar System due to the potential confluence of liquid water, chemical building blocks, and energy sources [90, 50].

Titan's subterranean oceans may be seeded with the hydrogen and carbon necessary for Earth-like biochemistry. Evidence for this possibility comes from measurements of Titan's atmosphere. Nitrogen and methane dominate the bulk composition at 98% and 1.8% respectively, with a multitude of trace gases probably including some oxygen species [28, 53]. Ultraviolet photons from the Sun, energetic particles from Saturn's magnetosphere, and galactic cosmic rays drive photolysis of methane and nitrogen throughout Titan's atmosphere, leading to various chemical reactions that create a variety of complex organic compounds [155, 29, 154, 107]. Eventually, the products of this chemistry form haze particles about a micron in size [145] that fall to the surface, blanketing Titan's terrain in an organic-rich layer

estimated at over 2 × 105 km3 [89]. However, the current rate of photolysis could not have been sustained over the lifetime of the Solar System without exhausting the current atmospheric inventory of methane. Forward modeling suggests that all the methane in Titan's atmosphere today would be consumed within 30 Myr [166, 160]. One way to bala nce this consumption with the 1 Gyr age suggested by the 12C/13C ratio observed by both Cassini [93, 109] and Huygens [107] is through continuous or episodic replenishing of methane from Titan's interior [144, 25]. In such a scenario, methane from the rocky interior dissolves in the subsurface ocean before outgassing [91] and could be delivered to the surface via cyrovolcanoes [143]. The possibility of organics going down into the ocean is compelling—especially given the evidence that amino acids can quickly form when liquid water interacts with lab analogs of Titan's haze [106, 105]—but beyond the scope of this paper and will be left to future research. Additionally, Titan's subsurface ocean is thought to contain ammonia to ensure the ocean remains liquid over the age of Titan [27, 43, 38]. Thus, Titan's subsurface ocean may contain hydrogen, carbon, nitrogen. (The presence of ammonia in Titan's subsurface ocean should not prevent Earth-like biochemistry, as terrestrial extremophiles can thrive in high pH [9.0–11.6] solvents [120].)

At 5150 km in diameter and with a bulk density of 1881 kg/m3, a high-pressure ice (ice VI) likely separates Titan's ocean from the rocky core today [91, 138, 44]; however, this layer may not inhibit water-rock interactions. Journaux et al., for example, demonstrated how brines at certain high temperatures and pressures can be transported through an ice VI layer to the ocean [67, 151]. It is also possible that earlier in Titan's history the ocean may have both been in direct contact with the rocky core (facilitating hydrothermal reactions) as well as received exogenic cometary and chondritic material from meteorite falls and impacts [130]. Fortes [37] outlines several similarities between Titan's evolving ocean environment and extreme environments on Earth.

Previous work has been done to look at the effects of pressure and temperature on protein flexibility [126, 70, 57] and adaptation [23, 131, 101]. Piezophiles, organisms that live in high pressures, are not well understood, but the high pressure limit of known life is around 1.1 kbar [133, 100, 58]. Alkaliphiles, organisms that typically live in pH values between 9 and 12, can thrive in a variety of environments as long as the basicity is optimum [72, 78, 79]. Combining the traits of these two extremophiles would result in an organism able to live and thrive in an environment similar to that of the subsurface ocean of Titan.

In this work, we build on the hypotheses of Fortes [37]. Instead of answering whether molecules can form in Titan's ocean (today or in some earlier epoch), we explore how biologically relevant molecules might behave in Titan's ocean. Some studies have been done that show the folding of proteins in different organic solvents can lower folding free energy [165] or can function similarly to water in an aqueous-organic solvent [136]. This study does not analyze the folding or binding of proteins but rather the integrity of

a protein over the course of a simulation. In Section 2, we describe the molecular dynamics simulations that we used to study Earth-based proteins in both Earth and Titan-like conditions. In Section 3, we analyze the results, measuring protein compactness, flexibility, and backbone dihedral angle distributions. Section 4 discusses the implications of the results for Titan and other Ocean Worlds, and we conclude in Section 5.

3.2 Materials and Methods

3.2.1 Protein folding

Inside biological cells on Earth, ribosomes manufacture proteins by sequentially adding individual amino acids in a chain according to instructions in messenger RNA as copied from the cell's nuclear DNA. Knowledge of the amino acid sequence alone does not allow us to directly infer the resulting protein's chemical behavior, however. After their manufacture, proteins fold in on themselves in a manner governed by their own self-affinity and electrical interactions with the surrounding solvent. Therefore, in Titan's high-pressure water-ammonia subsurface ocean proteins might behave differently than they do in biological systems on Earth.

Those different shapes and the variability thereof would necessarily alter the functionality of the protein. We see similar situations in extreme environments on Earth, where, for instance, high-temperature single-celled organisms living in deep-sea vents have evolved distinctly different versions of common proteins that specifically tailor their amino acid sequence to match their environment (at high temperature, for instance, deletions of large swaths of protein helps those proteins behave similarly to non-deleted ones at room temperature [6]). In that high-temperature situation, proteins typically work better with fewer large ancillary residues so as to not have excessive conformal variation with high internal kinetic energy.

As a first step toward understanding how the conditions in Titan's subsurface ocean affect protein folding and movement, we simulate the behavior of representative proteins computationally. The goal of these initial calculations is not to design an alien biochemistry compatible with Titan's ocean but rather to broadly determine the effects of aspects of that ocean such as high pressure and ammonia content.

3.2.2 Protein selections

Three types of proteins were chosen to simulate the two most common secondary structures on Earth: alpha helices and beta sheets [10]. These proteins were chosen from the CATH protein database [132, 80] by selecting the most common motif of that secondary structure. For example, one of our chosen motifs containing primarily alpha helices was selected by using the following steps: (1) Opened the website http://www.cathdb.info/browse/tree. (2) Chose "1 Mainly Alpha." (3) Chose the largest number of folds,

"1.10 Orthogonal Bundle." (4) Chose the largest number of superfamilies, "1.10.287 Helix Hairpins." (5) Chose the largest number of Domains, "1.10.287.610 Helix hairpin bin." (6) Chose example protein , "3uq8." (7) Looked at sequence to ensure that chain A amino acids 3–61 is the appropriate fragment. (8) Downloaded protein structure from the RCSB protein data bank and edited file to include only the amino acids that are within the appropriate fragment. The proteins chosen were 3uq8 and 1xmk (mainly alpha), 3ulj and 4unu (mainly beta), and 4g1q and 2gxq (mixed alpha/beta).

3.2.3 System preparation

The idea is to create a system that replicates an Earth environment at sea level and a Titan environment at the bottom of its subsurface ocean. We chose a practical temperature of 300 K and a pressure of 1 bar. For Titan we also chose 300 K. This decision was made based on an assumption that there would be hydrothermal vents, noting that a similar assumption has been made about Titan's neighbor, Enceladus [54]. The pressure chosen for the Titan environment was 1000 bar. This would correspond to a depth of about 10 km below Earth's ocean's surface. On Titan, using an aqueous ammonia solution density of 0.89 g/cm3 [111] and a gravitational constant of 1.35 m/s2 [64], we would get a pressure of 1 kbar at a depth of about 80 km below Titan's surface. This does not consider several variables such as ice thickness, ice density, or compressibility. However, this shows that a plausible pressure and temperature were used for the Titan environment.

Therefore, setting up the simulations, the Earth environment was set to a temperature of 300 K, pressure of 1 bar, and pure water. The deep Titan ocean environment used the same temperature as Earth but with 3 orders of magnitude larger pressure (1000 bar) and a eutectic water-ammonia mixture [37]. The three protein types were separately placed in each environment for a total of six simulations.

3.2.4 Molecular dynamics simulations

The software package GROMACS 5.1.2 was used for all molecular dynamics simulations with the Amber99sb*-ildnp forcefield [52]. The Earth system was placed in a dodecahedral box of TIP3P water [66]. A TIP3P water model was chosen, as it is a good balance of accuracy and computational efficiency. The Titan system was placed in a dodecahedral box of a eutectic mixture of TIP3P water and 32% weight ammonia at a density of 0.89 g/cm3 [37]. Each system's energy was minimized by using steepest descent for 1000 steps. To allow for some equilibration of the water around the proteins, each system was then simulated for 1 ns with the positions of all heavy atoms in the complex restrained via a harmonic potential, and then simulated for another 1 ns with no restraints. During the restrained simulations, the temperature of the system was increased linearly from 100 to 300 K, and the pressure was maintained

at 1 atm using the Berendsen algorithm. Production simulations for each system were then carried out for 100 ns with pressure maintained using Parrinello-Rahman coupling. For all simulations, the LINCS algorithm [51] was used to constrain all bonds to their ideal lengths allowing for a timestep of 2 fs. The temperature was controlled using the v-rescale option. Reaction-Field-zero was used for electrostatics with a real-space cutoff of 1.2 nm. The Van der Waals interaction cutoff was set to 1.2 nm with the Potential-shift-Verlet method for smoothing interactions.

3.2.5 Analyzing data

Radius of gyration of a protein is a measure of its compactness; smaller values correspond to more compact protein shape. The radius of gyration measurements were obtained using the GROMACS command gmx gyrate, and histograms were then generated using python. The root-mean-square fluctuation (RMSF) is a measure of how much an atom fluctuates about its average position, that is, the secondary structure flexibility. RMSF plots were obtained by running the GROMACS command gmx rmsf. RMSF values were converted to log form and used to color the protein using PyMOL 1.7 (blue for low RMSF values and red for high). Secondary structures plots as a function of simulation time were generated using the GROMACS command gmx do_dssp [98, 147]. Ramachandran plots show the distribution of backbone dihedrals that largely determine a protein's conformation and secondary structure preference during simulation. These results were generated using the GROMACS command gmx rama, and histograms were generated by subtracting the Earth measurements from the Titan measurements.

3.3 Results

Radius of gyration histograms are shown in fig. 3.1 and illustrate the compactness of the proteins. The peaks of each plot represent the most likely radius of gyration value for each protein in each environment. A difference in peak location represents either a more or a less compact protein (smaller radius of gyration is more compact). Our results show that proteins in the Titan environment experience a shift in the peaks toward a lower radius of gyration value in both the alpha and mixed alpha/beta as compared to the Earth environment. The beta shows no or negligible shifting of its peaks. The shifted plateau in the mixed alpha/beta 4g1q Earth environment is caused by the C-terminus of the protein moving to a new conformation.

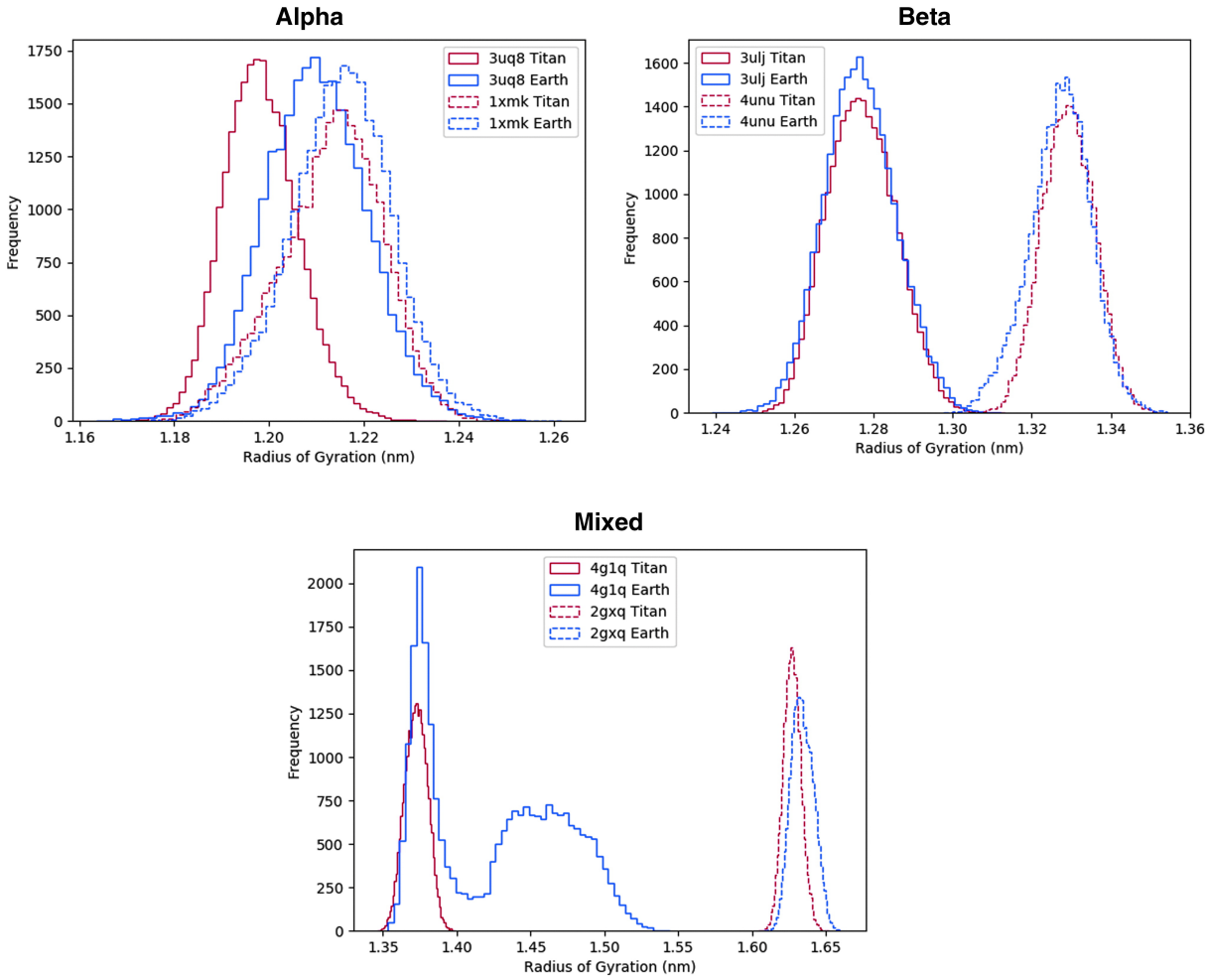


Figure 3.1: Radius of gyration for all proteins over the simulation. A shift to the left is a more compact protein, while a shift to right is less compact. The Titan environment experiences a shift in its peaks toward a lower radius of gyration value in both the alpha and mixed alpha/beta compared to the Earth environment. The beta shows no or negligible shifting of its peaks. The shifted plateau in the mixed alpha/beta 4g1q Earth environment is caused by the C-terminus of the protein moving to a new conformation.

Root-mean-square fluctuation results are in fig. 3.2 and show how much atoms fluctuate about their average position. The RMSF value was converted to a log form and overlaid onto a frame from the largest cluster of the simulation. The gradient ranges from blue to white to red, with blue being most stable and red being least stable. Proteins in the Titan environment experience larger maximum RMSF values as compared to the Earth environment but lower RMSF values on average.

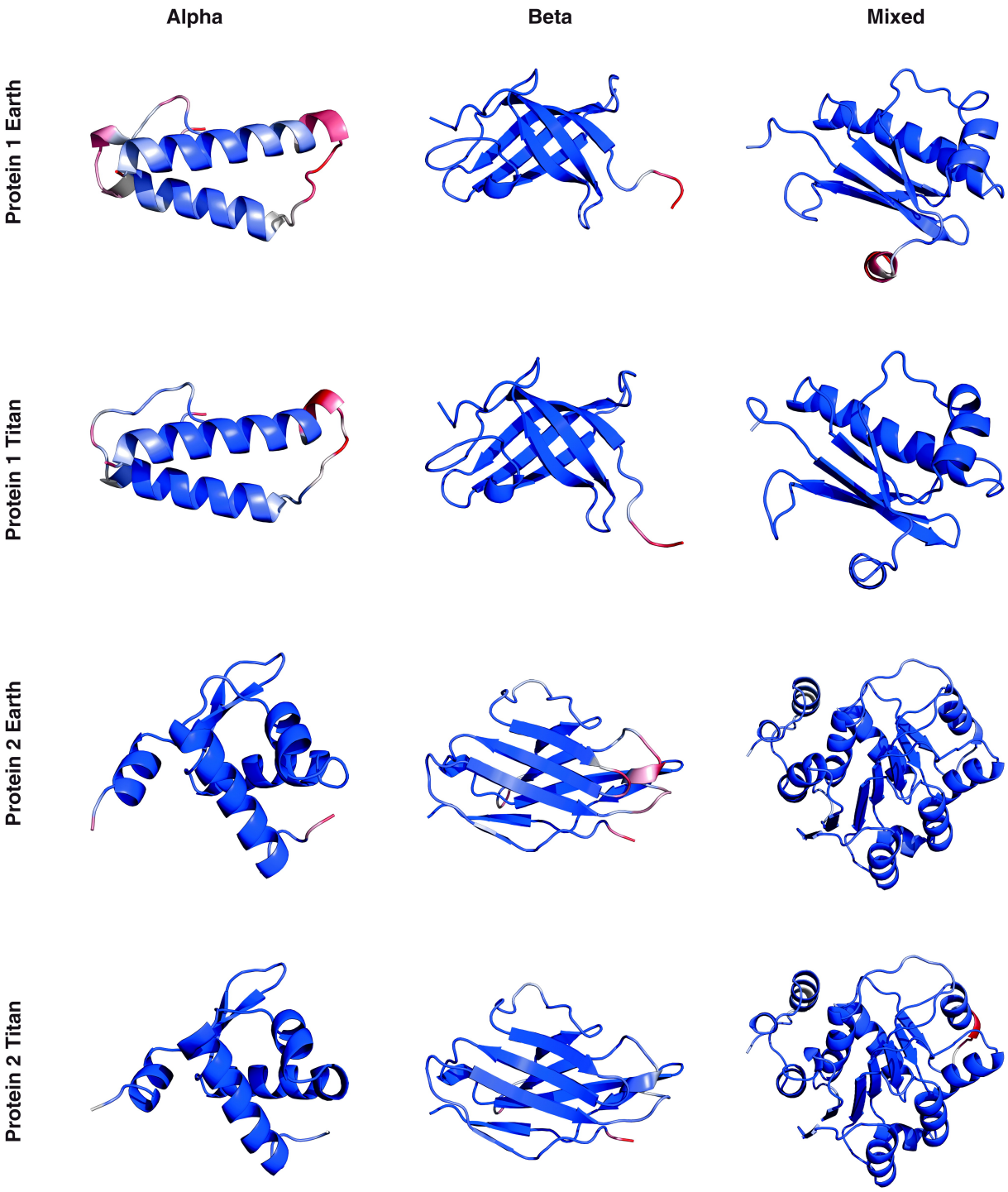
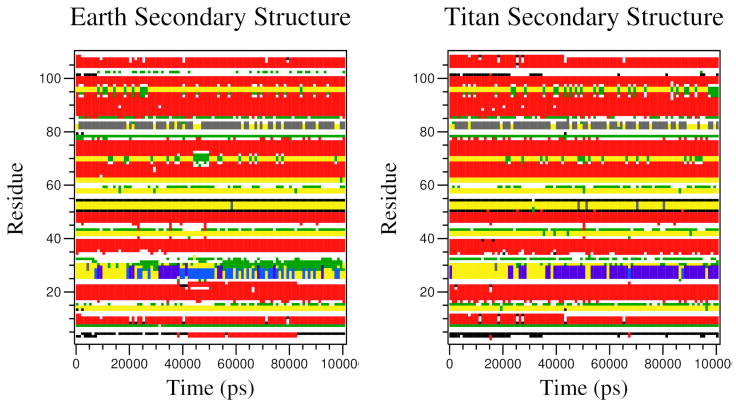


Figure 3.2: Root-mean-square fluctuations for each amino acid in the protein systems. Larger fluctuations are shown in red and smaller in blue. Proteins in the Titan environment experience larger maximum RMSF values as compared to the Earth environment but lower RMSF values on average.

Secondary structure results are shown in fig. 3.3 and demonstrate how the local structure of each

amino acid in the protein changes over the course of the simulation. The vertical axis is the amino acids in the protein, the horizontal axis is the simulation time, and Table 1 shows the color definitions for each secondary structure type. The emphasized region highlights that the protein does not become a specific secondary structure type in the Earth environment, in contrast to the Titan environment where the same region becomes a pi helix.



Color	Structure
White	Coil
Red	Beta sheet
Black	Beta bridge
Green	Bend
Yellow	Turn
Blue	Alpha helix
Purple	Pi helix
Gray	3-helix

Figure 3.3: Secondary structures for each amino acid of one of the protein systems over the simulation. Table 1 shows the color definitions used for each secondary structure type. The emphasized region highlights that the protein does not stabilize into a specific secondary structure type in the Earth environment in contrast to the Titan environment where the same region stabilizes into a pi helix.

Ramachandran plots are shown in fig. 3.4 and represent the distribution of backbone dihedral angles phi and psi that largely determine the protein's conformation for each protein as a two-dimensional histogram. These plots are made from subtracting the Earth dihedral distribution values from the Titan values. The gradient ranges from brown to blue-green; brown shows dihedrals that are favored in the Titan environment, and blue-green shows dihedrals that are favored in the Earth environment. In the Titan environment, the phi angle propensity is similar to the Earth environment, but the psi angle shifts from about -50 degrees to -25 degrees.

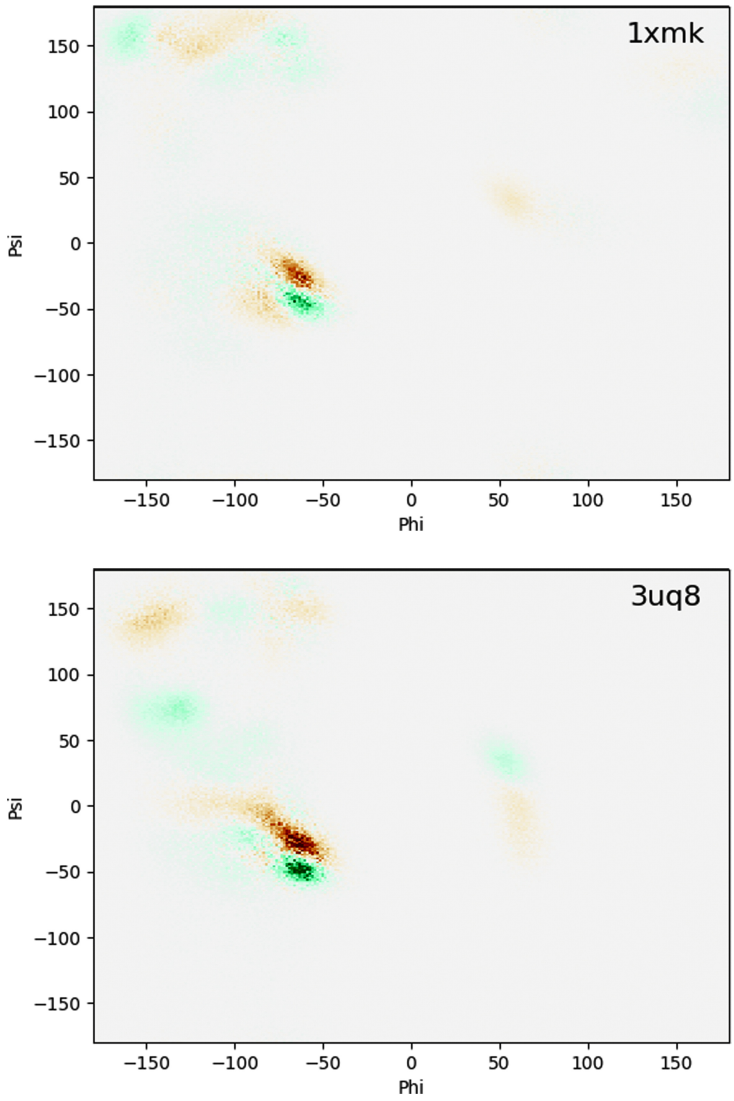


Figure 3.4: Ramachandran plots for all protein systems in the current study. Generally, the top left quadrant is beta sheets, and the bottom left quadrant is alpha helices. The gradient ranges from brown to blue-green; brown shows angles that are favored in the Titan environment, and blue-green shows angles that are favored in the Earth environment. In the Titan environment, the phi angle remains similar to the Earth environment, but the psi angle shifts from -50 degrees to -25 degrees.

3.4 Discussion

The radius of gyration results in fig. 3.1 show that the beta sheet proteins simulated in this study are unaffected by the difference in environment between Titan and Earth. By contrast, our results show the Titan environment leads to slightly more compact alpha helix proteins. We believe that the beta sheet proteins are less affected due to the higher number of stabilizing hydrogen bonds present in beta secondary structures. The beta sheets in these proteins have on average five hydrogen bonds compared

to alpha helices that have three hydrogen bonds.

The results in fig. 3.2 show that proteins in the Titan environment have lower average RMSF values than the Earth environment likely due to a combination of the high pressure and water-ammonia of the Titan environment suppressing atom fluctuations. Even though alpha helices are generally less long-lasting in the Titan environment (as shown in fig. 3.3), one protein had an alpha helix stabilize into a long-lasting pi helix (see fig. 3.2, beta, protein 2). Further studies will be needed to understand the mechanisms behind these differences.

The secondary structure results in fig. 3.3 show that the alpha helix region of one of the three proteins is more long-lasting in the Titan environment. In the Earth environment the helix region changes between different types of secondary structures: bends, turns, alpha helices, and pi helices. Turns and bends are similar in structure, but turns have hydrogen bonds whereas bends do not. In the Titan environment the helix region changes to a turn and then becomes a pi helix. The pi helix could be important because it shows Titan favoring a conformation that is not common for Earth conditions [36]. A pi helix is a helix with five hydrogen bonds compared to an alpha helix that has three. Due to the increase in hydrogen bonds, the pi helix is much more energetically favorable than an alpha helix. This suggests that proteins in the Titan environment may interact with other biomolecules with different biochemistry compared to Earth.

The Ramachandran plots in fig. 3.4 show that the Titan environment favors slightly different angles for the alpha helices. These different angles prefer a psi angle of about -25 degrees, whereas the phi angle remains the same. This psi angle difference is likely the result of differences in the preferred secondary structure types in the Titan environment, for example, a pi helix instead of an alpha helix. From these results, it appears that life on Titan would have similar beta sheets as Earth, allowing comparable proteins to form on Titan as on Earth.

3.5 Conclusion

In summary, protein secondary structure elements have different properties in a Titan environment compared to Earth. In the Titan environment alpha helices tended to be less flexible (fig. 3.2) and preferred slightly smaller psi dihedral backbone angles compared to an Earth environment (fig. 3.4). Protein structures were more compact in the Titan environment compared to an Earth environment (fig. 3.1). Most secondary structure elements were less long-lasting on Titan, but on rare occasions alpha helices were more long-lasting (see fig. 3.3) compared to the Earth environment. This study should be considered a starting point for a larger study to understand how proteins, protein-ligand complexes, and protein-protein complexes could fold, interact, and function in subsurface oceans such as those on Titan.

3.6 Acknowledgments

This research was supported by the Center for Modeling Complex Interactions sponsored by the National Institutes of Health (P20 GM104420), the National Science Foundation EPSCoR program (OIA-1736253), and the National Aeronautics and Space Administration Earth and Space Science Fellowship Program (NNX14AO30H). Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (P30 GM103324). The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4: Analysis of Software Methods for Estimation of Protein-Protein Relative Binding Affinity

Kyle P. Martin,[1,2] Tawny R. Gonzalez,[2] Jagdish S. Patel,[2] F. Marty Ytreberg,[1,2]

[1]Department of Physics, University of Idaho, [2]Institute for Modeling Complex Interactions, University of Idaho

Will be submitted to a TBD journal at a later date. As co-first author, I performed molecular dynamics simulations for 6 of the 16 complexes, ran all mutations for the iSEE method, wrote the in-house script to parse and analyze the data, and wrote the paper. This paper was previously published in Astrobiology and falls under the NIH public access policy (see https://publicaccess.nih.gov/) and can be used freely in this thesis. Final publication is available from Mary Ann Liebert, inc., publishers https://doi.org/10.1089/ast.2018.1972 .

4.1 Introduction

Biophysical modeling of protein-protein interactions provides insight into how and why proteins behave in specific ways. Mutations of the amino acids that make up these proteins play an essential role by introducing diversity into genomes. The more accurately mutation-induced changes in protein-protein binding can be determined, the more accurately we can predict loss-of-function phenotypes for previously uncharacterized point mutations. To understand living organisms, it is thus vital to have a comprehensive knowledge of how protein complexes interact under physiological conditions, that is, to determine their binding affinities and how these affinities can be modified[5]. Many researchers have developed and utilized computational methods to predict $\Delta\Delta G$ values, including values caused by single- or multiple-amino acid mutations. Experimental biophysical methods can quantitatively measure $\Delta\Delta G$ values for protein interactions, but these methods are typically costly, laborious, and time-consuming since all mutants must be expressed and purified[35, 71, 77].

Recently developed computational methods attempt to address the computational cost while accurately predicting $\Delta\Delta G$ values by including more precalculation informational variety and using prediction algorithms that fall into several categories. Some of these methods rely on known protein structures using functions that predict the energetic perturbation introduced by the mutation. Other methods train

machine learning methods on large data sets to combine selected physical, statistical, and empirical features for stability predictions. The most promising in terms of accuracy are rigorous methods based on statistical mechanics that use molecular dynamics (MD) simulations and are capable of addressing conformational flexibility and entropic effects; however these approaches are computationally highly expensive [11, 139, 47]. By contrast, non-rigorous, computationally less expensive, methods have been developed using the static all-atom protein complex structure.

In this work, we evaluate the ability of eight non-rigorous methods to predict amino acid mutation-induced free energy changes in protein binding in cases both for which an atomic-resolution structure is available and for which binding affinities of wild-type and mutant forms have been measured. To investigate whether any of these methods have a good trade-off between speed and accuracy, we chose 16 protein-protein test complexes with empirical $\Delta\Delta G$ values for observed mutations. Each test complex had at least 10 mutations occurring at different sites with varying empirical $\Delta\Delta G$ values. We calculated the $\Delta\Delta G$ values for each mutation with each method and compared the results with empirical $\Delta\Delta G$ values. Our hypothesis is that software methods using a wider variety of information will provide more accurate binding affinity and interface destabilization predictions than those relying on a single descriptive energy function and will do so with far less computational expense.

4.2 Methods

4.2.1 Compilation of Experimental $\Delta\Delta G$ Values

To assess the performance of a range of protein-protein binding affinity prediction methods, we first assembled a dataset containing mutations with known experimental relative binding affinity change ($\Delta\Delta G$) values. This list was assembled from Structural Kinetic and Energetic database of Mutant Protein Interaction (SKEMPI) version 2.0[65]. While generating this list, we considered four aspects: (i) type of the protein-protein complex; (ii) availability of quality 3-D structural information; (iii) range of experimental $\Delta\Delta G$ values; and (iv) the type of mutations at differing sites on the complex. Our final dataset contained 654 mutations from 16 protein-protein complexes and their respective experimental $\Delta\Delta G$ values. We further categorized these 16 complexes as either non-antibody-antigen (non-Ab) or antibody-antigen (Ab). Table 1 shows the complexes in our dataset with their respective non-Ab and Ab categories and the number of mutations associated with each complex. The dataset contains a total of 401 non-Ab mutations and 253 Ab mutations.

Non-Ab		Ab	
PDB	# Mutations	PDB	# Mutations
1a4y[114]	32	1bj1[103]	10
1brs[20]	30	1jrh[137]	42
1cbw[127]	31	1mlc[15]	11
1iar[48]	36	1vfb[12]	48
1jtg[87]	37	1yy9[86]	16
1lfd[55]	19	2jel[118]	43
1ppf[13]	190	3hfm[113]	71
2wpt[99]	26	4i77[149]	12

Table 4.1: Dataset containing all 16 protein complexes listed by PDB IDs and number of experimental mutants per complex for both Ab and Non-Ab categories.

4.2.2 Selection of Protein-Protein Binding Affinity Programs

Binding affinity prediction programs were chosen as those that had both a distinct approach to binding affinity calculation, and for which software was functional in October 2019 and available either online or upon request to the author. We also selected methods that utilize 3-D structural information of the protein complex. Table 2 summarises each method selected in this study, its approach, and its type of scoring function to calculate binding affinities. For simplicity, we categorized scoring functions, mathematical functions to calculate $\Delta\Delta G$ values of the selected programs as semi-empirical, statistical, or physics-based. Semi-empirical methods replace as many calculations as possible with pre-calculated data that are an integral part of the program. These semi-empirical methods were calibrated using existing crystal structures. Statistical methods use pre-calculated data and consider changes in coarse structural features such as the change in overall volume. Physics-based methods use energy functions to estimate enthalpic binding contributions. These programs were used to predict $\Delta\Delta G$ values for each mutation on our experimental list shown in Table 1. Detailed protocols for predicting $\Delta\Delta G$ values using each selected method is provided in the Supplemental Information.

4.2.3 Comparing Experimental and Predicted $\Delta\Delta G$ Values

To carry out statistical analysis of our results we built an in-house Python script that uses a combination of libraries including matplotlib, numpy, pandas, statistics, scipy, and sklearn. (Supplemental information file X) Using this script, we compared predicted $\Delta\Delta G$ values to experimental for each method.

To evaluate the predictive ability of each method tested, we compared concordance, Pearson, Kendall, and Spearman rank correlation coefficients using our script. Note that we distinguish methods that were calibrated to predict $\Delta\Delta G$ values from methods that compute metrics that are expected to linearly correlate with $\Delta\Delta G$ values. This distinction is important, as for optimal performance we expect a

Name	Brief Description	Scoring Function	Calibrated	Runtime (CPU hours)
BindProfX[163, 16]	Interface profile score based on conservation of homologous interfaces	Semi-Empirical	X	1ppf = 3.03 CPUh 1yy9 = 3.03 CPUh
BindProfX plus FoldX v3[163, 16]	Profile score weighted and combined with FoldX energy potential	Semi-Empirical	X	1ppf = 1.81 CPUh 1yy9 = 1.81 CPUh
iSEE[41]	Random forest model using structural, evolutionary, and energy-based features	Statistical		1ppf = 0 CPUh * 1yy9 = 0 CPUh *
DCOMPLEX v2[88]	Structural ideal-gas reference state potential	Physics-Based		1ppf = 0.013 CPUh 1yy9 = 0.001 CPUh
EasyE v1.0[152, 56]	GMEC-based method utilizing the Rosetta[2, 116] energy function	Statistical		1ppf = 0.48 CPUh 1yy9 = 0.09 CPUh
JayZ v1.0[152, 56]	Partition-function method utilizing Rosetta[2, 116] energy function	Statistical		1ppf = 0.14 CPUh 1yy9 = 0.21 CPUh
FoldX v4[46, 129]	Empirical energy score based on various energy parameters (e.g. van der Waals, solvation, electrostatics, hydrogen bonding)	Semi-Empirical	X	1ppf = 0.42 CPUh 1yy9 = 0.16 CPUh
MD[1] + FoldX v4[46, 129]	Molecular dynamics used to explore conformation space and generate snapshots; FoldX score calculated for each snapshot and averaged	Semi-Empirical	X	1ppf = 941 CPUh 1yy9 = 4093 CPUh

Table 4.2: Selected programs with a short summary of their approach and scoring function. Runtimes are listed for a representative protein complex for Ab (1yy9, 1058 AA) and Non-Ab (1ppf, 274) categories. 1yy9 is roughly four times bigger than 1ppf, which may or may not affect the total runtime.
* Runtime is significantly less than a second (note: preparation time is non-trivial and requires additional steps).

regression line that passes through the coordinate origin and has a slope of 1. In such a case all correlation coefficients would be equal to 1.

To compare the discriminating power of the methods, we obtained ROC curves. These curves quantify the ability of a method to correctly classify point mutations as destabilizing ($\Delta\Delta G$ ¡ -0.5 kcal/mol) or neutral/stabilizing ($\Delta\Delta G$ ¿ -0.5 kcal/mol). ROC curves that are skewed toward a higher true positive rate (sensitivity) classify mutations more accurately, as quantified by area under curve (ranging between 1.0 and 0.5 for perfect and chance classification, respectively).

We also used our script to parse the results on the basis of several physico-chemical and structural features to allow us to evaluate the methods based on these characteristics: wildtype amino acid type, mutant amino acid type, protein-protein interacting versus antibody-antigen, secondary structure classification of the mutation[142, 68], coordination number[148], Sneath index[135], mostly α-helical proteins versus mostly β-sheet proteins versus a mix of both α-helical and β-barrel proteins, percent exposure, location of the mutation, change in charge, change in polarity, change in volume, and whether or not the mutation location is predicted as an active or passive residue[150, 32, 156]. The script also outputs scatter plots, correlation plots, receiver operating characteristic (ROC) curves, and box plots to visualize the data, as well as correlations and standard deviations for each method. All plots in this manuscript were generated using this script.

4.3 Results and Discussion

The purpose of our study was to assess the ability of eight different relative binding affinity calculation methods (see table 4.2) to compare the estimated $\Delta\Delta G$ values with experiment (see fig. 4.1. We tested the performance of these methods using 16 different protein complexes (see table 4.2) with a total of 654 single amino acid mutations. We also looked at the computational speed of each method in the context of accuracy to determine its efficiency. Here, we analyze and discuss the results for Ab and non-Ab categories separately.

4.3.1 Non-Antibody-Antigen Results

The most correlating software is EasyE and the least correlating is iSEE (See figure 4.1 & 4.2). JayZ and EasyE are similar and consistently have the best correlation for non-Ab mutations. JayZ ran the 190 mutations of 1ppf in 507 seconds for an average of one mutation every 2.7 seconds. EasyE ran the 190 mutations of 1ppf in 1732 seconds for an average of one mutation every 9.1 seconds.The distribution of experimental $\Delta\Delta G$ values for all non-Ab complexes is as follows: 13% of point mutations resulted in $\Delta\Delta G$ values of less than -0.5 kcal/mol, considered destabilizing; 31% between -0.5 and 0.5 kcal/mol,

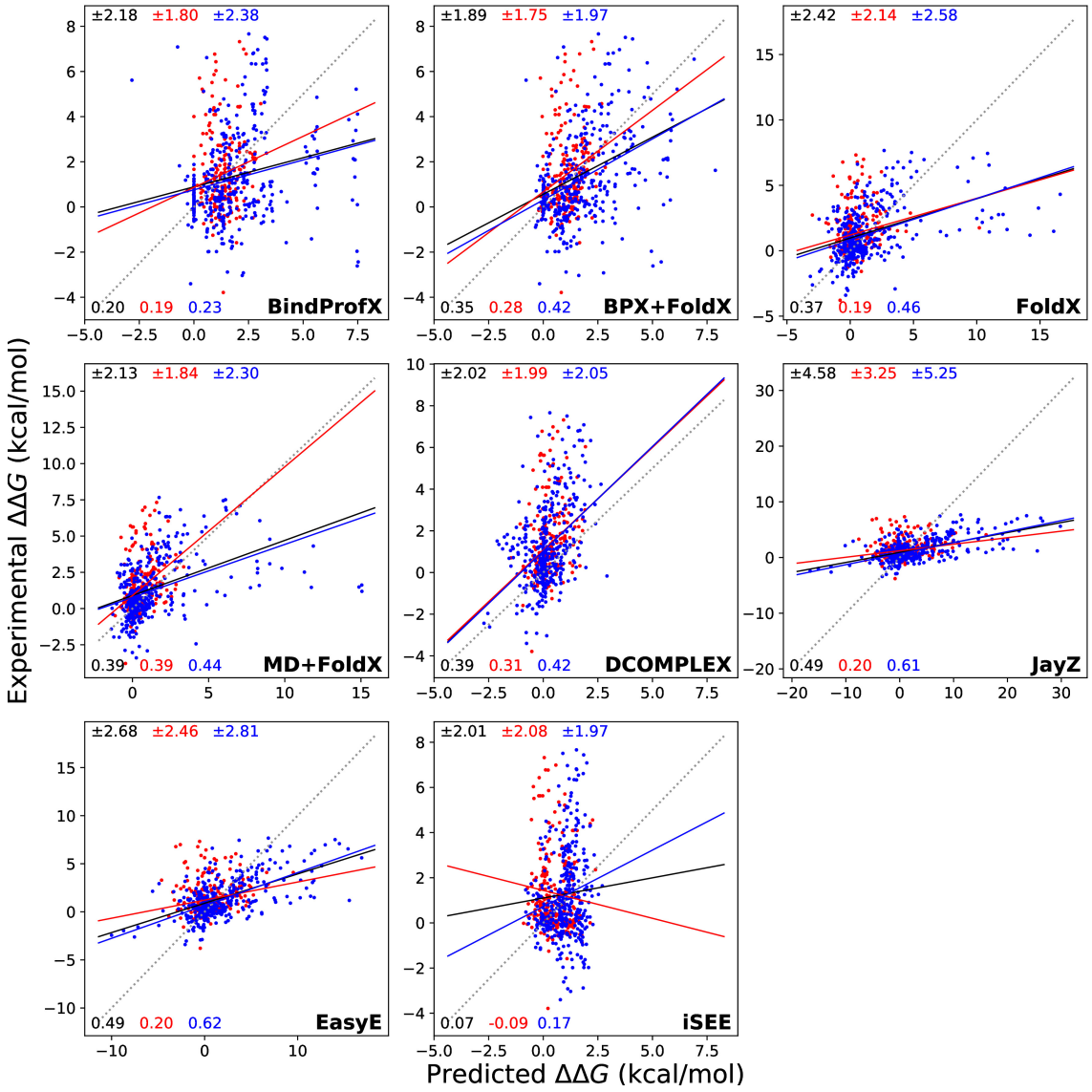


Figure 4.1: Calculated $\Delta\Delta G$ values (x-axis) compared to experimental $\Delta\Delta G$ values (y-axis) for each method tested in this study. Black, red, and blue lines are simple linear regressions from which Pearson correlations are derived. The red points are a scatter for Ab complexes and the blue points are for non-Ab complexes. The dashed line is the y = x line measuring perfect agreement between predicted $\Delta\Delta G$ and the experimental $\Delta\Delta G$ values. The solid black, red, and blue lines indicate a linear relationship between calculated and experimental observations for all data points, antibody-antigen complexes, and non-antibody-antigen complexes respectively. The top values in black, red, and blue match the root-mean-square error and the bottom values match that correlation coefficients for all values, Ab values, and non-Ab values respectively.

considered neutral; and 56% greater than 0.5 kcal/mol, considered stabilizing.

Figure 4.3 shows the ROC plot for all software and non-Ab complexes. JayZ [0.84], EasyE [0.83], DCOMPLEX [0.82], FoldX [0.79], and MD+FoldX [0.76] are the highest areas under the curve (AUC).

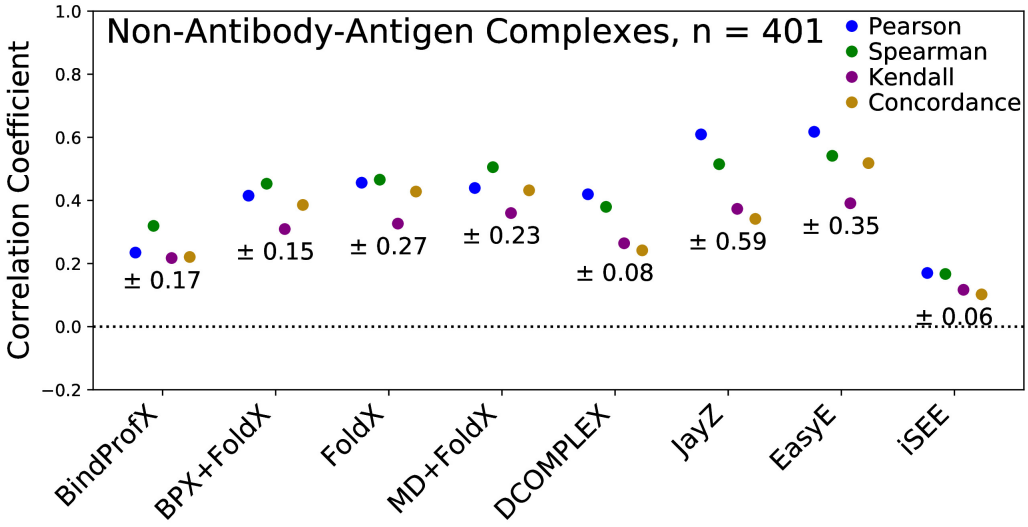


Figure 4.2: Performance of each method in predicting true $\Delta\Delta G$ values (concordance correlation coefficient), linearly correlated $\Delta\Delta G$ values (Pearson correlation coefficient), and rank order (Spearman and Kendall rank order correlation coefficient). The error for each method is reported under the correlation points. In total, there were 401 Non-Ab point mutations as designated by n=401.

Of all of these, DCOMPLEX has a much faster runtime. If the goal is to determine stabilizing and destabilizing non-Ab mutations, DCOMPLEX offers results similar to JayZ and EasyE, but at a fraction of the time. As above, JayZ runs one mutation every 2.7 seconds, EasyE every 9.1 seconds, but DCOMPLEX runs one mutation every 0.25 seconds.

All methods tested were trained on databases consisting mostly of non-Ab complexes, thus the correlation should be higher for these complexes as opposed to antibody-antigen complexes. It is possible that the true correlation is a combination of linear and rank correlation because Pearson and Spearman consistently correlated the best or second best. EasyE is the best option for balancing accuracy and speed. For destabilization/stabilization prediction, DCOMPLEX is recommended for its combination of speed and accuracy.

Table 4.3 shows eight different data subsets with two correlation coefficients per method. Can identify certain subsets that perform well in certain areas but not others. By removing worse performing subsets, a better idea of which method performs better for different subsets.

Table 4.3 also gives an idea of which methods perform better for non-Ab or Ab. EasyE has the highest correlation for non-Ab for five out of eight subsets. For the subsets it did not have the highest correlation, it had either the second or third highest correlation. This shows EasyE performs fairly well with non-Ab complexes compared to the other methods tested. Ab complexes are less straightforward. MD+FoldX

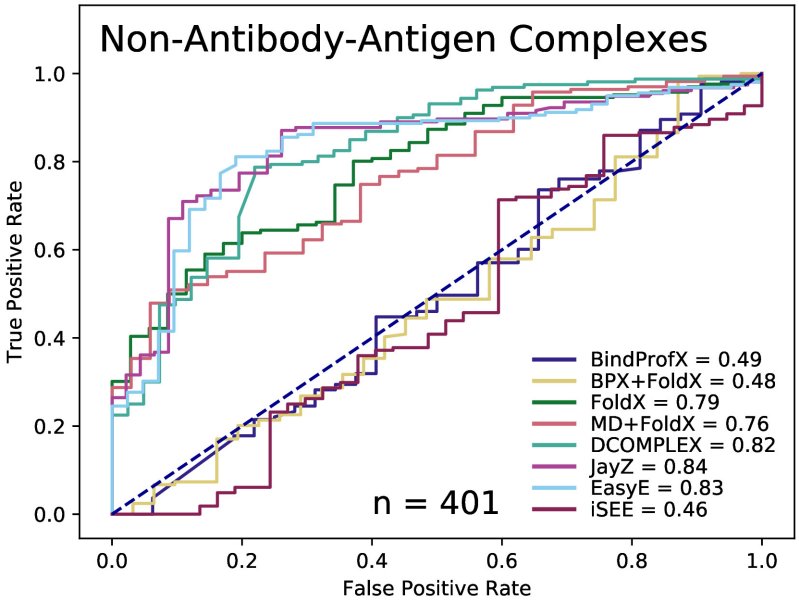


Figure 4.3: Receiver operating characteristic curves of the classification of variants that are more destabilized or less destabilized than 0.5 kcal/mol. The values in the legend represent the area-under-curve (AUC). The higher the value, the better the prediction capability of the method.

has the highest correlation for Ab complexes for three of the eight subsets. BindProfX+FoldX and DCOMPLEX both have the highest correlation for two of the eight subsets. An interesting trend is that the methods that have the highest correlation for non-Ab complexes do not have the highest correlation for Ab complexes.

4.3.2 Antibody-Antigen Results

The most correlating software for Ab complexes was MD+FoldX and the least correlating was iSEE (See figure 4.4). MD+FoldX is the most computationally expensive method, but can provide much better correlation than other software.

Figure 4.5 shows the ROC plot for all software and Ab complexes. The distribution of experimental $\Delta\Delta G$ values for only antibody-antigen complexes were as follows: 5% of point mutations resulted in $\Delta\Delta G$ values of less than -0.5 kcal/mol, considered destabilizing; 40% between -0.5 and 0.5 kcal/mol, considered neutral; and 55% greater than 0.5 kcal/mol, considered stabilizing. JayZ [0.97], EasyE [0.98], FoldX [0.85], and MD + FoldX [0.82] all had the highest AUC. JayZ and EasyE do remarkably well at determining if the mutation in an Ab complex is destabilizing or not. Compared to non-Ab complexes, all methods performed better for antibody-antigen complexes except for FoldX and DCOMPLEX which

Method	WT Gly or Pro	WT Non-Gly and Non-Pro	Alpha Helix	Beta Sheet	Surface Exposure	Neutral Charge	Hydrophobic to Polar	Large Vol Changes
BindProfX	Non-Ab: 0.11 Ab: -0.03	Non-Ab: 0.33 Ab: 0.23	Non-Ab: 0.29 Ab: 0.16	Non-Ab: 0.29 **Ab: 0.52**	Non-Ab: 0.22 Ab: 0.09	**Non-Ab: 0.37** Ab: 0.28	Non-Ab: 0.33 Ab: 0.17	Non-Ab: 0.13 Ab: 0.42
BindProfX + FoldX	**Non-Ab: 0.81** Ab: 0.09	Non-Ab: 0.45 Ab: 0.34	Non-Ab: 0.43 Ab: 0.39	Non-Ab: 0.43 **Ab: 0.54**	Non-Ab: 0.32 Ab: 0.21	Non-Ab: 0.52 Ab: 0.41	Non-Ab: 0.41 Ab: 0.26	Non-Ab: 0.71 **Ab: 0.50**
FoldX	**Non-Ab: 0.85** Ab: -0.11	Non-Ab: 0.45 Ab: 0.25	Non-Ab: 0.39 Ab: 0.25	Non-Ab: 0.39 Ab: 0.31	Non-Ab: 0.50 Ab: 0.26	Non-Ab: 0.42 **Ab: 0.41**	Non-Ab: 0.41 Ab: 0.11	Non-Ab: 0.63 Ab: -0.32
MD+FoldX	**Non-Ab: 0.83** **Ab: 0.71**	Non-Ab: 0.49 **Ab: 0.42**	Non-Ab: 0.44 **Ab: 0.54**	Non-Ab: 0.44 Ab: 0.49	Non-Ab: 0.47 Ab: 0.35	Non-Ab: 0.46 Ab: 0.46	**Non-Ab: 0.46** **Ab: 0.31**	Non-Ab: 0.71 Ab: 0.35
DCOMPLEX	**Non-Ab: 0.65** **Ab: 0.89**	Non-Ab: 0.34 Ab: 0.37	Non-Ab: 0.33 Ab: 0.31	Non-Ab: 0.33 Ab: 0.30	Non-Ab: 0.52 Ab: 0.27	Non-Ab: 40.36 **Ab: 0.56**	Non-Ab: 0.38 Ab: 0.16	Non-Ab: 0.62 Ab: 0.28
JayZ	Non-Ab: 0.80 **Ab: 0.54**	Non-Ab: 0.49 Ab: 0.24	Non-Ab: 0.44 Ab: -0.06	Non-Ab: 0.44 Ab: 0.16	Non-Ab: 0.59 **Ab: 0.36**	**Non-Ab: 0.62** Ab: 0.26	Non-Ab: 0.41 Ab: 0.01	**Non-Ab: 0.83** Ab: 0.19
EasyE	Non-Ab: 0.80 Ab: 0.29	**Non-Ab: 0.51** Ab: 0.22	**Non-Ab: 0.51** Ab: 0.06	**Non-Ab: 0.51** Ab: 0.03	**Non-Ab: 0.60** **Ab: 0.35**	Non-Ab: 0.61 Ab: 0.23	Non-Ab: 0.45 Ab: 0.02	**Non-Ab: 0.84** Ab: 0.18
iSEE	**Non-Ab: 0.43** Ab: -0.43	Non-Ab: 0.28 Ab: -0.16	Non-Ab: 0.05 Ab: -0.04	Non-Ab: 0.05 Ab: -0.24	Non-Ab: 0.15 **Ab: 0.11**	Non-Ab: 0.15 Ab: -0.11	Non-Ab: 0.14 Ab: -0.02	Non-Ab: 0.24 Ab: -0.44

Table 4.3: All methods correlation coefficients with respect to certain subsets. “”WT Gly or Pro” are wildtype amino acids that are either glycine or proline. “WT Non-Gly and Non-Pro” are wildtype amino acids that are neither glycine nor proline. “Alpha Helix” are mutations that occur in a helix structure. “Beta Sheet” are mutations that occur in a beta structure. “Surface Exposure” are mutations that occur in an amino acid that is up to 10% solvent accessible. “Neutral Charge” is a neutrally charged wildtype amino acid mutating to a neutrally charged mutant amino acid. “Hydrophobic to Polar” is a hydrophobic or polar wildtype amino acid mutating to a polar or hydrophobic mutant amino acid, respectively. “Larger Vol Changes” is a mutant amino acid that is greater than 40% larger than the wildtype amino acid. Values that are bolded are the highest correlation coefficients for each method and protein type. Values that are red or blue are the highest correlation coefficients for each subset, red for non-Ab and blue for Ab. The red and blue are the dominant representations.

were marginally worse. EasyE is recommended for a better stabilization prediction and offers a fairly quick solution. EasyE ran 16 mutations of 1yy9 in 336 seconds for an average of one mutation every 21 seconds.

MD+FoldX ran 16 mutations of 1yy9 in 941 CPUh for an average of one mutation every 58.8 CPUhs. DCOMPLEX offers a slightly lower correlation but is much less computationally expensive. It ran 16 mutations of 1yy9 in 5.0 seconds for an average of one mutation every 0.35 seconds.

MD+FoldX does much better with accuracy, at a much larger cost of runtime. For destabilization/stabilization prediction, EasyE or JayZ are the best options for balancing accuracy and speed.

4.3.3 Discussion

We hypothesized that methods utilizing a wide variety of information to predict relative binding affinity and interface destabilization would be more accurate than methods based on a single descriptive

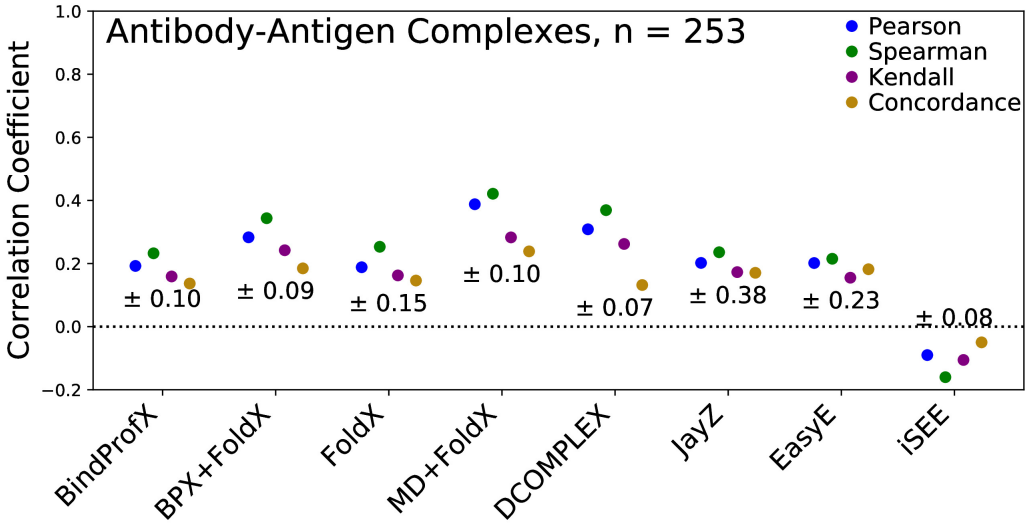


Figure 4.4: Performance of each evaluated method in predicting true $\Delta\Delta G$ values (concordance correlation coefficient), linearly correlated $\Delta\Delta G$ values (Pearson correlation coefficient), and rank order (Spearman and Kendall rank order correlation coefficient). The error for each method is reported under the correlation points. In total, there were 253 Ab point mutations as designated by n=253.

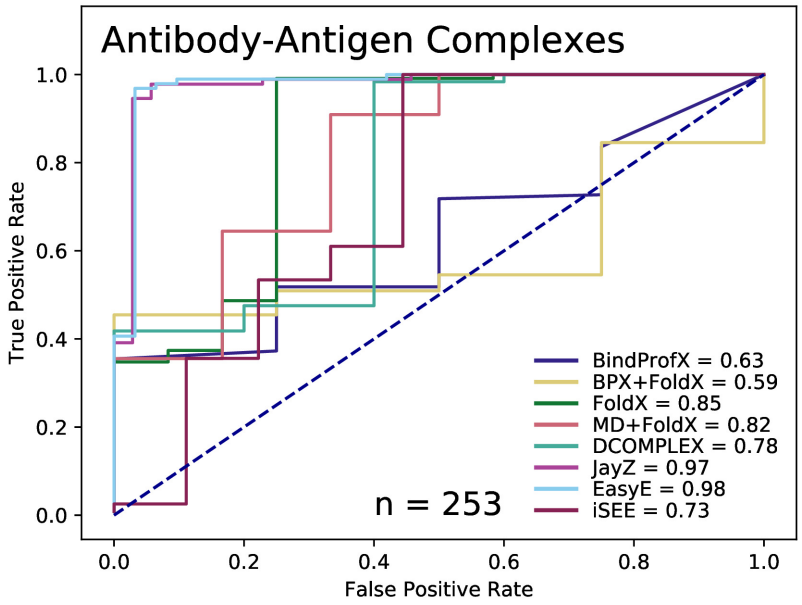


Figure 4.5: Figure 4.5: Receiver operating characteristic curves of the classification of variants that are more destabilized or less destabilized than 0.5 kcal/mol. The values in the legend represent the area-under-curve (AUC). The higher the value, the better the prediction capability of the method.

energy function and would do so in a computationally efficient manner. Based on the subsets we were able to analyze, our results suggest a more complex picture as each method has both strengths and limitations in predicting binding behavior. There is likely an intricate interplay of protein characteristics, as shown in Table 3, that limits the ability of current tools to predict binding affinity and interface destabilization with both speed and accuracy in multiple protein subsets.

Figure 4.1 summarizes our ability to estimate $\Delta\Delta G$ values for single mutations. None of the methods show a high correlation, highlighting the continued need for better implementation of our growing knowledge of protein characteristics into reliable and efficient tools for predicting protein behavior. JayZ and EasyE both have correlations of 0.49 for all single mutations, and have higher correlations of 0.61 and 0.62 (respectively) for non-Ab complexes. MD+FoldX had a correlation of 0.39 for Ab complexes. MD+FoldX appears to do better at non-trivial complexes. Software is typically trained on non-Ab complexes which is why MD+FoldX performed somewhat lower than others in that category. But for the Ab complexes, it had the best correlation. All methods have higher correlations for non-Ab complexes than Ab complexes. In general, the Pearson correlation coefficient for the various methods are all relatively average. However, JayZ and EasyE have high Pearson correlation coefficients, meaning the data is most likely linear in nature. The concordance correlation coefficient performed the worst or second worst of all correlation coefficients. Spearman rank correlation had better correlation than Kendall rank correlation in every method. It is possible that the true correlation is a combination of linear and rank correlation because Pearson and Spearman consistently correlated the best or second best. Based on predictive accuracy alone, JayZ and EasyE appear to be the best overall predictor of relative binding affinity and perform especially well with non-Ab complexes. MD+FoldX outperforms all other methods when analyzing Ab complexes.

While accuracy is generally the main driver for choosing a computational method, computational efficiency is also important. Here, we discuss each method as it performs in both speed and accuracy for predicting $\Delta\Delta G$ values. For all single mutations and our non-Ab subset, EasyE and JayZ both performed better than FoldX, a finding that appears to fit our hypothesis given both utilize a layered approach to energetic computations. However, while JayZ appears to be a faster method, EasyE is computationally equivalent to FoldX. DCOMPLEX, a physics-based method arguably on par with FoldX in terms of informational variety, performs better than FoldX for all single mutations and almost as well as FoldX for non-Ab mutations in a fraction of the time. MD+FoldX is on par with DCOMPLEX for all single mutations and similarly to FoldX in non-Ab mutations, but is by far the most computationally expensive method analyzed. Although BPX+FoldX implements several factors in its algorithm, computation time was longer than all but MD+FoldX without a concomitant improvement in correlation. It must be

noted, however, this method is perhaps the most accessible given the easy-to-use online server interface and performs similarly to FoldX for all single mutations and non-Ab mutations. BindProfX utilizes the same scoring profile as BPX+FoldX without the FoldX calculations. In this case, accuracy decreased while calculation speed remained similar to BPX+FoldX. iSEE, the least correlating method, employs the widest variety of information to obtain relative binding affinity predictions and is the fastest of all methods, if the non-trivial preparation time is ignored.

For Ab complexes, MD+FoldX completely negates our hypothesis in that it was the most correlating method, followed by DCOMPLEX. iSEE is again the fastest of all methods but also the least correlating. BindProfX, JayZ, and EasyE utilize a wide variety of information and have similar correlations but do not approach the accuracy of MD+FoldX. Here, BPX+FoldX and DCOMPLEX performed similarly, but DCOMPLEX has the advantage of speed. For Ab complexes, no method can really be recommended as both accurate and efficient, although DCOMPLEX appears to do surprisingly well for not being calibrated given its fast runtime.

Figures 4.3 and 4.5 show the ROC curves for each method's ability to classify mutations as either destabilizing or neutral/stabilizing. For non-Ab complexes, JayZ and EasyE have the best predictive ability followed by DCOMPLEX. In terms of speed, DCOMPLEX is faster than either JayZ or EasyE. Here, given the similarities in predictive ability, the tradeoff between speed and accuracy is best determined by the user. For Ab complexes, JayZ and EasyE performed very well, followed by MD+FoldX. If mutation classification is the primary need, JayZ or EasyE are both recommended over MD+FoldX due to their much faster runtime.

Some of these software have much longer runtimes but similar correlations. Thus, if the outcome only needs to be so good, the faster software would be ideal. In the case of the fastest runtime, iSEE computes all mutations instantly, but performs poorly for both non-Ab and Ab. DCOMPLEX does much better at stabilization prediction and correlation. If speed is the most important parameter, DCOMPLEX would be a good choice for general binding affinity calculations. As can be seen in Table 3, some of the aforementioned methods performed extremely well, and even in some cases efficiently, in certain subcategories but were less effective in others, highlighting the need for understanding how the interplay of these protein characteristics actually affect computational accuracy and resource usage.

Our main result is that we have identified several computational methods that predict relative binding affinity. While some complexes can be easily approximated using quick and less rigorous, more novel and unique complexes will most likely have improved accuracy utilizing molecular dynamics. The in-house script can parse any aforementioned parameters, and combined with the known runtimes above, one of the identified softwares should be used to get the fastest and most accurate relative binding affinity. The

dataset and accompanying python script should be enough for a user to determine correlations and ROC curves for all factors shown in Table 3. The script could also be used to elucidate strengths and potential problem areas for any given method and protein subset, allowing users to determine the best predictive method for the types of mutations in which they are interested beyond what is discussed in this section.

4.4 Conclusion

In this article, we have described several computational methods and tested them to predict the effects of single mutations on protein-protein binding affinity. We have shown that some methods perform better than others depending on the complex such as EasyE for non-Ab complexes and MD+FoldX for Ab complexes. Moreover, JayZ and EasyE predicted the signs of relative binding free energy ($\Delta\Delta G$) values of studied mutations with high accuracy compared to any other method. In future work, we could look at more complexes or different methods to expand and better refine our conclusion on the predictive capability of each method.

Chapter 5: DmORC

5.1 Introduction

Drosophila melanogaster (Dm) is a species of fly in the family *Drosophilidae*. The origin recognition complex (ORC) directs DNA replication throughout the genome and is required for its initiation. ORC is central for the duplication of a cell and is necessary for the maintenance of the eukaryotic genome [83]. DmORC is a protein complex with six chains.

In this study we were looking at mutations in ORC2 (Chain B, PDB ID 4xgc) and ORC3 (Chain C, PDB ID 4xgc) and how they affect the relative binding affinity with a DNA complex. Dm samples were taken along the east coast of the United States and two mutations were found in some populations [9]. The populations that contained the mutations depended on the physical location where the sample was collected. ORC2 contained the mutation T321P and ORC3 contained the mutation S339N. Both T321P and S339N appeared more often in species that existed at a colder and lower (towards the North Pole) latitude.

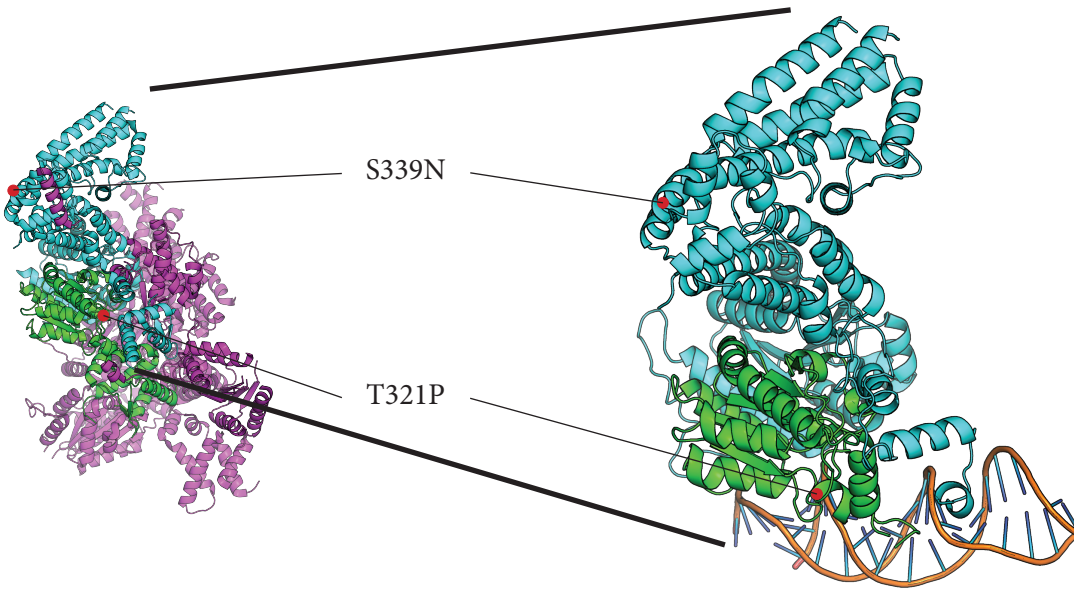


Figure 5.1: The *Drosophila melanogaster* origin recognition complex (DmORC) is of immense interest in the scientific community due to its similarity to the human ORC. (PDB ID 4xgc[31]). The 3D image on the left is the complete DmORC structure. The 3D image on the right is ORC2 and ORC3 bound to DNA. S339N and T321P are non-synonymous amino acid variants. Residue 339 in chain C has been observed to mutate from a Serine to an Asparagine and Residue 321 in chain B has been observed to mutate from Threonine to a Proline.

5.2 Methods

Using Schrodinger, DmORC was trimmed to only contain ORC2 and ORC3, since these are the only chains with observed mutations. Each ORC structure was divided into a main section and a winged hinge (WH). The WH was assumed not to contribute to the complex since the mutations were located in the main sections. Protein Preparation Wizard was used to preprocess, optimize, and minimize the structure [92]. The Structure Prediction tool was used to find and compare homologs. It then applied homolog characteristics and rebuilt portions of the intrinsically disorder portions of the structure using an energy-based method.

After the primary structure was built, we focused on secondary structures. Prime was used to run Monte Carlo simulations that forced secondary structures into disordered regions of the protein chains [62, 63]. Secondary structure was predicted for each chain using several intrinsically disordered structure predictors (I-TASSER[164], SpotFold[22], NetSurfP[75]). The structure was then processed through a 10 ns molecular dynamics simulation to ensure there were no steric clashes and confirm stability.

The disordered regions of DmORC were also rebuilt using Modeller[158]. This method was a quicker compared to Schrodinger, but turned out to be less effective. After analyzing the structure, it was decided that the rebuilt disordered regions were less reliable. Modeller has ways of refining the loops it has rebuilt, but it does not compare the structure to homologs. This means it only using statistical mechanics to rebuild the disordered regions, and not similar known protein sequence structures.

5.3 Results and Discussion

The results from the MD simulation showed some promise for understanding the effects of the two temperature sensitive mutations on DmORC. However, after discussion with our experimental collaborator, it was determined that the structure needs to be built again to include the WH portions. This will be left for future work.

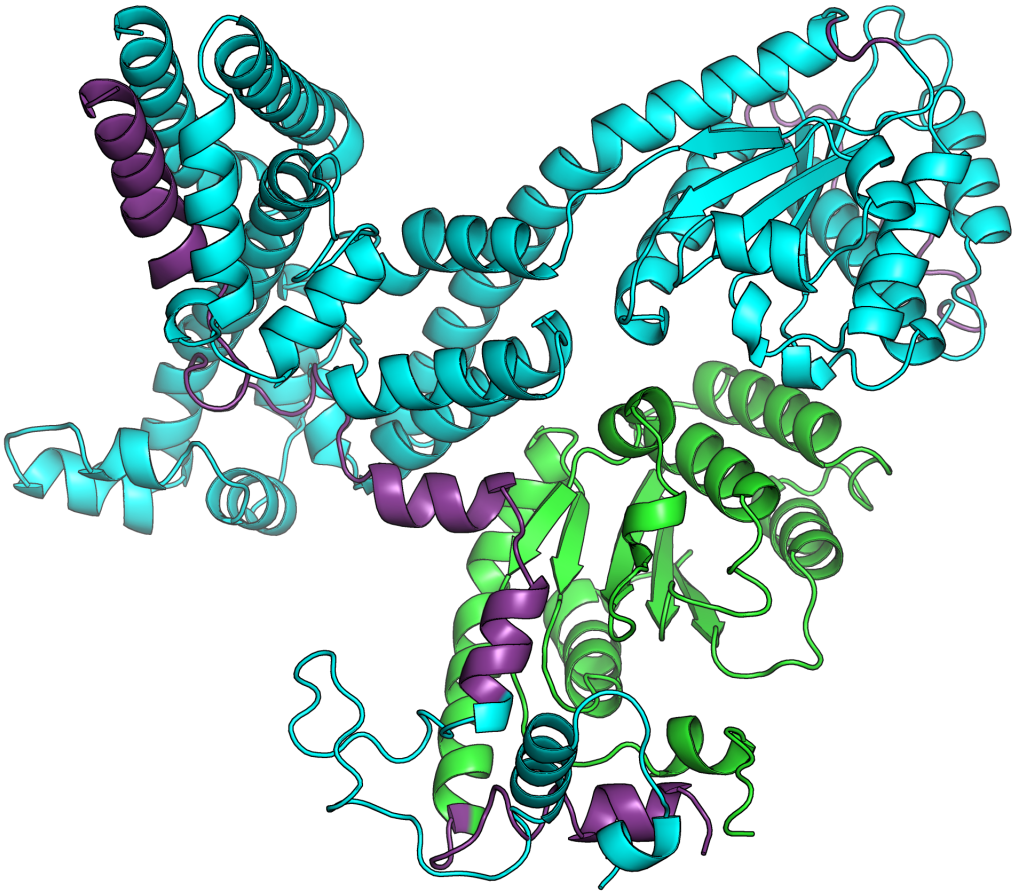


Figure 5.2: The green region represents ORC2 and the cyan region represents ORC3. The purple region is what was rebuilt using Schrodinger.

Chapter 6: Conclusion

6.1 Summary

In our study of Ebola (see chapter 2), we showed that molecular modeling can be used to simulate the interactions between antibodies and antigens. In-silico experiments, while less accurate than in-vitro or in-vivo, are much faster. We calculated the effects of thousands of mutations to identify those that disrupt antibody binding. This type of modeling has applications in drug design and development of vaccines.

In our study of Titan's subsurface ocean (see chapter 3), we showed that molecular modeling can be applied to the extra-terrestrial search for life. An ammonia-water solvent was made to simulate an alien environment, and Earth-based protein structures were simulated in this environment. Although we have no concrete clues as to what exists in the subsurface ocean of Titan, we've shown that Earth-based proteins containing beta sheets are stable and could exist in a Titan environment. This type of research has applications in the universal search for life.

In our study of binding affinity software (see chapter 4), we've shown the accuracy of binding free energy calculations for eight different methods. By comparing the raw data, the ideal software can be chosen for a specific system or mutation. This research is leading to a better understanding of both computational models and the importance of the physical attributes of the mutations and how it impacts appropriate software choice.

6.2 Future Research

Our study of Ebola set a precedence for understanding how evolution can impact treatment of a viral disease. This same method could be used to generate watch lists for any virus with known structures. The method could also be refined by adding more simulation time to allow for more accuracy, or adding other methods depending on the mutation type or location (see chapter 4).

Future research for the our Titan project could include calculating binding and folding energies, protein functionality changes, or intrinsically disordered proteins. The approach is very general; different solvent or proteins could be used. Binding energies could shed some light on whether or not these new environments help or hinder protein functionality, or whether novel binding mechanisms could exist in other environments.

The research we did looking at mutation binding energy accuracy with different software is general and thus can be applied to many different binding systems. Two ideas, in particular, can be readily

expanded. More methods could be used to offer even more options. Or, more systems and experimental data can be included to better refine and optimize the results. The script we used to parse data is general and can be easily applied to other data sets. Additionally, some future research could involve improving the script by adding plots, features, or a better interface.

DmORC is important to our understanding of DNA replication and cell division. The next step for this project is to build the DmORC structure with the winged hinges of ORC2 and ORC3. Ideally the entire DmORC structure (six chains) should be rebuilt, but it's unclear if the other chains have any effect on the mutations. Once the structure is built, it can be simulated and analyzed. Ideally, DmORC should be simulated at different temperatures to compare to the experimental data.

References

[1] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, September 2015.

[2] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Jr, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031, June 2017.

[3] C. B. Anfinsen. The formation and stabilization of protein structure. *Biochem. J.*, 128(4):737–749, Jul 1972.

[4] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 11 2005.

[5] Marc Baaden and Siewert J Marrink. Coarse-grain modelling of protein–protein interactions. *Current Opinion in Structural Biology*, 23(6):878–886, December 2013.

[6] Wayne M. Barnes. The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene*, 112(1):29–35, mar 1992.

[7] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, J. L. Milne, and S. Subramaniam. 2.2 Ã... resolution cryo-EM structure of Î²-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, Jun 2015.

[8] Pierre Becquart, Tanel Mahlakõiv, Dieudonné Nkoghe, and Eric M. Leroy. Identification of continuous human b-cell epitopes in the VP35, VP40, nucleoprotein and glycoprotein of ebola virus. *PLoS ONE*, 9(6):e96360, June 2014.

[9] Alan O. Bergland, Emily L. Behrman, Katherine R. O'Brien, Paul S. Schmidt, and Dmitri A. Petrov. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in drosophila. *PLOS Genetics*, 10(11):1–19, 11 2014.

[10] Helen M Berman. The Protein Data Bank / Biopython. *Presentation*, 28(1):235–242, 2000.

[11] R. C. Bernardi, M. C. R. Melo, and K. Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta*, 1850(5):872–877, May 2015.

[12] T. N. Bhat, G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dall'Acqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences*, 91(3):1089–1093, February 1994. Publisher: National Academy of Sciences Section: Research Article.

[13] W. Bode, A. Z. Wei, R. Huber, E. Meyer, J. Travis, and S. Neumann. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *The EMBO journal*, 5(10):2453–2458, October 1986.

[14] Luciana Borio, Edward Cox, and Nicole Lurie. Combating emerging threats—accelerating the availability of medical therapies. *New England Journal of Medicine*, 2015.

[15] Bradford C. Braden, Hélène Souchon, Jean-Luc Eiselé, Graham A. Bentley, T. Narayana Bhat, Jorge Navaza, and Roberto J. Poljak. Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. *Journal of Molecular Biology*, 243(4):767–781, November 1994.

[16] J. R. Brender and Y. Zhang. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput. Biol.*, 11(10):e1004494, Oct 2015.

[17] Melinda A Brindley, Laura Hughes, Autumn Ruiz, Paul B McCray, Anthony Sanchez, David A Sanders, and Wendy Maury. Ebola virus glycoprotein 1: identification of residues important for binding and postbinding events. *Journal of virology*, 81(14):7702–7709, 2007.

[18] J Rodney Brister, Yiming Bao, Sergey A Zhdanov, Yuri Ostapchuck, Vyacheslav Chetvernin, Boris Kiryutin, Leonid Zaslavsky, Michael Kimelman, and Tatiana A Tatusova. Virus variation resource—recent updates and future directions. *Nucleic acids research*, page gkt1268, 2013.

[19] A. M. Buckle, G. Schreiber, and A. R. Fersht. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry*, 33(30):8878–8889, Aug 1994.

[20] Ashley M. Buckle, Gideon Schreiber, and Alan R. Fersht. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.ANG. resolution. *Biochemistry*, 33(30):8878–8889, August 1994. Publisher: American Chemical Society.

[21] Ondřej Čadek, Gabriel Tobie, Tim Van Hoolst, Marion Massé, Gaël Choblet, Axel Lefèvre, Giuseppe Mitri, Rose Marie Baland, Marie Běhounková, Olivier Bourgeois, and Anthony Trinh. Enceladus's internal ocean and ice shell constrained from Cassini gravity, shape, and libration data, jun 2016.

[22] Y. Cai, X. Li, Z. Sun, Y. Lu, H. Zhao, J. Hanson, K. Paliwal, T. Litfin, Y. Zhou, and Y. Yang. SPOT-Fold: Fragment-Free Protein Structure Prediction Guided by Predicted Backbone Structure and Contact Map. *J Comput Chem*, 41(8):745–750, Mar 2020.

[23] Stefano Campanaro, Laura Treu, and Giorgio Valle. Protein evolution in deep sea bacteria: An analysis of amino acids substitution rates. *BMC Evolutionary Biology*, 8(1):1–15, 2008.

[24] Michael H. Carr, Michael J.S. Belton, Clark R. Chapman, Merton E. Davies, Paul Geissler, Richard Greenberg, Alfred S. McEwen, Bruce R. Tufts, Ronald Greeley, Robert Sullivan, James W. Head, Robert T. Pappalardo, Kenneth P. Klaasen, Torrence V. Johnson, James Kaufman, David Senske, Jeffrey Moore, Gerhard Neukum, Gerald Schubert, Joseph A. Burns, Peter Thomas, and Joseph Veverka. Evidence for a subsurface ocean on Europa. *Nature*, 391(6665):363–365, jan 1998.

[25] Julie C. Castillo-Rogez and Jonathan I. Lunine. Evolution of Titan's rocky core constrained by Cassini observations. *Geophysical Research Letters*, 37(20), oct 2010.

[26] Centers for Disease Control. `http://www.cdc.gov/vhf/ebola/`.

[27] Mathieu Choukroun, Olivier Grasset, Gabriel Tobie, and Christophe Sotin. Stability of methane clathrate hydrates under pressure: Influence on outgassing processes of methane on Titan. *Icarus*, 205(2):581–593, 2010.

[28] Athena Coustenis. Formation and Evolution of Titan's Atmosphere. In *The Outer Planets and their Moons*, pages 171–184. Springer-Verlag, Berlin/Heidelberg, 2005.

[29] F J Crary, B A Magee, K Mandt, J H Waite, J Westlake, and D T Young. Heavy ions, temperatures and winds in Titan's ionosphere: Combined Cassini CAPS and INMS observations. *Planetary and Space Science*, 57(14-15):1847–1856, 2009.

[30] Edgar Davidson, Christopher Bryan, Rachel H. Fong, Trevor Barnes, Jennifer M. Pfaff, Manu Mabila, Joseph B. Rucker, and Benjamin J. Doranz. Mechanism of binding to ebola virus glycoprotein by the zmapp, zmab, and mb-003 cocktail antibodies. *Journal of Virology*, 89(21):10982–10992, 2015.

[31] DmORC. `https://www.rcsb.org/structure/4XGC`.

[32] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK:- A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7):1731–1737, February 2003.

[33] William Dowling, Elizabeth Thompson, Catherine Badger, Jenny L Mellquist, Aura R Garrison, Jeffery M Smith, Jason Paragas, Robert J Hogan, and Connie Schmaljohn. Influences of glycosylation on antigenicity, immunogenicity, and protective efficacy of ebola virus gp dna vaccines. *Journal of virology*, 81(4):1821–1837, 2007.

[34] Derek Dube, Matthew B Brecher, Sue E Delos, Sean C Rose, Edward W Park, Kathryn L Schornberg, Jens H Kuhn, and Judith M White. The primed ebolavirus glycoprotein (19-kilodalton gp1, 2): sequence and residues critical for host cell binding. *Journal of virology*, 83(7):2883–2891, 2009.

[35] Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia, and Michael L. Tress. Progress and challenges in predicting protein–protein interaction sites. *Briefings in Bioinformatics*, 10(3):233–246, 04 2009.

[36] M N Fodje and S Al-Karadaghi. Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering*, 15(5):353–358, 2002.

[37] A D Fortes. Exobiological Implications of a Possible Ammonia–Water Ocean inside Titan. *Icarus*, 146:444–452, 2000.

[38] A. D. Fortes. Titan's internal structure and the evolutionary consequences. *Planetary and Space Science*, 60(1):10–17, jan 2012.

[39] Michael Garbutt, Ryan Liebscher, Victoria Wahl-Jensen, Steven Jones, Peggy Möller, Ralf Wagner, Viktor Volchkov, Hans-Dieter Klenk, Heinz Feldmann, and Ute Ströher. Properties of replication-competent vesicular stomatitis virus vectors expressing glycoproteins of filoviruses and arenaviruses. *Journal of virology*, 78(10):5458–5465, 2004.

[40] Y Geiß and U Dietrich. Catch me if you can-the race between hiv and neutralizing antibodies. *AIDS reviews*, 17(2):107–113, 2014.

[41] Cunliang Geng, Anna Vangone, Gert E. Folkers, Li C. Xue, and Alexandre M. J. J. Bonvin. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics*, 87(2):110–119, February 2019.

[42] S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P.-P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, September 2014.

[43] O. Grasset, A. Coustenis, W. B. Durham, H. Hussmann, R. T. Pappalardo, S. Sasaki, and D. Turrini. Satellites of the outer solar system: Exchange processes involving the interiors, jun 2010.

[44] O Grasset and J Pargamin. The ammonia-water system at high pressures: Implications for the methane of Titan. *Planetary and Space Science*, 53(4):371–384, 2005.

[45] Yi Gu and Ming Li. *Handbook of Benzoxazine Resins*, chapter Chapter 3 - Molecular Modeling, pages 103–110. Elsevier, Amsterdam, 2011.

[46] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, 320(2):369–387, 2002.

[47] J. C. Gumbart, B. Roux, and C. Chipot. Efficient determination of protein-protein standard binding free energies from first principles. *J Chem Theory Comput*, 9(8), Aug 2013.

[48] Thorsten Hage, Walter Sebald, and Peter Reinemer. Crystal Structure of the Interleukin-4/Receptor α Chain Complex Reveals a Mosaic Binding Interface. *Cell*, 97(2):271–281, April 1999. Publisher: Elsevier.

[49] Ana Maria Henao-Restrepo, Ira M Longini, Matthias Egger, Natalie E Dean, W John Edmunds, Anton Camacho, Miles W Carroll, Moussa Doumbia, Bertrand Draguez, Sophie Duraffour, et al. Efficacy and effectiveness of an rvsv-vectored vaccine expressing ebola surface glycoprotein: interim results from the guinea ring vaccination cluster-randomised trial. *The Lancet*, 2015.

[50] Amanda R Hendrix, Terry A Hurford, Laura M Barge, Michael T Bland, Jeff S Bowman, William Brinckerhoff, Bonnie J Buratti, Morgan L Cable, Julie Castillo-Rogez, Geoffrey C Collins, Serina Diniega, Christopher R German, Alexander G Hayes, Tori Hoehler, Sona Hosseini, Carly J.A. Howett, Alfred S. McEwen, Catherine D Neish, Marc Neveu, Tom A Nordheim, G Wesley Patterson, D Alex Patthoff, Cynthia Phillips, Alyssa Rhoden, Britney E Schmidt, Kelsi N Singer, Jason M Soderblom, and Steven D Vance. The NASA Roadmap to Ocean Worlds. *Astrobiology*, 19(1):1–27, 2018.

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

[52] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, Mar 2008.

[53] S. M. Hörst. Titan's atmosphere and climate, mar 2017.

[54] Hsiang-Wen Wen Hsu, Frank Postberg, Yasuhito Sekine, Takazo Shibuya, Sascha Kempf, Mihály Horányi, Antal Juhász, Nicolas Altobelli, Katsuhiko Suzuki, Yuka Masaki, Tatsu Kuwatani, Shogo Tachibana, Sin-iti Iti Sirono, Georg Moragas-Klostermeyer, and Ralf Srama. Ongoing hydrothermal activities within Enceladus. *Nature*, 519(7542):207–210, mar 2015.

[55] Lan Huang, Franz Hofer, G. Steven Martin, and Sung-Hou Kim. Structural basis for the interaction of Ras with RaIGDS. *Nature Structural Biology*, 5(6):422–426, June 1998. Number: 6 Publisher: Nature Publishing Group.

[56] Barry Hurley, Barry O'Sullivan, David Allouche, George Katsirelos, Thomas Schiex, Matthias Zytnicki, and Simon de Givry. Multi-language evaluation of exact solvers in graphical model discrete optimization. *Constraints*, 21(3):413–434, July 2016.

[57] Toshiko Ichiye. What makes proteins work: Exploring life in P-T-X. *Physical Biology*, 13(6), 2016.

[58] Toshiko Ichiye. Enzymes from piezophiles. *Seminars in Cell and Developmental Biology*, 84:138–146, 2018.

[59] L. Iess, R. A. Jacobson, M. Ducci, D. J. Stevenson, J. I. Lunine, J. W. Armstrong, S. W. Asmar, P. Racioppa, N. J. Rappaport, and P. Tortora. The Tides of Titan. *Science*, 337(6093):457–459, 2012.

[60] L Iess, D J Stevenson, M Parisi, D Hemingway, R A Jacobson, J I Lunine, F Nimmo, J W Armstrong, S W Asmar, M Ducci, and P Tortora. The gravity field and interior structure of Enceladus. *Science*, 344(6179):78–80, apr 2014.

[61] Hiroshi Ito, Shinji Watanabe, Anthony Sanchez, Michael A Whitt, and Yoshihiro Kawaoka. Mutational analysis of the putative fusion domain of ebola virus glycoprotein. *Journal of virology*, 73(10):8907–8912, 1999.

[62] Matthew P. Jacobson, Richard A. Friesner, Zhexin Xiang, and Barry Honig. On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology*, 320(3):597–608, 2002.

[63] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. Day, Barry Honig, David E. Shaw, and Richard A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.

[64] R. A. Jacobson, P. G. Antreasian, J. J. Bordi, K. E. Criddle, R Ionasescu, J. B. Jones, R. A. Mackenzie, M. C. Meek, D Parcher, F. J. Pelletier, W. M. Owen, Jr., D. C. Roth, I. M. Roundhill, and J. R. Stauch. The Gravity Field of the Saturnian System from Satellite Observations and Spacecraft Tracking Data. *The Astronomical Journal*, 132(6):2520–2526, 2006.

[65] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, February 2019.

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

[67] B. Journaux, I. Daniel, R. Caracas, G. Montagnac, and H. Cardon. Influence of NaCl on ice VI and ice VII melting curves up to GPa, implications for large icy moons. *Icarus*, 226(1):355–363, sep 2013.

[68] W Kabsch and C Sander. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22(12):2577–2637, 1983.

[69] Masahiro Kajihara, Eri Nakayama, Andrea Marzi, Manabu Igarashi, Heinz Feldmann, and Ayato Takada. Novel mutations in marburg virus glycoprotein associated with viral evasion from antibody mediated immune pressure. *Journal of General Virology*, 94(Pt 4):876–883, 2013.

[70] Andrey Karshikoff, Lennart Nilsson, and Rudolf Ladenstein. Rigidity versus flexibility: The dilemma of understanding protein thermal stability, 2015.

[71] P. L. Kastritis and A. M. Bonvin. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.*, 9(5):2216–2225, May 2010.

[72] Vadim V. Kevbrin, Tatjana N. Zhilina, Fred A. Rainey, and George A. Zavarzin. Tindallia magadii gen. nov., sp. nov.: An alkaliphilic anaerobic ammonifier from Soda Lake Deposits. *Current Microbiology*, 37(2):94–100, 1998.

[73] K K Khurana, M G Kivelson, D J Stevenson, G Schubert, C T Russell, R J Walker, and C Polanskey. Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto. *Nature*, 395(6704):777–780, oct 1998.

[74] M. G. Kivelson, K. K. Khurana, C. T. Russell, R. J. Walker, J. Warnecke, F. V. Coroniti, C. Polanskey, D. J. Southwood, and G. Schubert. Discovery of Ganymede's magnetic field by the Galileo spacecraft. *Nature*, 384(6609):537–541, dec 1996.

[75] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjærgaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, and Paolo Marcatili. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

[76] F Kostolanskỳ, E Varečková, T Betáková, V Mucha, G Russ, and SA Wharton. The strong positive correlation between effective affinity and infectivity neutralization of highly cross-reactive monoclonal antibody iib4, which recognizes antigenic site b on influenza a virus haemagglutinin. *Journal of General Virology*, 81(7):1727–1735, 2000.

[77] Brett M. Kroncke, Amanda M. Duran, Jeffrey L. Mendenhall, Jens Meiler, Jeffrey D. Blume, and Charles R. Sanders. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. *Biochemistry*, 55(36):5002–5009, 2016.

[78] Terry Ann Krulwich and Masahiro Ito. *Alkaliphilic Prokaryotes*. Springer, Berlin, 2013.

[79] Snehal Kulkarni, Kusum Dhakar, and Amaraja Joshi. Alkaliphiles: Diversity and Bioprospection. *Microbial Diversity in the Genomic Era*, pages 239–263, jan 2019.

[80] Su Datt Lam, Natalie L Dawson, Sayoni Das, Ian Sillitoe, Paul Ashford, David Lee, Sonja Lehtinen, Christine A Orengo, and Jonathan G Lees. Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research*, 44, 2016.

[81] Jeffrey E. Lee, Marnie L. Fusco, Ann J. Hessell, Wendelien B. Oswald, Dennis R. Burton, and Erica Ollmann Saphire. Structure of the ebola virus glycoprotein bound to an antibody from a human survivor. *Nature*, 454(7201):177–182, July 2008.

[82] Kuo Hao Lee, Craig R. Miller, Anna C. Nagel, Holly A. Wichman, Paul Joyce, and F. Marty Ytreberg. First-step mutations for adaptation at elevated temperature increase capsid stability in a virus. *PLOS ONE*, 6:1–8, 09 2011.

[83] Adam R. Leman and Eishi Noguchi. The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. *Genes*, 4(1):1–32, Mar 2013.

[84] N. J. Lennemann, M. Walkner, A. R. Berkebile, N. Patel, and W. Maury. The role of conserved n-linked glycans on ebola virus glycoprotein 2. *Journal of Infectious Diseases*, June 2015.

[85] Eric M Leroy, Pierre Rouquet, Pierre Formenty, Sandrine Souquiere, Annelisa Kilbourne, Jean-Marc Froment, Magdalena Bermejo, Sheilag Smit, William Karesh, Robert Swanepoel, et al. Multiple ebola virus transmission events and rapid decline of central african wildlife. *Science*, 303(5656):387–390, 2004.

[86] Shiqing Li, Karl R. Schmitz, Philip D. Jeffrey, Jed J. W. Wiltzius, Paul Kussie, and Kathryn M. Ferguson. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell*, 7(4):301–311, April 2005. Publisher: Elsevier.

[87] Daniel Lim, Hyeon Ung Park, Liza De Castro, Sung Gyun Kang, Hyun Sook Lee, Susan Jensen, Kye Joon Lee, and Natalie C. J. Strynadka. Crystal structure and kinetic analysis of β-lactamase

inhibitor protein-II in complex with TEM-1 β-lactamase. *Nature Structural Biology*, 8(10):848–852, October 2001. Number: 10 Publisher: Nature Publishing Group.

[88] Song Liu, Chi Zhang, Hongyi Zhou, and Yaoqi Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101, April 2004.

[89] Ralph D Lorenz, Karl L Mitchell, Randolph L Kirk, Alexander G Hayes, Oded Aharonson, Howard A Zebker, Phillipe Paillou, Jani Radebaugh, Jonathan I Lunine, Michael A Janssen, Stephen D Wall, Rosaly M Lopes, Bryan Stiles, Steve Ostro, Giuseppe Mitri, and Ellen R Stofan. Titan's inventory of organic surface materials. *Geophysical Research Letters*, 35(2), 2008.

[90] Jonathan I. Lunine. Ocean worlds exploration. *Acta Astronautica*, 131(September 2016):123–130, 2017.

[91] Jonathan I. Lunine and David J. Stevenson. Clathrate and ammonia hydrates at high pressure: Application to the origin of methane on Titan. *Icarus*, 70(1):61–77, apr 1987.

[92] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, 2013.

[93] Kathleen E Mandt, J. Hunter Waite, Benjamin Teolis, Brian A Magee, Jared Bell, Joseph H Westlake, Conor A Nixon, Olivier Mousis, and Jonathan I Lunine. The 12C/13C ratio on titan from Cassini INMS measurements and implications for the evolution of methane. *Astrophysical Journal*, 749(2), 2012.

[94] Kyle P. Martin, Shannon M. MacKenzie, Jason W. Barnes, and F. Marty Ytreberg. Protein stability in titan's subsurface water ocean. *Astrobiology*, 20(2):190–198, 2020. PMID: 31730377.

[95] Toshiaki Maruyama, Luis L. Rodriguez, Peter B. Jahrling, Anthony Sanchez, Ali S. Khan, Stuart T. Nichol, C. J. Peters, Paul W. H. I. Parren, and Dennis R. Burton. Ebola virus can be effectively neutralized by antibody produced in natural human infection. *Journal of Virology*, 73(7):6024–6030, 1999.

[96] Andrea Marzi, Flora Engelmann, Friederike Feldmann, Kristen Haberthur, W Lesley Shupert, Douglas Brining, Dana P Scott, Thomas W Geisbert, Yoshihiro Kawaoka, Michael G Katze, et al. Antibodies are necessary for rvsv/zebov-gp–mediated protection against lethal ebola virus challenge in nonhuman primates. *Proceedings of the National Academy of Sciences*, 110(5):1893–1898, 2013.

[97] Wayne Materi and David S. Wishart. Computational systems biology in drug discovery and development: methods and applications. *Drug Discovery Today*, 12(7):295–303, 2007.

[98] Christopher Mclendon, F. Jeffrey Opalko, Heshan I. Illangkoon, and Steven A. Benner. Solubility of Polyethers in Hydrocarbons at Low Temperatures. A Model for Potential Genetic Backbones on Warm Titans. *Astrobiology*, 15(3):200–206, mar 2015.

[99] Nicola A. G. Meenan, Amit Sharma, Sarel J. Fleishman, Colin J. MacDonald, Bertrand Morel, Ruth Boetzel, Geoffrey R. Moore, David Baker, and Colin Kleanthous. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proceedings of the National Academy of Sciences*, 107(22):10080–10085, June 2010. Publisher: National Academy of Sciences Section: Biological Sciences.

[100] Lars Meinhold, Jeremy C. Smith, Akio Kitao, and Ahmed H. Zewail. Picosecond fluctuating protein energy landscape mapped by pressure-temperature molecular dynamics simulation. *Proceedings of the National Academy of Sciences of the United States of America*, 104(44):17261–17265, 2007.

[101] Grégoire Michoud and Mohamed Jebbar. High hydrostatic pressure adaptive strategies in an obligate piezophile Pyrococcus yayanosii. *Scientific Reports*, 6(May):1–10, 2016.

[102] Craig R. Miller, Kuo Hao Lee, Holly A. Wichman, and F. Marty Ytreberg. Changing folding and binding stability in a viral coat protein: A comparison between substitutions accessible through mutation and those fixed by natural selection. *PLoS ONE*, 9(11):e112988, November 2014.

[103] Yves A. Muller, Yvonne Chen, Hans W. Christinger, Bing Li, Brian C. Cunningham, Henry B. Lowman, and Abraham M. de Vos. VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 å resolution and mutational analysis of the interface. *Structure*, 6(9):1153–1167, September 1998. Publisher: Elsevier.

[104] National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov/.../ebola/`.

[105] Catherine D. Neish, Árpád Somogyi, and Mark A. Smith. Titan's Primordial Soup: Formation of Amino Acids via Low-Temperature Hydrolysis of Tholins. *Astrobiology*, 10(3):337–347, apr 2010.

[106] C.D. Neish, Á. Somogyi, H. Imanaka, J.I. Lunine, and M.A. Smith. Rate Measurements of the Hydrolysis of Complex Organic Macromolecules in Cold Aqueous Solutions: Implications for Prebiotic Chemistry on the Early Earth and Titan. *Astrobiology*, 8(2):273–287, apr 2008.

[107] H. B. Niemann, S. K. Atreya, J. E. Demick, D. Gautier, J. A. Haberman, D. N. Harpold, W. T. Kasprzak, J. I. Lunine, T. C. Owen, and F. Raulin. Composition of Titan's lower atmosphere and simple surface volatiles as measured by the Cassini-Huygens probe gas chromatograph mass spectrometer experiment. *Journal of Geophysical Research E: Planets*, 115(12):E12006, dec 2010.

[108] E G Nisbet and C M R Fowler. Archaean metabolic evolution of microbial mats. *Proceedings of the Royal Society B: Biological Sciences*, 266(1436):2375–2382, 1999.

[109] C A Nixon, B Temelso, S Vinatier, N A Teanby, B Bézard, R K Achterberg, K E Mandt, C D Sherrill, P. G.J. Irwin, D E Jennings, P N Romani, A Coustenis, and F M Flasar. Isotopic ratios in titan's methane: Measurements and modeling. *Astrophysical Journal*, 749(2), 2012.

[110] Abayomi S. Olabode, Xiaowei Jiang, David L. Robertson, and Simon C. Lovell. Ebolavirus is evolving but not changing: No evidence for functional change in EBOV from 1976 to the 2014 outbreak. *Virology*, 482:202–207, August 2015.

[111] M.J. O'Neil. *The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals*. Merck, Whitehouse Station N.J., 14th ed. edition, 2013.

[112] Wendelien B Oswald, Thomas W Geisbert, Kelly J Davis, Joan B Geisbert, Nancy J Sullivan, Peter B Jahrling, Paul W. H. I Parren, and Dennis R Burton. Neutralizing antibody fails to impact the course of ebola virus infection in monkeys. *PLOS Pathogens*, 3(1):1–5, 01 2007.

[113] E. A. Padlan, E. W. Silverton, S. Sheriff, G. H. Cohen, S. J. Smith-Gill, and D. R. Davies. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proceedings of the National Academy of Sciences*, 86(15):5938–5942, August 1989. Publisher: National Academy of Sciences Section: Research Article.

[114] Anastassios C. Papageorgiou, Robert Shapiro, and K.Ravi Acharya. Molecular recognition of human angiogenin by placental ribonuclease inhibitor—an X-ray crystallographic study at 2.0 Å resolution. *The EMBO Journal*, 16(17):5162–5177, September 1997. Publisher: John Wiley & Sons, Ltd.

[115] Daniel J Park, Gytis Dudas, Shirlee Wohl, Augustine Goba, Shannon LM Whitmer, Kristian G Andersen, Rachel S Sealfon, Jason T Ladner, Jeffrey R Kugelman, Christian B Matranga, et al. Ebola virus epidemiology, transmission, and evolution during seven months in sierra leone. *Cell*, 161(7):1516–1526, 2015.

[116] Hahnbeom Park, Philip Bradley, Per Greisen, Jr, Yuan Liu, Vikram Khipple Mulligan, David E. Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy function

on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201, December 2016.

[117] Paul W. H. I. Parren, Tom W. Geisbert, Toshiaki Maruyama, Peter B. Jahrling, and Dennis R. Burton. Pre- and postexposure prophylaxis of ebola virus infection in an animal model by passive transfer of a neutralizing human antibody. *Journal of Virology*, 76(12):6408–6412, 2002.

[118] Lata Prasad, E. Bruce Waygood, Jeremy S Lee, and Louis T. J Delbaere. The 2.5 Å resolution structure of the Jel42 Fab fragment/HPr complex11Edited by I. A. Wilson. *Journal of Molecular Biology*, 280(5):829–845, July 1998.

[119] V. P. Prasher, Matthew J. Farrer, Anna M. Kessling, Elizabeth M. C. Fisher, R. J. West, P. C. Barber, and A. C. Butler. Molecular mapping of alzheimer-type dementia in down's syndrome. *Annals of Neurology*, 43(3):380–383, 1998.

[120] Laura Preiss, David B. Hicks, Shino Suzuki, Thomas Meier, and Terry Ann Krulwich. Alkaliphilic bacteria with impact on industrial applications, concepts of early life forms, and bioenergetics of ATP synthesis, jun 2015.

[121] X. Qiu, J. Audet, G. Wong, S. Pillet, A. Bello, T. Cabral, J. E. Strong, F. Plummer, C. R. Corbett, J. B. Alimonti, and G. P. Kobinger. Successful treatment of ebola virus-infected cynomolgus macaques with monoclonal antibodies. *Science Translational Medicine*, 4(138):138ra81–138ra81, June 2012.

[122] Xiangguo Qiu, Judie B Alimonti, P Leno Melito, Lisa Fernando, Ute Ströher, and Steven M Jones. Characterization of zaire ebolavirus glycoprotein-specific monoclonal antibodies. *Clinical immunology*, 141(2):218–227, 2011.

[123] Xiangguo Qiu, Gary Wong, Jonathan Audet, Alexander Bello, Lisa Fernando, Judie B Alimonti, Hugues Fausther-Bovendo, Haiyan Wei, Jenna Aviles, Ernie Hiatt, et al. Reversion of advanced ebola virus disease in nonhuman primates with zmapp. *Nature*, 2014.

[124] Jean-Paul Renaud, Chun-wa Chung, U. Helena Danielson, Ursula Egner, Michael Hennig, Roderick E. Hubbard, and Herbert Nar. Biophysics in drug discovery: impact, challenges and opportunities. *Nature Reviews Drug Discovery*, 15(10):679–698, 2016.

[125] Andrej Sali and Tom L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[126] Lucas Sawle and Kingshuk Ghosh. How do thermophilic proteins and proteomes withstand high temperature? *Biophysical Journal*, 101(1):217–227, 2011.

[127] Axel J. Scheidig, Thomas R. Hynes, Laura A. Pelletier, James A. Wells, and Anthony A. Kossiakoff. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of alzheimer's amyloid β-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): Engineering of inhibitors with altered specificities. *Protein Science*, 6(9):1806–1824, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560060902.

[128] Stephanie J Schrag, Paul A Rota, and William J Bellini. Spontaneous mutation rate of measles virus: direct estimation based on mutations conferring monoclonal antibody resistance. *Journal of virology*, 73(1):51–54, 1999.

[129] Joost W H Schymkowitz, Frederic Rousseau, Ivo C Martins, Jesper Ferkinghoff-Borg, Francois Stricher, and Luis Serrano. Prediction of water and metal binding sites and their affinities by using the fold-x force field. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10147–10152, July 2005.

[130] Everett L. Shock and William B. McKinnon. Hydrothermal Processing of Cometary Volatiles-Applications to Triton. *Icarus*, 106(2):464–477, dec 1993.

[131] KS Siddiqui and T Thomas. *Protein Adaptation in Extremophiles*. Nova Science, New York, 2008.

[132] Ian Sillitoe, Tony E Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L Dawson, Nicholas Furnham, Roman A Laskowski, David Lee, Jonathan G Lees, Sonja Lehtinen, Romain A Studer, Janet Thornton, and Christine A Orengo. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43, 2015.

[133] Francesca Simonato, Stefano Campanaro, Federico M. Lauro, Alessandro Vezzi, Michela D'Angelo, Nicola Vitulo, Giorgio Valle, and Douglas H. Bartlett. Piezophilic adaptation: a genomic point of view, 2006.

[134] Derek J Smith, Alan S Lapedes, Jan C de Jong, Theo M Bestebroer, Guus F Rimmelzwaan, Albert DME Osterhaus, and Ron AM Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[135] P. H. A. Sneath. Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology*, 12(2):157–195, November 1966.

[136] Cláudio M. Soares, Vitor H Teixeira, and António M. Baptista. Protein structure and dynamics in nonaqueous solvents: Insights from molecular dynamics simulation studies. *Biophysical Journal*, 84(3):1628–1641, 2003.

[137] Satoshi Sogabe, Fiona Stuart, Christoph Henke, Angela Bridges, Geoffrey Williams, Ashley Birch, Fritz K. Winkler, and John A. Robinson. Neutralizing epitopes on the extracellular interferon γ receptor (IFNγR) α-chain characterized by homolog scanning mutagenesis and X-ray crystal structure of the A6 Fab-IFNγR1-108 complex11Edited by R. Huber. *Journal of Molecular Biology*, 273(4):882–897, November 1997.

[138] Christophe Sotin and Gabriel Tobie. Internal structure and dynamics of the large icy satellites, 2004.

[139] Vojtech Spiwok, Zoran Sucur, and Petr Hosek. Enhanced sampling techniques in biomolecular simulations. *Biotechnology Advances*, 33(6, Part 2):1130–1140, 2015.

[140] Thomas Splettstoesser. Cc by-sa 3.0.

[141] Daphne A Stanley, Anna N Honko, Clement Asiedu, John C Trefry, Annie W Lau-Kilby, Joshua C Johnson, Lisa Hensley, Virginia Ammendola, Adele Abbate, Fabiana Grazioli, et al. Chimpanzee adenovirus vaccine generates acute and durable protective immunity against ebolavirus challenge. *Nature medicine*, 2014.

[142] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE*, 8(11):e80635, November 2013.

[143] G Tobie, M Choukroun, O Grasset, S Le Mouélic, J.I Lunine, C Sotin, O Bourgeois, D Gautier, M Hirtzig, S Lebonnois, and L Le Corre. Evolution of Titan and implications for its hydrocarbon cycle. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1889):617–631, feb 2008.

[144] Gabriel Tobie, Jonathan I. Lunine, and Christophe Sotin. Episodic outgassing as the origin of atmospheric methane on Titan. *Nature*, 440(7080):61–64, mar 2006.

[145] M G Tomasko, L Doose, S Engel, L E Dafoe, R West, M Lemmon, E Karkoschka, and C See. A model of Titan's aerosols based on measurements made inside the atmosphere. *Planetary and Space Science*, 56(5):669–707, 2008.

[146] Yi-Gang Tong, Wei-Feng Shi, Di Liu, Jun Qian, Long Liang, Xiao-Chen Bo, Jun Liu, Hong-Guang Ren, Hang Fan, Ming Ni, Yang Sun, Yuan Jin, Yue Teng, Zhen Li, David Kargbo, Foday Dafae, Alex Kanu, Cheng-Chao Chen, Zhi-Heng Lan, Hui Jiang, Yang Luo, Hui-Jun Lu, Xiao-Guang Zhang, Fan Yang, Yi Hu, Yu-Xi Cao, Yong-Qiang Deng, Hao-Xiang Su, Yu Sun, Wen-Sen Liu, Zhuang Wang, Cheng-Yu Wang, Zhao-Yang Bu, Zhen-Dong Guo, Liu-Bo Zhang, Wei-Min Nie, Chang-Qing Bai, Chun-Hua Sun, Xiao-Ping An, Pei-Song Xu, Xiang-Li-Lan Zhang, Yong Huang, Zhi-Qiang Mi, Dong Yu, Hong-Wu Yao, Yong Feng, Zhi-Ping Xia, Xue-Xing Zheng, Song-Tao Yang, Bing Lu, Jia-Fu Jiang, Brima Kargbo, Fu-Chu He, George F. Gao, Wu-Chun Cao, Yi-Gang Tong, Jun Qian, Yang Sun, Hui-Jun Lu, Xiao-Guang Zhang, Fan Yang, Yi Hu, Yu-Xi Cao, Yong-Qiang Deng, Hao-Xiang Su, Yu Sun, Wen-Sen Liu, Zhuang Wang, Cheng-Yu Wang, Zhao-Yang Bu, Zhen-Dong Guo, Liu-Bo Zhang, Wei-Min Nie, Chang-Qing Bai, Chun-Hua Sun, Yong Feng, Jia-Fu Jiang, and George F. Gao. Genetic diversity and evolutionary dynamics of ebola virus in sierra leone. *Nature*, May 2015.

[147] Wouter G. Touw, Coos Baakman, Jon Black, Tim A H Te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368, 2015.

[148] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, February 2014. arXiv: 1310.0980.

[149] Mark Ultsch, Jack Bevers, Gerald Nakamura, Richard Vandlen, Robert F. Kelley, Lawren C. Wu, and Charles Eigenbrot. Structural Basis of Signaling Blockade by Anti-IL-13 Antibody Lebrikizumab. *Journal of Molecular Biology*, 425(8):1330–1339, April 2013.

[150] G. C. P. van Zundert, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. de Vries, and A. M. J. J. Bonvin. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4):720–725, February 2016.

[151] Steven D. Vance, Mark P. Panning, Simon Stähler, Fabio Cammarano, Bruce G. Bills, Gabriel Tobie, Shunichi Kamata, Sharon Kedar, Christophe Sotin, William T. Pike, Ralph Lorenz, Hsin Hua Huang, Jennifer M. Jackson, and Bruce Banerdt. Geophysical Investigations of Habitability in Ice-Covered Ocean Worlds. *Journal of Geophysical Research: Planets*, 123(1):180–205, jan 2018.

[152] Clément Viricel, Simon de Givry, Thomas Schiex, and Sophie Barbe. Cost function network-based design of protein–protein interactions: predicting changes in binding affinity. *Bioinformatics*, 34(15):2581–2589, August 2018.

[153] B. F. Volkman, S. L. Alam, J. D. Satterlee, and J. L. Markley. Solution structure and backbone dynamics of component IV Glycera dibranchiata monomeric hemoglobin-CO. *Biochemistry*, 37(31):10906–10919, Aug 1998.

[154] V Vuitton, P Lavvas, R V Yelle, M Galand, A Wellbrock, G R Lewis, A J Coates, and J. E. Wahlund. Negative ion chemistry in Titan's upper atmosphere. *Planetary and Space Science*, 57(13):1558–1572, 2009.

[155] J H Waite, D T Young, T E Cravens, A J Coates, F J Crary, B Magee, and J Westlake. The process of tholin formation in Titan's upper atmosphere. *Science (New York, N.Y.)*, 316(5826):870–5, may 2007.

[156] Tsjerk A. Wassenaar, Marc van Dijk, Nuno Loureiro-Ferreira, Gijs van der Schot, Sjoerd J. de Vries, Christophe Schmitz, Johan van der Zwan, Rolf Boelens, Andrea Giachetti, Lucio Ferella, Antonio Rosato, Ivano Bertini, Torsten Herrmann, Hendrik R. A. Jonker, Anurag Bagaria, Victor Jaravine, Peter Güntert, Harald Schwalbe, Wim F. Vranken, Jurgen F. Doreleijers, Gert Vriend, Geerten W. Vuister, Daniel Franke, Alexey Kikhney, Dmitri I. Svergun, Rasmus H. Fogh, John Ionides, Ernest D. Laue, Chris Spronk, Simonas Jurkša, Marco Verlato, Simone Badoer, Stefano Dal Pra, Mirco Mazzucato, Eric Frizziero, and Alexandre M. J. J. Bonvin. WeNMR: Structural Biology on the Grid. *Journal of Grid Computing*, 10(4):743–767, December 2012.

[157] Shinji Watanabe, Ayato Takada, Tokiko Watanabe, Hiroshi Ito, Hiroshi Kida, and Yoshihiro Kawaoka. Functional importance of the coiled-coil of the ebola virus glycoprotein. *Journal of virology*, 74(21):10194–10201, 2000.

[158] B. Webb and A. Sali. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*, 54:1–5, 06 2016.

[159] Ralph Weissleder. Scaling down imaging: molecular mapping of cancer in mice. *Nature Reviews Cancer*, 2(1):11–18, 2002.

[160] E. H. Wilson and S. K. Atreya. Current state of modeling the photochemistry of Titan's mutually dependent atmosphere and ionosphere. *Journal of Geophysical Research E: Planets*, 109(6):E06002, jun 2004.

[161] J. A. Wilson, M. Hevey, R. Bakken, S. Guest, M. Bray, A. L. Schmaljohn, and M. K. Hart. Epitopes involved in antibody-mediated protection from ebola virus. *Science (New York, N.Y.)*, 287(5458):1664–1666, March 2000.

[162] Tatiana J Wittmann, Roman Biek, Alexandre Hassanin, Pierre Rouquet, Patricia Reed, Philippe Yaba, Xavier Pourrut, Leslie A Real, Jean-Paul Gonzalez, and Eric M Leroy. Isolates of zaire ebolavirus from wild apes reveal genetic lineage and recombinants. *Proceedings of the National Academy of Sciences*, 104(43):17123–17127, 2007.

[163] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. *Journal of Molecular Biology*, 429(3):426–434, February 2017.

[164] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, 12(1):7–8, Jan 2015.

[165] Yuqi Yu, Jinan Wang, Qiang Shao, Jiye Shi, and Weiliang Zhu. The effects of organic solvents on the folding pathway and associated thermodynamics of proteins: A microscopic view. *Scientific Reports*, 6, 2016.

[166] Y. L. Yung, M. Allen, and J. P. Pinto. Photochemistry of the atmosphere of Titan - Comparison between model and observations. *The Astrophysical Journal Supplement Series*, 55:465, jul 1984.

[167] Yingqian Ada Zhan and F. Marty Ytreberg. The cis conformation of proline leads to weaker binding of a p53 peptide to mdm2 compared to trans. *Archives of Biochemistry and Biophysics*, 575:22–29, 06 2015.

[168] Xiaodong Zhao, Enmei Liu, Fu-Ping Chen, and Wayne M Sullender. In vitro and in vivo fitness of respiratory syncytial virus monoclonal antibody escape mutants. *Journal of virology*, 80(23):11651–11657, 2006.