Using modeling to understand structure-function relationships in proteins

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

Jonathan E. Barnes

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

Committee Members: Matthew M. Hedman, Ph.D.; Andreas E. Vasdekis, Ph.D.; Craig R. Miller, Ph.D.

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

December 2021

Authorization to Submit Dissertation

This Dissertation of Jonathan E. Barnes, submitted for the degree of Doctor of Philosophy with a Major in Physics and titled "Using modeling to understand structure-function relationships in proteins," 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:	Matthew M. Hedman, Ph.D.	Date
	Andreas E. Vasdekis, Ph.D.	Date
	Craig R. Miller, Ph.D.	Date
Department Chair:	John R. Hiller, Ph.D.	Date

Abstract

Proteins are critical to the function of cells and to life. It is well established that changes to the DNA sequence (genotype) of a protein can have a significant impact on how they function or interact within the cell. Understanding the mapping between changes in a protein genotype and how those changes modify an organism phenotype is a largely unsolved problem in biology. Solving this problem will require integration of experimental methods with computational and mathematical approaches. In this thesis, we utilize both computational and mathematical methodologies. We start by using statistical methods to investigate potential physical features that can explain epistasis in proteins. Here we find a number of intuitive features that play a role, but we can only explain ∼30% of the observed epistasis in both protein binding and folding. Next, we use molecular dynamics to inform statistical models and predict the spectral sensitivity of opsin proteins with high accuracy. Following that, we investigate a suite of fast methods for predicting protein-protein binding affinity, finding their performance to be largely context dependent. Lastly, we explore using two different molecular modeling techniques to calculate free energies and build a watch list of antibody escape mutations for the current COVID-19 pandemic.

Acknowledgements

Most importantly I would like to thank the members of my committee. Dr. F. Marty Ytreberg has been my major professor and mentor the past 5 years. He has been instrumental in my success, supporting and encouraging me every step of the way, without whom I wouldn't be here defending today. I thank Matthew M. Hedman for his wide breadth of knowledge and expertise he brings to the committe. I thank Craig R. Miller for his pateience, statistical expertise, and assistance with co-authoring my first first-author manuscript. I thank Andreas E. Vasdekis for his expertise and perspective he brings to the committee. 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 Epistasis paper was supported by the Center for Modeling Complex Interactions sponsored by the National Institute General Medical Sciences (https://www.nigms.nih.gov) under award number P20 GM104420 and the National Science Foundation (https://www.nsf.gov) EPSCoR TrackII under award number OIA1736253. Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (NIH P30 GM103324). The Sws2 paper was supported by the Center for Modeling Complex Interactions (CMCI) sponsored by the NIGMS under award number NIH P20 GM104420 and was also supported in part by National Science Foundation EPSCoR Track-II grant under award number OIA1736253 and was also supported in part by NIH R01 EY012146 and NSF DEB 1638567. Computer resources were provided by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (NIH P30 GM103324). This research also made use of the computational resources provided by the high-performance computing center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. DOE and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517. WILD was supported by the Australian Research Council (ARC) in the form of a Future Fellowship (FT110100176) and a Discovery Project grant (DP140102117), and is currently supported by a JC Kempe Memorial Scholarship from the Kempe Foundation, Sweden. The Methods Paper was supported by the National Science Foundation (OIA1736253) and the Institute for Modeling Collaboration and Innovation (P20 GM104420). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this dissertation.

I would also like to thank Jagdish Patel, Chris Mirabzadeh, Peik Lund, Caleb Quates, Kyle Martin, Casey Beard, Tawny Gonzalez, and Dharmesh Patel.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Statistical and molecular modeling of proteins

Proteins play a crucial role in nearly all biological processes in cells. They are mediators for processes such as energy transport, immune response, and color vision. When proteins are generated in a cell, it is possible for errors to occur in their sequence; called amino acid substitutions or mutations. When these mutations occur, the proteins can malfunction and lead to disease or alternatively provide better immune response or other positive outcomes. Conversely, when mutations occur in viruses or antigens, they can change in such a way to escape immune response yet still be functional enough to wreak havoc on their host. These points highlight that understanding and predicting how proteins function and interact — especially as they evolve and mutate — is key to developing better drugs and therapeutics. While there are many existing experimental methods designed to better understand proteins, they are often limited by low throughput and resolution and are unable to determine atomistic mechanisms.

Molecular modeling provides a conduit to understand mechanisms that experimental methods do not have the resolution to decipher. This form of modeling can be used to simulate protein systems under physiological condition at atomistic resolution, with force field inaccuracy being one of the greatest limitations. Molecular dynamics is a form of molecular modeling that uses full physical descriptors of atoms, bonds, and forces to replicate how a given system progresses in its environment (e.g., membrane, water and ions). One caveat however, is that molecular dynamics can require long simulation times and large computation power to obtain accurate results. While still less costly compared to some experimental methods, these simulations can still take time on the order of days, weeks, or months depending on the system and desired outcomes.

Statistical modeling on the other hand relies on a priori information and statistical principles rather than rigorous simulation, thus saving time. Such methods can be used to predict the effects of mutations on proteins for quantifiable phenotypes, like binding affinity. Statistical methods can also be used to build mechanistic pictures based on what elements are most important or contribute most strongly to a given effect. While these methods are fast, they tend to exclude relevant information like flexibility or conformation — a limitation since proteins are dynamic systems. To improve on this, it is possible to utilize the rigor of molecular dynamics simulations and resulting trajectories on the wildtype structure as input files for statistical methods to ascertain the effects of mutation sets quickly.

1.2 Quantification of protein function and interaction

All proteins start out as an unfolded string of amino acids. In order to function, most proteins must fold into a well-defined 3D structure. There is an energy associated with this folding process that can be quantified by comparing the Gibbs free energy (G) between the folded and unfolded state. Statistical mechanics tells us that systems in equilibrium will spend more time in lower free energy states. For most proteins to be functional, the folded state must have a lower Gibbs free energy than the unfolded state. Said another way, the Gibbs free energy difference between the folded and unfolded states determines stability and can be given by Equation 1.1 and shown visually by Figure 1.1:

$$\Delta G_{\text{fold}} = G_{\text{folded}} - G_{\text{unfolded}}. \tag{1.1}$$

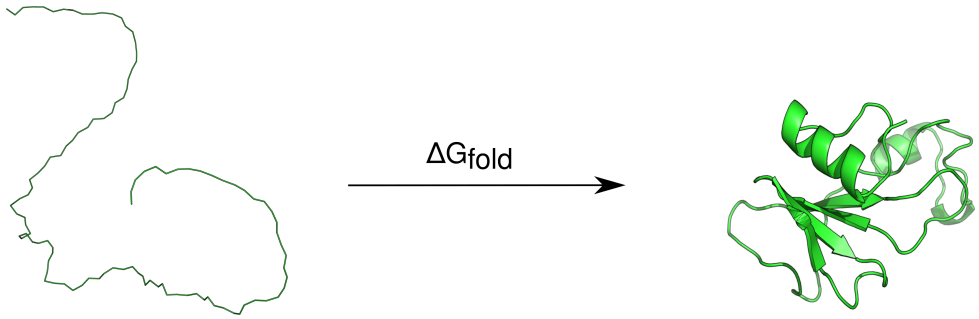

Figure 1.1: Example of change in the Gibbs free energy due to protein folding

Analogous to folding, proteins can interact with other bio-molecules by binding to them with a certain free energy. In order for an antibody to fulfill it's purpose of inhibiting an antigen it needs to be able to bind to same receptors said antigen uses to affect a human, or animal, cell. As a corollary to folding, we can use the change in Gibbs free energy of binding (also called binding affinity) as a metric for binding strength. The free energy difference will indicate whether the bound state is preferred or not (Equation 1.2, Figure 1.2):

$$\Delta G_{\text{bind}} = G_{\text{bound}} - G_{\text{unbound}}. \tag{1.2}$$

Here, ΔG_{bind} is the change in the Gibbs free energy of binding and G_{bound}, G_{unbound} correspond to the bound and unbound state respectively.

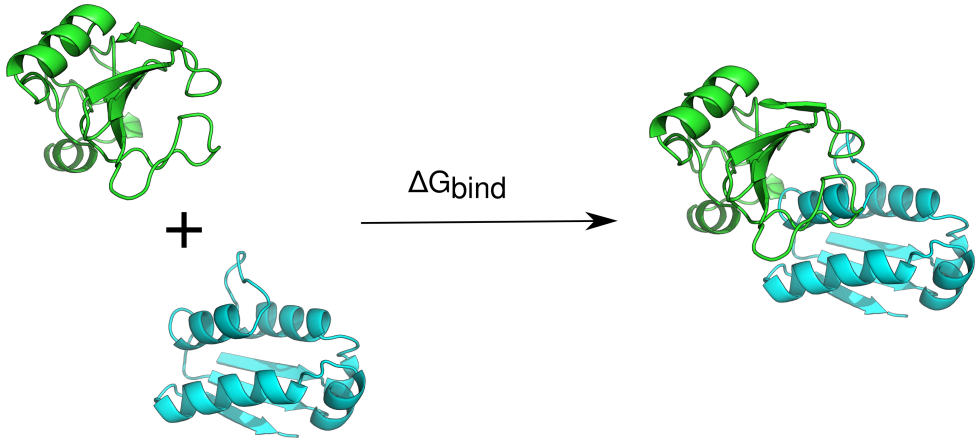

Figure 1.2: Example of change in binding free energy (affinity)

We are particularly interested in how these free energy differences change upon mutation. Most notably, whether or not a given mutation will disrupt binding or folding. To calculate this, we compare the difference in the changes of free energy, or the $\Delta\Delta$G value. More specifically:

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}. \tag{1.3}$$

Where "mut" and "wt" correspond to the mutant and wild-type sequences respectively. A smaller value of $\Delta\Delta$G for folding/binding indicates that the mutation has very little impact on the ability of the protein to fold or bind.

1.3 Linear models and their applications on protein function

Statistical methods can help simplify and explain complex phenomena, including complex aspects of protein interactions that may only be dependent on structural effects or intrinsic attributes. For those cases, it's possible to use molecular dynamics to develop a picture of the conformational changes the molecule goes through or utilize a priori information about the proteins or amino acids themselves as inputs for statistical models. Linear models assume that a given response variable is directly proportional to a sum of independent variables, scaled by a coefficient. More succinctly,

$$Y = c_1 \cdot X_1 + c_2 \cdot X_2 + c_3 \cdot X_3 + \dots \tag{1.4}$$

where Y is the response variable, c_i are scalar coefficients, and X_i is the value of a given independent variable. The assumption here is that the response variables are truly independent with no multicollinearity. In the case of proteins, the response variable could be the Gibbs free energy difference, as is the case

with some the aforementioned models that are mostly properties like the entropy, VDW forces, and other metrics. Or in the case of epistasis, it can be the difference between the Gibbs free energy of different mutational states. Alternatively, it could be another phenotype of the system such as in the case of opsins where parameters generated via molecular dynamics were used as inputs to predict their spectral sensitivity.

1.4 Overview of dissertation

In this dissertation I will demonstrate how to use molecular and statistical modeling to determine structure function relationships in proteins and protein systems:

1.4.1 Searching for a mechanistic description of pairwise epistasis in protein systems

When two or more amino acid mutations occur in protein systems, they can interact in a non-additive fashion termed epistasis. One way to quantify epistasis between mutation pairs in protein systems is by using free energy differences: $\epsilon = \Delta\Delta G_{1,2}$ - ($\Delta\Delta G_1$ + $\Delta\Delta G_2$) where $\Delta\Delta G$ refers to the change in the Gibbs free energy, subscripts 1 and 2 refer to single mutations in arbitrary order and 1,2 refers to the double mutant. In this study, we explore possible biophysical mechanisms that drive pairwise epistasis in both protein-protein binding affinity and protein folding stability. Using the largest available datasets containing experimental protein structures and free energy data, we derived statistical models for both binding and folding epistasis (ϵ) with similar explanatory power (R^2) of 0.299 and 0.258, respectively. These models contain terms and interactions that are consistent with intuition. For example, increasing the Cartesian separation between mutation sites leads to a decrease in observed epistasis for both folding and binding.

1.4.2 Short-wavelength-sensitive 2 (Sws2) visual photopigment models combined with atomistic molecular simulations to predict spectral peaks of absorbance

For many species, vision is one of the most important sensory modalities for mediating essential tasks including navigation, predation and foraging, predator avoidance, and numerous social behaviors. The vertebrate visual process begins when photons of light interact with rod and cone photoreceptors that are present in the neural retina. Vertebrate visual photopigments are housed within these photoreceptor cells and are sensitive to a wide range of wavelengths that peak within the light spectrum, the latter of which is a function of the type of chromophore used and how it interacts with specific amino acid residues

found within the opsin protein sequence. Minor differences in the amino acid sequences of the opsins are known to lead to large differences in the spectral peak of absorbance (i.e. the λ_{max} value). In our prior studies, we developed a new approach that combined homology modeling and molecular dynamics simulations to gather structural information associated with chromophore conformation, then used it to generate statistical models for the accurate prediction of λ_{max} values for photopigments derived from Rh1 and Rh2 amino acid sequences. To build a model that can predict the λ_{max} using our approach presented in our prior studies, we selected a spectrally-diverse set of 11 teleost Sws2 photopigments with known amino acid sequences and λ_{max} values are known. The final first-order regression model, consisting of three terms associated with chromophore conformation, was sufficient to predict the λ_{max} of Sws2 photopigments with high accuracy.

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

Here, eight non-rigorous computational methods were assessed using eight antibody-antigen and eight non-antibody-antigen complexes for their ability to accurately predict relative binding affinities ($\Delta\Delta G$) for 654 single mutations. We found that Rosetta-based JayZ and EasyE methods classified mutations as destabilizing ($\Delta\Delta G <$ -0.5 kcal/mol) with high (83–98%) accuracy and a relatively low computational cost for non-antibody-antigen complexes. Some of the most accurate results for antibody-antigen systems came from combining molecular dynamics with FoldX with a correlation coefficient (r) of 0.46, but this was also the most computationally expensive method. Overall, our results suggest these methods can be used to quickly and accurately predict stabilizing versus destabilizing mutations but are less accurate at predicting actual binding affinities. This study highlights the need for continued development of reliable, accessible, and reproducible methods for predicting binding affinities in antibody-antigen proteins and provides a recipe for using current methods.

1.4.4 Unfinished Works

1.4.4.1 The Effect of Mutations on Binding Interactions Between the SARS-CoV-2 Receptor Binding Domain and Neutralizing Antibodies

1.4.4.2 An EvoDevo study of visual opsin dynamics and spectral modeling in salmonids

Chapter 2: Searching for a Mechanistic Description of Pairwise Epistasis in protein systems

Jonathan E. Barnes,[1,2] Craig R. Miller,[2,3], F. Marty Ytreberg,[1,2]

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

Currently under peer review for publication. As first author, I contributed the following:

- Wrote the manuscript.
- Wrote all code and scripts.
- Curated and pre-processed all data.
- Performed all statistical analysis
- Generated all tables and figures.

All code is freely available on github: https://github.com/YtrebergPatelLab/EpistasisStats.

2.1 introduction

Multiple amino acid mutations can interact in biological systems, leading to nonadditive effects termed epistasis. While a general understanding of the concept of epistasis has existed for many years, the prevalence of epistasis, or its importance in biological systems, is still a matter of debate [3, 4, 5, 6, 7]. Some believe it is a major force in evolution, either by constraining the available pathways for systems to evolve, by counteracting mutations that reduce fitness through compensatory effects, or by contributing to a more rugged fitness landscape [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Others have explored the epistatic effect between sets of beneficial mutations, finding that epistasis is pervasive and a key aspect of adaption, but leading to diminishing returns or negative epistasis [12, 21, 22, 23, 24]. Other studies using RNA viruses have shown that epistasis is prevalent and likely a mechanism for their evolution [25, 26, 27, 28, 29, 30]. Epistasis has also been shown to be a likely contributing factor to drug and antibody resistance of influenza A, HIV-1 and other pathogens [14, 26, 31, 32], and for general disease susceptibility in humans [33]. Finally, the complexity that epistasis provides in understanding mutation

effects must be accounted for in protein engineering and design [34, 35, 36, 37].

For pairs of simultaneous mutations in proteins (we will refer to these as "double mutations"), epistasis can be expressed in terms of free energy differences:

$$\epsilon = \Delta\Delta G_{1,2} - (\Delta\Delta G_1 + \Delta\Delta G_2) \tag{2.1}$$

Where $\Delta\Delta G_{1,2}$ corresponds to the change in the folding or binding free energy due to the double mutation, and $\Delta\Delta G_1 + \Delta\Delta G_2$ refers to the sum of the constituent single mutation free energy changes. This nonadditivity can be caused by direct interactions between mutational sites, or by indirect effects such as conformational perturbations. Epistasis is positive when the double mutant is more stabilizing than the sum of the constituent singles ($\epsilon < 0$) and negative when the double mutant is more destabilizing than the sum of the constituent singles ($\epsilon > 0$).

Despite its importance to understanding biological systems, a comprehensive mechanistic picture of the drivers of epistasis in proteins is not known. An early attempt to explain epistasis mechanisms is a study by Wells [38]; they concluded that features like separation distance, electrostatic interactions, and conformational perturbations were likely contributors. However, this conclusion was based on a small data set containing a total of 12 folding and binding systems, with less than 75 total multiple mutations. More recent studies have examined specific protein systems like TEM-1 β-lactamase [39, 40] and the IgG-binding domain of protein G [41], finding pervasive negative epistasis. Long-range epistasis has also received attention. Gromiha et al. proposed that distant residues that are part of a specific local group (they defined this as a rigid cluster) could lead to epistasis [42]. Other researchers have used tools like molecular dynamics to analyze if networks of interactions can mediate long-range epistasis [43]. Classification systems have also been developed. Jemimah et al. used structural features to build a model to classify whether mutational pairs would be additive (i.e., not epistatic) [44]. These previous studies provide a basis for understanding possible contributors to epistasis and some even offer predictive capability, however they do not provide a complete understanding of epistasis mechanisms and their interactions.

In this study, we determine biophysical drivers of pairwise epistasis in protein systems and rank their contribution to the observed epistasis, ϵ (Equation 2.1). We used protein structural data, protein-protein binding affinities, and protein folding stabilities from the largest, most diverse datasets currently available. We explored possible relationships between the observed epistasis and features that are intrinsic to both the proteins and the mutated residues. A statistical model selection procedure was performed to determine the features that are most important to explaining the observed epistasis. The models

determined for binding and folding have similar and modest predictive power. Both models contain similar features that include separation distance and charge interactions. Our work serves as a stepping stone to further our understanding of the biophysical drivers of epistasis, and to build future models with more complex features and interactions.

2.2 Methods

2.2.1 Curating experimental data

Experimental binding affinity data was obtained from SKEMPI v2.0 [45] and folding stability data from ProTherm 4 [46]. Since the focus of our study is pairwise epistasis, we extracted a subset of the data consisting of all instances where there was data for a double mutant and the corresponding constituent singles. For both folding and binding data, values were converted to kcal/mol. A temperature of 298 K was used if not specified in the dataset. Averages were calculated for mutations that included multiple free energy values. The attributes in the resulting curated folding and binding datasets used in our study include the PDB ID, protein complex name, the mutation(s), and either binding or folding free energy values. The total number of data points for double mutants with constituent single mutants were 572 from 58 protein-protein complexes for binding, and 204 from 30 protein systems for folding. Epistasis was calculated for each double mutation data point using Equation 2.1, that is, by taking the difference between the free energy change due to the double mutation and the sum of the free energy changes due to the constituent single mutations. Protein structures used for analysis were acquired from the RCSB Protein Data Bank (PDB) [47].

2.2.2 Extracting Features as Possible Drivers of Epistasis

For electrostatics, and other categorical features described below, the explicit wildtype-mutant pairs are henceforth denoted separated by a semicolon for simplicity: wt_1wt_2;mut_1mut_2.

2.2.2.1 Amino Acid Properties

To investigate the effect of electrostatics on epistasis, we classified amino acids as positively charged (+), negatively charged (-), or neutral (0). To incorporate every wildtype-mutant pair state would be infeasible due to overparameterization, as it would result in 3^4=81 possible categories (++;--, ++;-+, ++;+-, +-;--, ...). To avoid overparameterization, we explored various abstractions of this data, incorporating this into our model selection process (detailed below). The resulting charge contribution was given by a simplified charge-interaction scheme with pairs belonging to one of three categories: attractive (+- or -+, denoted "A"), repulsive (-- or ++, denoted "R"), and neutral (all other cases,

denoted "0"). The reverse of each wildtype-mutant states were classified as the same (e.g. 0;A = A;0), resulting in four categories: 0A, 0R, AR, and 00 to capture all possible electrostatic interactions. Note that the AR case was not present in either dataset.

To include the change in size for the constituent amino acids we used the van der Waals volume in Å. To capture the net effect due to the change in size for both sites we used the metric (referred to as size_{net}.

$$\text{size}_{\text{net}} = |\ \text{size}_{\text{m1}} - \text{size}_{\text{wt1}}| + |\ \text{size}_{\text{m2}} - \text{size}_{\text{wt2}}| \tag{2.2}$$

where wt and m correspond to the wildtype and mutant amino acids, respectively, and 1 and 2 denote the amino acid sites in an arbitrary order. Under this scheme, if one or both sites undergo a large/small change in volume occupancy the corresponding metric will be large/small respectively, even if they are in opposing directions.

To include the effect of hydrophobicity, each residue is classified as either "H" for hydrophobic or "P" for polar. Using all possible 16 categories would be possible, but risk overfitting. We instead found the following abstraction: a boolean value ("0" or "1") that denotes whether the net hydrophobicity of the pair changed upon mutation. For example, HP;PH would give 0 since the net hydrophobicity remained the same. By contrast, PP;HP or PP;HH would both give 1, since the net hydrophobic state changed upon mutation.

2.2.2.2 Structural Properties

Separation distance was defined as the Cartesian separation between the alpha carbons for each mutational site. This Euclidean distance r was calculated using the x, y, z coordinates for the mutation sites via the standard formula:

$$r = \sqrt{(x_{wt1} - x_{wt2})^2 + (y_{wt1} - y_{wt2})^2 + (z_{wt1} - z_{wt2})^2} \tag{2.3}$$

Secondary structure information was included by considering whether a given mutational site was located in an alpha helix ("H"), beta sheet ("S"), or loop ("L"). Secondary structure content was determined using a PyMol script [48]. As with other categorical features overparameterization may be a concern, though in this case the explicit consideration only has nine possible cases. We tested the possible abstractions, ranging from explicit consideration of the structures at each site (e.g., HL,LL,LS,...) to the simplest case of a boolean value denoting whether both sites belong to the same type of structure ("0") or different structures ("1").

We also considered the effect of solvent accessible surface area (SASA): a metric describing whether a residue is exposed or buried. To calculate the SASA, we first prepared the PDB files using *pdbfixer* from the OpenMM software suite [49], to add missing residues, replace non-standard residues with their standard equivalents, and add missing hydrogens. The repaired structures were then processed with FoldX [50] to generate mutations using the *BuildModel* command. DSSP v3.0.0 [51] was then used to calculate the absolute SASA (SASA_{abs}) for each residue of interest. Both absolute and relative SASA were considered, relative SASA (SASA_{rel}) was calculated using the empirical max accessible surface area (ASA_{max}) generated by Tien et al [52] via the formula:

$$\text{SASA}_{\text{rel}} = \text{SASA}_{\text{abs}} \ / \ \text{ASA}_{\text{max}} \tag{2.4}$$

Since SASA changes affect both wildtype and mutant residues, we used a modified version of 2.2 replacing size_{net} with SASA.

We also included classification information. For binding, we included the type of protein-protein complex broken into five categories, based on the information provided in the SKEMPI v2.0 database: antibody-antigen (AB/AG), T cell receptor-peptide bound major histocompatibility complex (TCR/pMHC), Cytokine-Cytokine receptor (Cyto/Cyto), GTPase-other, and non-specific protein-protein interaction (Pr/PI) which functioned as the reference category for the statistical models. We also included a boolean value indicating whether or not the mutational sites occur on the same ("0") or different ("1") protein chains, as sites which occur on the same chain may have a different effect on binding than if they occur on opposing chains. For folding, we included the system size given by the total number of residues acquired from the PDB.

2.2.3 Statistical Analysis

To analyze the relationship between epistatic effect and separation distance, we conducted a likelihood ratio test that compares a null model (where separation distance is unrelated to epistasis) against an alternative model (where epistasis decays with increasing separation). More precisely, we defined the null model to be that epistasis values are sampled from a normal distribution that is independent of the separation between residues. For the alternative model, epistasis values are sampled from a normal (same mean as the null case) with a standard deviation that decays exponentially as a function of separation according to $ae^{\alpha \cdot r}$ where r is the separation between residue site alpha-carbons (Equation 2.3) and a and α are the curve's parameters estimated by maximum likelihood for the dataset. This maximum likelihood was determined by a grid-search method, considering all possible a and α, taking the resulting model

with the largest likelihood. The likelihood ratio is given by the ratio of the log of the two likelihoods of the data under the two models:

$$\Lambda(\epsilon) = \frac{\mathcal{L}(\theta_0|\epsilon)}{\mathcal{L}(\theta_1|\epsilon)} \rightarrow log(\Lambda(\epsilon)) = log(\mathcal{L}(\theta_0|\epsilon)) - log(\mathcal{L}(\theta_1|\epsilon)) \tag{2.5}$$

where $\mathcal{L}$ refers to the likelihood, log is the natural logrithm, and $\theta_0\theta_1$ correspond to the null and alternative models respectively. Small values of Λ indicate that the alternative model has more explanatory power than the null. We first calculated the likelihood ratio for the experimental data, Λ_{exp}. In order to determine statistical significance of Λ_{exp} we then obtained the distribution of Λ under the null through parametric simulation. Specifically, we simulated datasets using the mean and standard deviation of the experimental epistasis data. We then repeated the fitting exercise used on the real dataset for the simulated dataset, using the same separation data, and calculated Λ. This process was repeated 1000 times to obtain the distribution of Λ under the null: Λ_{sim}. The p-value for the test was then calculated as the proportion of Λ_{sim} less than or equal to Λ_{exp}.

Linear statistical models were used to determine the biophysical features that are best able to explain the observed epistasis. The absolute value of the epistasis, ϵ, was used as a response variable for our model building. The choice to use the absolute value was necessary to ensure a monotonic relationship between the features and the response variable, as assumed when using linear models. One could imagine analyzing positive and negative epistasis separately; however, this was not possible due to small sample sizes. All features described above were considered in a standard model selection procedure, including all pairwise interactions terms. For any features where we considered more than one level of abstraction, only one level was included in any given model. To evaluate model performance, the corrected Akaike information criterion (AICc) was used. The corrected criterion was chosen over the standard AIC due to the potential for overfitting models that contain a large number of terms given a small amount of data [53]. Models were generated and tested using R software [54] by considering all permutations of abstracted and non-abstracted features. Model selection was performed using a modified form of stepAIC from the MASS [55] package to perform forward and backward selection based on AICc (further verified by the AICc function of AICcmodavg [56] and compared to standard AIC). Forward selection explores model space by starting with a term-less model and systematically adding terms to find the model with the best value for a given criterion. Conversely, backward selection starts with the complete full-term model and removes terms to find the best model. This model selection process was performed twice with randomized input terms to avoid potential ordering bias (terms treated differently based on their position in the initial list) and the lowest AICc values were compared for consistency. Once we verified that there

was no ordering bias, the model with the lowest AICc for both binding and folding was used for further analysis.

To rank the importance of features present in the final statistical models for their effect on epistasis we compared R^2 values with and without each feature and its interactions. Features with larger explanatory power of the observed epistasis will have a larger change in R^2 when removed.

2.2.4 Quantification of Experimental Error and Model Validation

In order to develop a model for epistasis, it is important to quantify how much of the observed epistasis could be attributed to error, or noise, in the experimental data. Quantification of overall error is based on the error in three values ($\Delta\Delta G_{1,2}$, $\Delta\Delta G_1$, $\Delta\Delta G_2$), each of which were determined using a broad range of techniques and conditions from diverse studies (e.g., 60+ for binding). A survey of six studies that contained some of the largest observed epistasis for binding showed the experimental standard error for $\Delta\Delta G$ to be in the range 0.05 - 0.3 kcal/mol [57, 58, 59]. However, some studies do explicitly include the error for epistasis (frequently termed the coupling energy). For example, in the case of barnase-barstar, Schreiber et al., reports errors in ϵ from 0.2 - 0.39 kcal/mol across 33 mutation pairs [60] and Goldman et al. reports an error of 0.3 kcal/mol across 13 pairs for an Idiotype-AntiIdiotype Protein-Protein complex [61]. There are outliers, such as the study from Pielak, et al. with six mutational pairs in the Iso-1-cytochrome C Peroxidase complex [62] found to have an error range of 0.4 - 1.0 kcal/mol with an average error of 0.75 kcal/mol for six samples; an unusually large error. In summary, the reported error for our curated binding and folding datasets are in the range of 0.2 - 1.0 kcal/mol, with mean around 0.4 kcal/mol. For the remainder of this study, we will use a slightly more conservative estimated error of 0.5 kcal/mol to quantify the amount of observed epistasis.

Since our binding and folding data comes from many different protein systems collected by a diversity of methodologies and laboratories, there is an inherent imbalance in the quantity and quality of data for each system. To test the robustness of our model to this bias, we applied a modified "leave-one-out" procedure. We randomly removed 10% of the protein systems and their data, creating a subset from the remaining 90% of systems. The model selection procedure was performed on this subset to generate a new model. This process of removing 10% of the systems and running model selection was repeated 100 times. The resulting 100 subset models were analyzed and compared to determine which terms appeared, their frequency of appearance, and average performance or ranking when present in a model.

2.3 Results

To build a statistical model for epistasis in proteins we used data for binding curated from SKEMPI v2.0 (572 mutation pairs), and for folding curated from ProTherm 4 (204 mutation pairs). We first considered the extent to which epistasis was present in our data set. To determine this, we defined an epistasis cutoff; values where $|\epsilon|$ is larger than the cutoff are considered epistatic, and other values are not. Ideally, the cutoff would be chosen based on the experimental error or uncertainty, however, given that our data come from a broad spectrum of methods and sources, this is not possible to determine for the dataset as a whole.

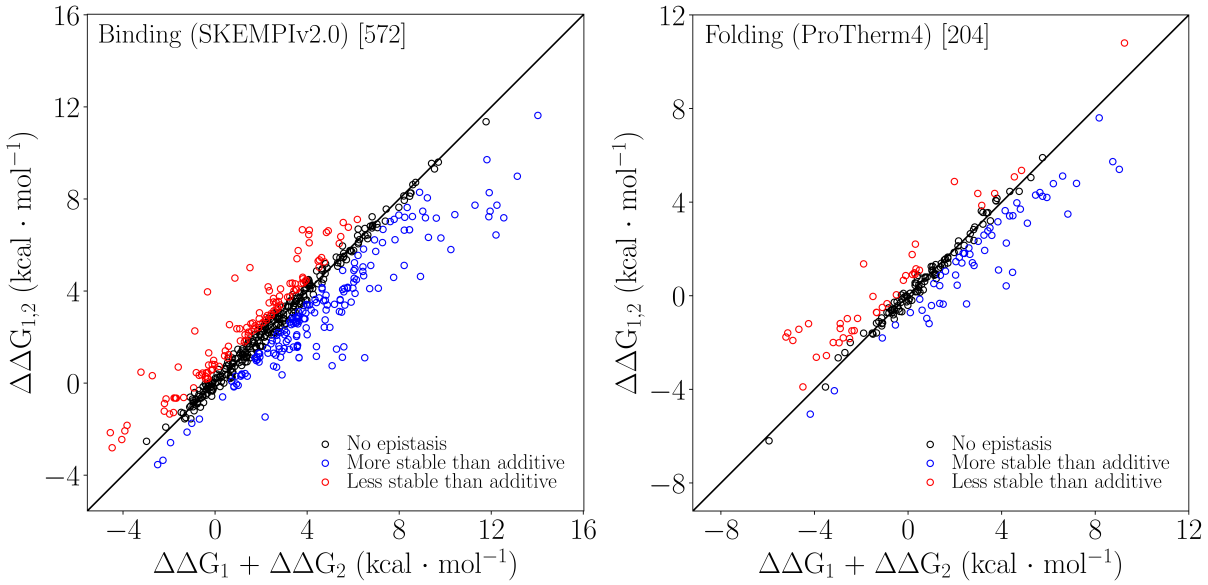

Figure 2.1: Epistasis scatterplots for binding (left) and folding (right). Both figures use a cutoff of 0.5 kcal/mol and show data characterized as no epistasis (black), positive epistasis (blue), and negative epistasis (red).

Figure 2.1 shows the free energy change of the double mutant as a function of the sum of individual free energies for both binding and folding datasets with a cutoff of 0.5 kcal/mol. In both datasets there is a marked trend for large sums of constituent single mutations (sum in Equation 2.1) to correspond to a double mutant with free energy falling below the 1:1 line (i.e., more stabilizing that predicted by additivity). The opposite is true for constituent mutations with smaller sums.

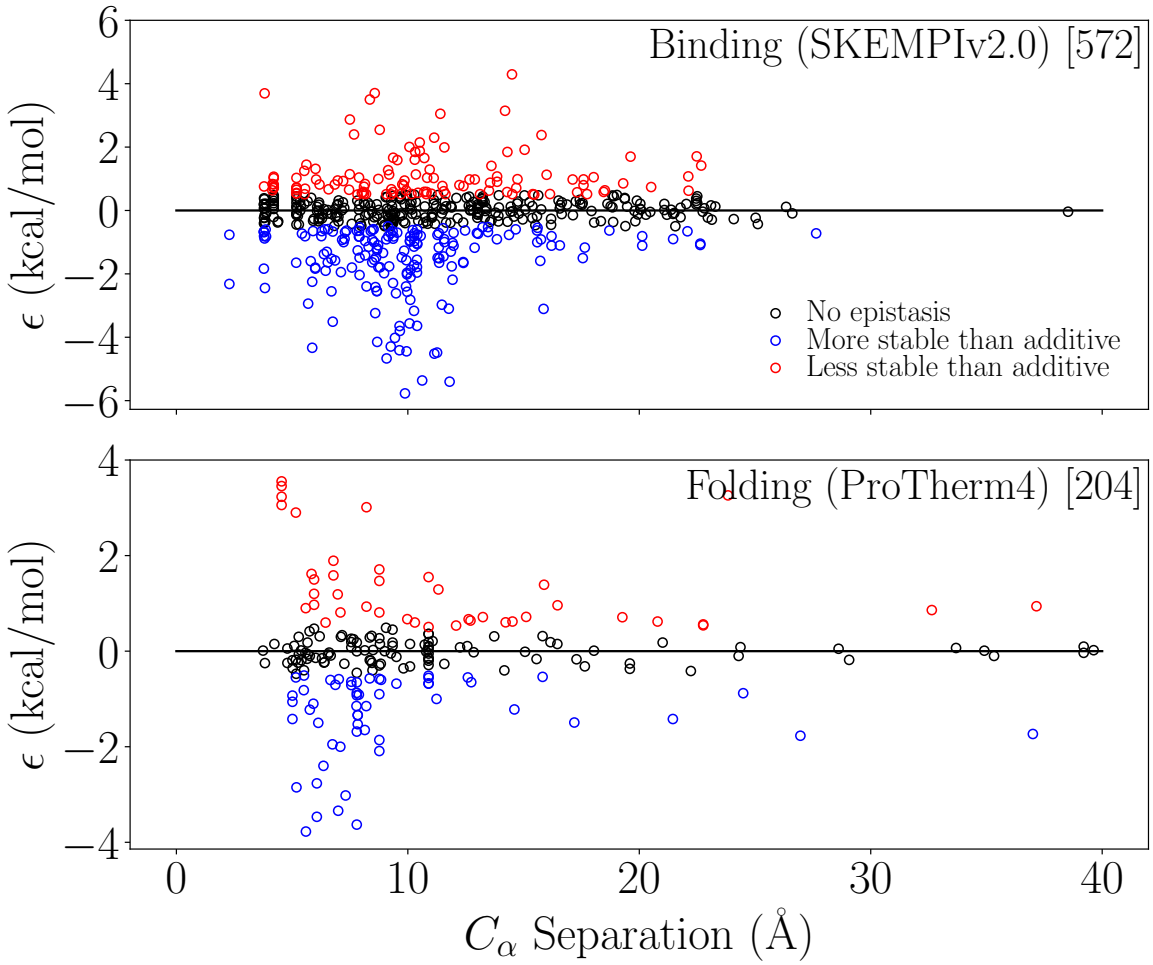

Figure 2.2: Observed epistasis as a function of alpha-carbon separation between mutation sites for binding (top) and folding (bottom). Black indicates no-epistasis using our cutoff of 0.5 kcal/mol, and blue and red indicate positive and negative epistasis, respectively.

After ascertaining the extent to which epistasis is present in our data, we investigated how well the separation between mutation sites could explain the epistatic effect. Figure 2.2 shows the relationship between separation distance and the observed epistasis for binding (top) and folding (bottom). Both show the general expected trend of less epistasis as separation increases. Both also show a larger number of data points for distances with the largest ϵ values, or spread in ϵ (around 6-10 Å).

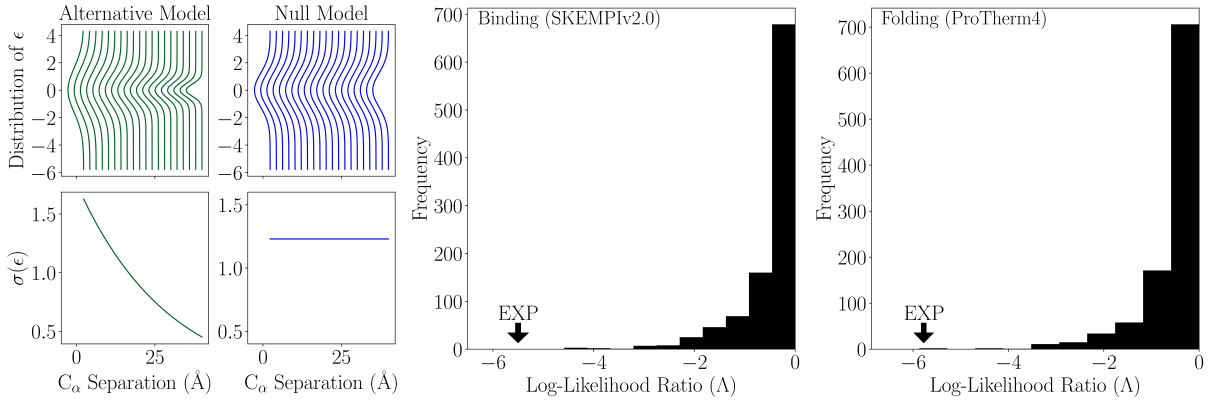

Figure 2.3: Comparison between the alternative and null models for epistatic effect, ϵ, as a function of separation distance, r (left). Results of log(likelihood) ratio test for separation distance with 1000 samples for simulated data for binding affinity (center) and folding stability (right). These plots show the alternative model is a significantly better explanation of the data than the null model.

Figure 2.3 shows our analysis to determine whether the apparent decrease in epistasis with increasing separation distance (Figure 2.2) is due to an actual relationship or a consequence of the larger number of data points at small distances. Figure 2.3 left shows null model ($\sigma(\epsilon)$ is not a function of r) and alternative model ($\sigma(\epsilon)$ exponentially decreases as a function of r) for the likelihood ratio analysis. Figure 2.3 center and right shows the simulated distribution of the likelihood ratio, Λ,from the analysis with 1000 samples for binding and folding respectively. The experimentally observed likelihood ratio is well outside the distribution of null ratios given by the label "EXP" and has a value of -5.50 compared with the tail of the simulation distribution minimum of -4.59. In simple terms, this results in a p-value of $p < 1/1000$ ($p<0.001$) in strong support of the alternative model.

Binding Model (AICc: 1418.65)				
Feature	**Categorical Breakdown**	**Removal (ΔR^2)**	**Coefficient**	**P-value**
Full Model	572	0.2991		
Intercept			0.6994	0.0000
Complex Type		0.1275		
AB/AG	66 (8)		-0.8401	0.0120
Cyto/Cyto	69 (4)		0.4677	0.3738
GTPase/other	85 (7)		-0.4134	0.3153
TCR/pMHC	58 (9)		-1.0090	0.0071
Pr/Pl	294 (30)		0.0000	0.0000
Charge		0.0778		
0	450		0.0000	0.0000
0A	69		1.0195	0.0041
0R	53		-0.7424	0.0131
Separation		0.0522	-0.0025	0.8074
Interaction Side		0.0464	0.5889	0.0055
1	399			
0	173			
$Size_{net}$		0.0427	0.0022	0.1345
$SASA_{abs}$		0.0327	-0.0060	0.0002
Secondary Structure		0.0162	0.5837	0.0014
1	315			
0	257			
Binding Model Interaction Terms				
Feature1:Feature2			**Coefficient**	**P-value**
Separation:Secondary Structure			-0.0331	0.0246
$Size_{net}$:Interaction Side			-0.0052	0.0081
Separation:Charge				
0A			-0.0882	0.0001
0R			0.0454	0.0401
$SASA_{abs}$:Complex Type				
AB/AG			-0.0140	0.0207
Cyto/Cyto			0.0072	0.2340
GTPase/other			0.0072	0.0789
TCR/pMHC			0.0062	0.1631
$Size_{net}$:Complex Type				
AB/AG			0.0130	0.0000
Cyto/Cyto			0.0023	0.4650
GTPase/other			0.0008	0.7843
TCR/pMHC			0.0060	0.0754
Interaction Side:Charge				
0A			0.7289	0.0089
0R			0.1460	0.5625
Interaction Side: Complex Type				
AB/AG			0.1376	0.6577
Cyto/Cyto			-1.4297	0.0004
GTPase/other			-0.3189	0.3331
TCR/pMHC			0.0062	0.9824

Table 2.1: Summary of epistasis model for binding. The leftmost column (column one) contains features. Any categorical abstractions are listed directly below the category with right justification. Column two gives the specific number of mutation pairs for a given category, where applicable. For complex type specifically, the number of complexes of that type are indicated in parentheses. Column three is the change in R^2 (ΔR^2), i.e., how much poorer the model fits the data after removing this feature. In the case of the full model, column three is the R^2. Removal of a feature also removes all subcategories and any interaction terms involving the feature. Column four lists coefficients for the feature/interaction term in the full model. The rightmost column contains p-values for the features,and features within a given category.

Folding Model (AICc: 481.60)				
Feature	**Categorical Breakdown**	**Removal (ΔR^2)**	**Coefficient**	**P-value**
Full Model	204	0.2578		
Intercept			-0.0607	0.8513
HP		0.1506	0.0746	0.0018
0	133			
1	71			
$Size_{net}$		0.0765	0.0120	0.0000
Charge		0.0695		
0	174		0.0000	0.0000
0A	13		1.8631	0.0015
0R	17		-0.7356	0.0133
Separation		0.0446	-0.0416	0.0054
$SASA_{rel}$		0.0383	0.3269	0.8546
Folding Model Interaction Terms				
Feature1:Feature2			**Coefficient**	**P-value**
$SASA_{rel}$:$Size_{net}$			-0.0437	0.0174
$SASA_{rel}$:Separation			0.1577	0.1324
$Size_{net}$:HP			-0.0103	0.0085
$SASA_{rel}$:HP			2.7022	0.0459
HP:Charge				
0A			-2.0467	0.0016
0R			0.8380	0.0391

Table 2.2: Summary of epistasis model for folding. The leftmost column (column one) contains features. Any categorical abstractions are listed directly below the category with right justification. Column two gives the specific number of mutation pairs for a given category, where applicable. Column three is the change in R^2 (ΔR^2), i.e., how much poorer the model fits the data after removing this feature. In the case of the full model, column three is the R^2. Removal of a feature also removes all subcategories and any interaction terms involving the feature. Column four lists coefficients for the feature/interaction term in the full model. The rightmost column contains p-values for the features,and features within a given category.

Tables 2.1 and 2.2 show a summary of the binding and folding statistical models for epistasis in protein systems, respectively. The final model for binding had an AICc value of 1418.65, with the other models considered having AICc values ranging from 1420.86 to 1493.66. The final model for folding had an AICc value of 481.60 with the other models considered having AICc values ranging from 482.20 to 507.90. Note that all other models consist of the remaining permutations of all features considered. Both models have similar predictive power in the range of 25-30%. The final selected binding model contains all features that we considered except for hydrophobicity (seven features, 28 terms including interactions) and depends on $SASA_{abs}$ and secondary structure in addition to binding specific features like the complex type. The

folding model is simpler (five features, 12 terms with interactions), and depends on hydrophobicity and $SASA_{rel}$. Features are listed in order according to their relative contribution to the explanatory power of the full model. That is, the highest-ranked feature is the one whose removal leads to the greatest reduction in R^2. For the binding epistasis model, the largest contributor was the complex type, with a change in R^2 of 0.128 upon removal followed by charge with a change in R^2 of 0.078 upon its removal. The remaining terms each have a ΔR^2 of $\sim$0.05 or less. For the folding epistasis model, the largest contributor was hydrophobicity with a change in R^2 of 0.151 upon removal, followed by both size and charge with similar contributions (change in R^2 of 0.0765 and 0.0695 with their removal respectively). The remaining terms each have a ΔR^2 of $\sim$0.045 or less.

Binding Validation					
Feature	**Mean Rank**	**Average ΔR^2**	**Number of Models (/100)**	**In Full Model**	**Full Model Rank**
Complex Type	1.04	0.134	100	Yes	1
Charge	2.01	0.082	100	Yes	2
Separation	3.66	0.0560	100	Yes	3
$Size_{net}$	4.45	0.047	100	Yes	5
Interaction side	4.66	0.046	100	Yes	4
SASA	5.64	0.0380	100	Yes	6
Secondary Structure	6.68	0.022	97	Yes	7
Hydrophobicity	7.469	0.018	32	No	N/A
Folding Validation					
Feature	**Mean Rank**	**Average ΔR^2**	**Number of Models (/100)**	**In Full Model**	**Full Model Rank**
Hydrophobicity	1.212	0.15	99	Yes	1
Charge	3.083	0.082	96	Yes	3
Secondary Structure	3.213	0.0810	47	Yes	N/A
$Size_{net}$	3.22	0.074	100	Yes	2
Separation	3.98	0.058	100	Yes	4
SASA	4.897	0.041	97	Yes	5
Number Residues	5.269	0.0380	26	No	N/A

Table 2.3: Results from 100 trials of our "leave-10%-out" model robustness test for binding (top) and folding (bottom). The feature is indicated by the first column. The second column indicates the average rank across all trials the given feature appeared in, lower numbers suggest more robust features. The third column indicates the average ΔR^2 from all trials the feature appeared in (higher numbers suggest more robustness), the fourth column indicates the total number of trials a given feature occurred in out of 100 possible, the fifth column indicates whether the feature was present in the full model, and the last column indicates the rank of the feature in the full model

Table 2.3 shows the results of 100 trials of our "leave-10%-out" robustness test where 10% of the available systems were randomly removed. These results show that both of our full models are highly robust -- with the binding model being slightly more robust than the folding model. All terms present in the full models are present in the "leave-10%-out" analysis, most occurring in all trials. Additionally, the mean ranks of most terms are identical to the full-data binding model with more variance in the folding model.

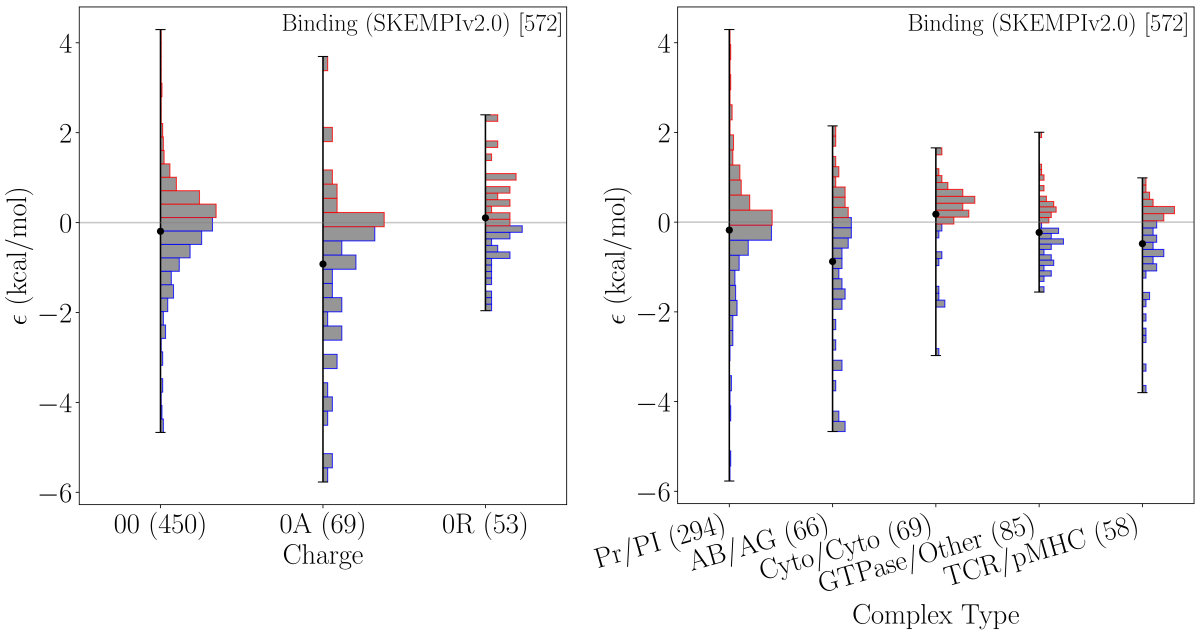

Figure 2.4: Comparison of binding model of epistasis for the categories of charge (left) and complex type (right). The mean value for a given subcategory is indicated by a black dot. The barplots show the histograms within the categories. In parenthesis is the number of mutation pairs belonging to each category. For the complex type, the number of complexes belonging to each category are shown in square brackets.

Figure 2.4 further illustrates the results of our statistical model for epistasis in binding. For charge, the subcategory for interactions involving an attractive pairing (0A) contains the most strongly epistatic mutations. While mutations in this subcategory cover a broad range of values, many tend towards positive epistasis; the largest value belongs to this subcategory. Neutral or constant charge states (00) show a near normal distribution centered on zero with some low levels of epistasis. Changes involving a repulsive interaction (0R) contain the least number of data and have a narrow distribution, with fewer large values for ϵ in either direction. For the complex type category, the antibody-antigen subcategory shows the most epistasis, including the most positive. TCR/pMHC also contains a large amount of positive epistasis. Cytokine-cytokine is the only subcategory with a negative mean suggesting that mutations in this subcategory tend to have negative epistasis. Generic protein-protein complexes show similar behavior to the neutral charge category; centered on zero, broad spread, but low numbers of epistatic data points.

2.4 Discussion

Before building linear models we first determined the extent to which our datasets contain meaningful epistasis. That is, considering there is uncertainty in the data, where should we draw the line between

epistatic and non-epistatic values of ϵ? We estimated (see Methods) that the error for both datasets fall between 0.2 - 1.0 kcal/mol with an average around 0.5 - 0.6 kcal/mol. From this, we estimated a cutoff of 0.5 kcal/mol, i.e., $|\epsilon| > 0.5$ kcal/mol are considered epistatic. There are further limitations of our dataset; the data is not from randomized studies. Instead, the experiments were generally conducted in a targeted fashion with a priori knowledge of function. This may explain why we find more positive epistasis (more stabilizing that additivity predicts) than negative epistasis (more destabilizing that additivity predicts) as shown in Figure 2.1. Alternatively, it is possible that more positive epistasis is present in the data because negative epistasis could lead to protein misfolding or non-binding events in the experiments. The former reasoning is an artifact of how the data was generated, and the latter is related to biophysical features of the proteins; both carry different implications for the dataset and warrant future work.

Separation distance is the most intuitive feature expected to contribute to epistasis, because residues that are near each other are more likely to interact than those far apart. Simple visual comparisons show a decreasing spread of epistasis with increasing distance (Figure 2.2). The folding data show this most strongly with a sharp peak around the shortest separation distances of approximately 6 Å, dropping to near zero at larger distances. The binding data show a possible peak around 10 Å, however, the trend is not as clear. Additionally, with binding there is a paucity of data from 25 Å to 40 Å with only one data point around 40 Å. Our tests using likelihood ratio methods (Figure 2.3) confirm that separation does play a role in epistasis for both binding and folding. Our alternative model (width of possible ϵ values depends on separation) was a better explanation than the null model (no relationship between separation and ϵ), with a p-value of $p < 0.001$ in the case of binding, and $p < 0.002$ in the case of folding. This indicates there is an inverse relationship between separation distance and epistasis, but does not quantify its significance or magnitude. This effect due to separation is confirmed and quantified in our models (Table 2.1) where both folding and binding models have negative coefficients for separation distance. In the folding model, a 10 Å increase in separation between residues results in a decrease in epistasis of 0.416. In the binding model, the effect of separation alone is an order of magnitude less than the folding model and has less significance in the model (p=0.8074). Instead, the effect of separation in the binding model is most strongly characterized by the interaction with charge. With charge alone, changes involving attractive pairings show an increase in epistatic effect whereas changes involving a repulsive pairing show a decrease. The interaction between charge and separation contributes an opposing effect: as separation between residues increases, changes involving attractive and repulsive pairings cause a decrease and increase in epistatic effect respectively. Intuitively, as separation between charged residues, regardless of categorization, increases the net effect of charge on epistasis tends towards zero ($\Delta\epsilon_{\text{charge}} + \Delta\epsilon_{\text{charge:separation}} \sim 0$).

In addition to separation distance, amino acid size is present in both models. Size is another feature one might intuitively expect to contribute to epistasis: large absolute changes in size imply that voids are created when residues change from larger to smaller, or that smaller to larger residues create steric clashes. In both models the coefficient is positive (increases in ϵ occur with change in size) with more of an effect in the case of folding (on the order of 10^{-2} vs 10^{-3} in the case of binding). Size interaction terms differ between binding and folding. In the case of binding, when there are changes in size that occur on different protein chains, there is a reduction in epistasis. Otherwise, for all complex types, changes in size lead to an increase in epistasis, most strongly with Antibody-antigen complexes. For folding, size_{net} interacts with hydrophobicity and SASA_{rel} leading to decreases in epistasis. This will be discussed further with the features specific to the folding model.

In the case of both binding and folding, there are unique features that contribute significantly to the observed epistasis. In the case of binding, these elements only apply to binding interactions such as the type of complex (defined by function) and whether both mutations occur on the same side of the binding interaction. Complex type is the most significant contributor to the observed epistasis ($\Delta R^2 = 0.17$) with most complexes showing less epistatic effect compared to the reference category of generic protein-protein complexes. There is an exception with Cytokine-cytokine complexes that shows a small increase in epistasis with a coefficient of +0.4677. The interaction side is a smaller contributor compared to complex type ($\Delta R^2 = 0.0465$), with a slight increase in epistatic effect when mutations occur on opposite sides of the binding interaction. This is consistent with intuition; if both mutations are near the binding interface and on opposite sides, they are more likely to directly interact, or propagate effects at the interface. Additional features that contribute to epistasis in binding are secondary structure and SASA_{abs}. Secondary structure has a minor contribution, with a slight increase in epistatic effect when residues belong to different secondary structure types. This is counterbalanced by an interaction with separation distance, where residues that occur in different secondary structures, and are also far apart, lead to a decrease in epistatic effect. This could be due to direct interactions between sites; if they are close together but belong to different secondary structures, they can change these structures either directly or indirectly. This is less likely to happen if they are further apart. SASA_{abs} is the penultimate feature in the model ranking with a very small coefficient (-0.006). This implies that changes in the total exposed surface area due to the two mutations lead to small reductions in the epistatic effect.

Unique to the folding model, hydrophobicity is present, and is the strongest contributor to epistasis with a ΔR^2 of 0.1506. Changes in the net hydrophobicity lead to an increase in the observed epistasis. This is consistent with other studies that have shown that hydrophobicity contributes to predicting folding stabilities with double mutations [63]. Most of the other terms present in the folding model interact with

hydrophobicity leading to a stronger effect on epistasis, and a reduction when paired with changes in size, and changes in charge involving attractive interactions.

Since our statistical models for both binding and folding explain approximately 25-30% of the observed epistasis, an important question is: what explains the other 70-75%? We believe the answer lies in dynamical properties that are beyond the scope of what we investigated here. Protein complexes are not static objects, thus static features like those considered in this study are only likely to capture some of the true physical effect they can have on these systems. While a tool like molecular dynamics could potentially help address this question, given the number of mutations and systems considered here, the computational cost would be unreasonably large and will be left as a topic for future study.

Given the size of our datasets, and the imbalanced nature of the data in terms of protein systems, we performed a "leave-10%-out" validation procedure to test the robustness of our models and determine whether there are system-specific effects (see Table 2.2). We found that our binding model was very robust; all terms appearing in the full model were also present in the validation trials effectively 100% of the time (the least significant term, secondary structure, was missing from three trials). The mean rank was also consistent between the validation trials and the full model ranking for the three most significant terms, the 4th and 5th are switched but close enough to be within a margin of error, the 6th and 7th were also consistently ranked. The folding model was slightly less robust. The effect of hydrophobicity was very robust being ranked first in the full model and appearing in 99 of the 100 validation trials with a mean rank of one. The remaining folding model terms appear between 96% to 100% of the time, however their mean rankings are generally inconsistent with their full model rank, indicating that while they are important to explaining epistasis we cannot be as certain of their relative contribution.

A limitation in the current study, that is also a limitation for all similar studies, is the lack of comprehensive, diverse, and unbiased datasets. Given the challenges associated with measuring binding or folding free energies for a large number of mutants, these datasets are built with narrow focus and small sample sizes. Such databases tend to be biased toward systems of particular interest. Additionally, they will not contain mutations that result in a nonviable protein or system. This does not make the data any less relevant since in nature proteins must be viable, and thus we should expect similar results (e.g., the preponderance of positive epistasis observed in this study). If we want to understand the nature of epistasis at the level of protein stability, we need to study it across more protein systems in a more systematic fashion. To build a truly predictive model of epistasis, dynamic properties would need to be considered and a larger, more representative sample of data would need to be accessible.

2.5 Conclusion

In this study we investigated possible mechanisms and determined statistical models for pairwise epistasis in proteins based on the largest, most diverse, experimental data available. Mechanistic features were investigated that are intrinsic to the mutating amino acids (e.g., charge, hydrophobicity) or to the proteins (e.g. secondary structure, distance between mutational sites). Using a model selection procedure we ranked these features by their power in explaining the observed epistasis. The resulting models for both binding and folding had similar explanatory power of 25-30% and were composed of similar high-ranked features. The features included in both models were charge, separation distance, and residue size. The largest contributing features were complex type for binding, and hydrophobicity for folding. Our results shed some light on the mechanisms for pairwise epistasis in proteins, and highlights the need for larger datasets. Our study also suggests that development of a truly predictive model for epistasis will likely require difficult to ascertain features such as conformational changes, bond formation, and other propagated mutational effects.

Chapter 3: Short-wavelength-sensitive 2 (Sws2) visual photopigment models combined with atomistic molecular simulations to predict spectral peaks of absorbance

Dharmeshkumar Patel[1], Jonathan E. Barnes[2], Wayne I. L. Davies[3-7], Deborah L. Stenkamp[8,9], Jagdish Suresh Patel[1,8*]

[1]Institute for Modeling Collaboration and Innovation (IMCI), University of Idaho, Moscow, ID, United States of America

[2]Department of Physics, University of Idaho, Moscow, ID, United States of America

[3]Umeå Centre for Molecular Medicine (UCMM), Umeå University, Umeå, Sweden

[4]School of Biological Sciences, University of Western Australia, Perth, WA 6009, Australia

[5]The Oceans Graduate School, University of Western Australia, Perth, WA 6009, Australia

[6]The Oceans Institute, University of Western Australia, Perth, WA 6009, Australia

[7]Lions Eye Institute, University of Western Australia, Perth, WA 6009, Australia

[8]Department of Biological Sciences, University of Idaho, Moscow, ID, United States of America

[9]Institute for Bioinformatics and Evolutionary Biology, University of Idaho, Moscow, ID, United States of America

Published in PLOS Computational Biology. As second author, I contributed the following:

- Shortlist of angles compared to experimental spectral peaks.
- Statistical model selection resulting in models given by equations: 3.2 - 3.5.
- Generated figures: 3.3C, 3.4, 3.5.
- Wrote relevant sections in methods, results, and discussion that pertained to the above.

Writing and editing was done with google docs and a word document 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 published by PLOS open access, with a creative commons license, and can be used freely in this dissertation. Final publication is available through PLOS: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008212 .

3.1 Introduction

For many animals, vision is a critical sensory modality that facilitates essential tasks that include navigation, predation and foraging, predator avoidance, and numerous social behaviors. In vertebrates, vision is initiated when photons enter the eye and interact with rod and cone photoreceptors found within the neural retina. Specifically, light is detected by photosensitive visual photopigments that are housed within the folded membrane of photoreceptor outer segments. Upon the absorbance of photons, these photopigments activate a specific phototransduction cascade that results in a change in membrane potential. Once hyperpolarized, visual photoreceptors cause a change in neurotransmitter release that propagates neural signals through other retinal neurons that ultimately lead the visual centers of the brain [64, 65, 66].

A vertebrate visual photopigment consists of a transmembrane opsin protein that is covalently linked to a vitamin A-derived chromophore. Indeed, it is the interaction of the chromophore with specific amino acids of the protein sequence of the opsin that results in a broad array of different spectral peaks of absorbance (i.e. the λ_{max} value) [64, 65, 66, 67, 68, 69]. In the visual system, the predominant chromophores are either based on 11-*cis* retinal (i.e. rhodopsins or vitamin-A_1 photopigments) or 11-*cis*-3,4-didehydro retinal (i.e. porphyropsins or vitamin-A_2 photopigments) [64, 65, 66, 70]. Throughout vertebrate evolution, the associated visual opsin genes and their gene products have diverged into five classes: these comprise a long-wavelength-sensitive (*LWS*) opsin gene class, two short-wavelength-sensitive (*SWS*) opsin gene classes (*SWS1* and *SWS2*), and two medium-wavelength-sensitive (*mws*) opsin gene classes called rhodopsin-like 1 (*RH1*) and rhodopsin-like 2 (*RH2*), respectively [64, 65, 66, 71]. In general, RH1 opsins form the highly sensitive visual photopigments (rod opsins) of rod photoreceptors, which typically mediate dim light or scotopic vision by maximally detecting wavelengths at around 500 nm [64, 65, 66], although λ_{max} values can shift to shorter wavelengths (e.g. 470-490 nm in some deep-sea fishes [72, 73, 74]. The remaining four opsin gene classes (*LWS*, *SWS1*, *SWS2* and *RH2*) form the visual photopigments expressed in cone photoreceptors that mediate bright light or photopic vision. Color vision is possible when at least two cone photopigments with distinct λ_{max} values and overlapping spectral absorbance profiles are present within distinct cone populations, thus providing differential input to other retinal neurons [75]. Cone opsins of the SWS1 class typically produce photopigments with λ_{max} values between 360 nm and 450 nm (i.e. perceived as ultraviolet (UV) to violet parts of the light spectrum), SWS2 photopigments present with λ_{max} values of ∼400-470 nm (perceived as blue), RH2 opsins produce photopigments with λ_{max} values of ∼480-530 nm (perceived as blue-green), whereas LWS photopigments have λ_{max} values that are largely sensitive to a range of wavelengths from 500-570 nm (perceived as red) [64, 65, 66, 71].

These spectral ranges are based on photopigments that possess a vitamin-A_1-derived chromophore (i.e. rhodopsins); whereas for porphyropsins, the presence of a retinal chromophore based on vitamin-A_2 shifts the λ_{max} value towards longer wavelengths (e.g. up to ~620 nm [64, 65, 66, 70, 76], a spectral property that is more pronounced the longer the wavelength [77].

There is a great deal of diversity within each class of vertebrate opsins, since visual photopigment proteins are under strong natural selection [78, 79, 80]. This diversity is particularly striking within the teleost fishes, where many of the opsin genes have been tandemly or otherwise replicated, followed by subfunctionalization or neofunctionalization to generate new photopigments with distinct λ_{max} values [73, 1, 81, 82, 83]. For example, many fish genomes harbor two or three copies of the *sws2* cone opsin gene [84], the subject of the present study. In some cases, only minor changes in the amino acid sequence have resulted in major changes in the λ_{max} value, as was demonstrated for the A269T (Amino acid numbering in this study is based upon bovine RH1 sequence numbering) change in the Sws2a opsin of the spotted flounder (*Verasper variegatus*) that resulted in a photopigment that is sensitive to longer wavelengths (i.e. λ_{max} = 485 nm instead of 466 nm) [2].

There is great interest in understanding the evolutionary, as well as the molecular mechanisms, that underlie the diversity of visual photopigments and their spectral peak absorbances. In general, there are two main experimental techniques for defining the λ_{max} value of visual photopigments, namely microspectrophotometry, which analyzes the spectral profile of photoreceptors directly, but only with fresh retinas, and spectral tuning site substitutions cannot be studied in isolation [85] or *in vitro* regeneration [65, 2, 86, 87, 88]. The latter is a popular, yet highly specialized and labor-intensive, approach that has been used to deduce the spectral properties of photopigments in isolation from wildtype and inferred ancestral protein sequences, followed by reconstitution with chromophore (usually 11-*cis* retinal) and experimental measurement of absorbance from 200-800 nm [86, 89, 90, 91, 92]. This technique, when combined with site-directed mutagenesis, has illuminated the contributions of specific amino acid substitutions to shifts in the λ_{max} value [65, 86, 87, 91]. Although success of such studies has permitted the accurate prediction of the spectral peak of absorbance for some opsin classes (specifically, LWS [93] and UV-sensitive SWS1 [87] photopigments), such efforts are far from sufficient to understand pigment function that allows prediction of the λ_{max} value based entirely upon the amino acid sequence [69, 94]; this is particularly the case for both SWS2 and RH2 photopigments, where experimental interventions are frequently employed. One of the long-term goals of our studies is to develop computational tools that result in straightforward, genome-to-phenome, predictive pipelines, as is the case for the application of spectral modeling and atomistic molecular simulations for both the Rh1 and Rh2 classes of visual photopigments [73, 95].

The λ_{max} value of any functional photopigment in its inactive form is determined by the conformation adopted by the chromophore in the dark state, a function that is dependent upon the shape and composition of the retinal binding pocket, as well as the counterions that stabilize the Schiff base linkage of the chromophore to lysine (K) 296 of the opsin protein [96, 97, 98]. Therefore, our aim of generating genome-to-phenome pipelines for predicting the λ_{max} values of visual photopigments from their amino acid sequences, has been to increase an understanding of chromophore conformation via atomistic molecular simulations and to use this structural information to generate predictive models [73, 95]. The approach includes: 1) building homology models for classes of visual photopigments, using the solved crystal structure of bovine rod opsin (RH1) as a template [99]; 2) carrying out atomistic molecular dynamics (MD) simulations using homology models of photopigments with experimentally-measured λ_{max} values; 3) identifying structural features of the chromophore and opsin that are correlated with a particular λ_{max} value; and 4) using these features to generate a statistical model that can in turn be used to predict λ_{max} values of other photopigments. This approach was successful for predicting λ_{max} values of teleost Rh1 photopigments [73] and a closely-related Rh2 class of teleost cone photopigments [95]. Notably, this approach also revealed structural features of the Rh1 protein, namely the presence vs. the absence of a C111-C188 disulfide bridge that powerfully predicts λ_{max} values >475 nm when present vs. λ_{max} values <475 nm when absent [73].

In this present study, we test the hypothesis that the approach outlined above and in our previous publications [73, 95], can also successfully predict λ_{max} values of a class of cone photopigments that are phylogenetically-distinct from the known bovine RH1 template, namely the teleost Sws2 class. The SWS2 opsins are more divergent from RH1 than RH2 vs. RH1 [71, 83] and are known to be notoriously difficult when attempting to successfully predict λ_{max} values from the amino acid sequence alone [64, 65]. Furthermore, teleost Sws2 opsins display only ∼48-51% amino acid sequence identity to the bovine rod opsin sequence (Table 3.1). By contrast, our previous studies tested this approach for teleost Rh1 photopigments, with ∼49-83% identity to bovine rod opsin [73], and Rh2 cone pigments, with ∼63-72% identity to bovine rod opsin [95]. Here we also test the hypothesis that this approach will work for a class of photopigments with a broader, and more short-wavelength-shifted range of λ_{max} values (i.e. 397-485 nm; Table 3.1) than the Rh1 (444-519 nm) or Rh2 (467-528 nm) photopigments used in our prior studies [73, 95].

In this study, we show that this approach was highly successful at predicting spectral peaks of absorbance values of 11 teleost Sws2 opsins for which sequences and λ_{max} values are known (Figure 3.1, Table 3.1). We identified three parameters of chromophore conformation that together accurately predict the λ_{max} value. Furthermore, we discuss these results in the context of known amino acid substitu-

tions that likely contribute to divergent λ_{max} values [84]. These studies, therefore, not only provide a valuable extension of our prior work, but also guidance and strategic directions for the improvement of functionally-predictive photopigment modeling.

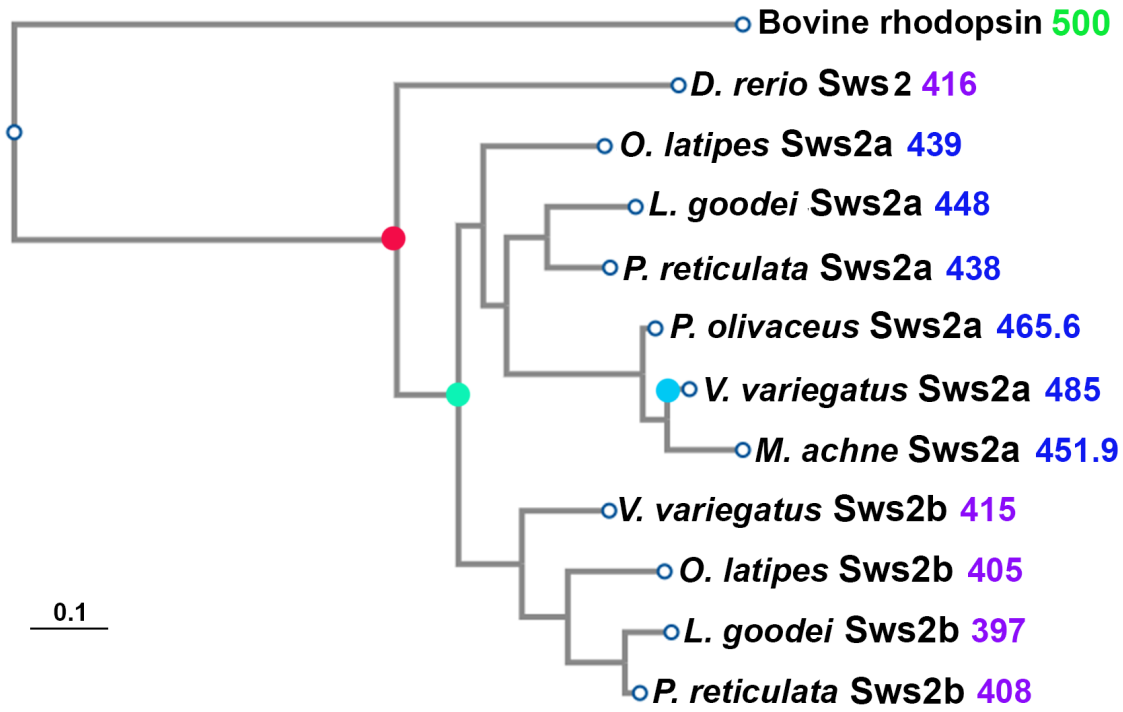

Figure 3.1: Evolutionary relationships of teleost Sws2 opsin proteins used in the simulations, as inferred by PhyML.Red filled circle indicates the speciation event that occurred prior to the duplication (green filled circle) of the sws2 opsin genes in teleosts [1]. The duplication generated the sws2a clade, which encode photopigments with λ_{max} values that are shifted to longer wavelengths, and the sws2b clade, which encode photopigments with short-wavelength-shifted λ_{max} values. The blue filled circle indicates the amino acid substitution A269T (with numbering standardized to the bovine rod opsin sequence) that is likely to be the spectral tuning site important for a further shift of the λ_{max} value of *V. variegatus* Sws2a to longer wavelengths [2]. Experimentally measured λ_{max} values (in nm) are indicated next to the name of each opsin and are color-coded for λ_{max} values that are <430 nm (violet) or >430 nm (blue)

Fish [Reference]	**Target Sws2 Sequences**	**UniProt Accession Number**	**Sequence Identity (%) vs. Bovine RH1**	**Experimentally Measured Spectral Peak of Absorbance (nm)**
Spotted flounder [2]	Sws2a	A0A1L7P076	51.27	485.4
	Sws2b	A0A1L7P082	49.30	415.8
Slime flounder [2]	Sws2a	A0A1L7P074	50.70	451.9
Olive flounder [2]	Sws2a	D7RP09	49.86	465.6
Guppy [100]	Sws2a	A0A140JTJ4	48.02	438
	Sws2b	A0A140JTJ5	48.31	408
Medaka[82]	Sws2a	Q2L6A3	50.00	439
	Sws2b	Q2L6A2	47.75	405
Bluefin killifish [101]	Sws2a	Q7T2U6	50.28	448
	Sws2b	Q7T2U7	49.30	397
Zebrafish [64]	Sws2	Q9W6A8	47.47	416

Table 3.1: Selected Sws2 opsin sequences used for homology modeling. Sequence UniProt accession numbers, percentage sequence identity compared to the bovine RH1 protein sequence, and experimentally measured spectral peaks of absorbance are indicated in third, fourth and fifth columns, respectively.

3.2 Results

To develop a model to predict spectral peaks of absorbance (i.e. λ_{max} values) from a diverse set of teleost Sws2 cone opsins, the following amino acid sequences were selected: two from *V. variegatus* (spotted flounder) [2], one from *Microstomus achne* (slime flounder) [2], one from *Paralichthys olivaceus* (olive flounder) [2], two from *Poecilia reticulata* (guppy) [100], two from *Oryzias latipes* (medaka) [82], two from *Lucania goodei* (bluefin killifish) [101], and one from *Danio rerio* (zebrafish) [64]. These Sws2 opsin amino acid sequences show a wide range of experimentally measured λ_{max} values that range from 397 nm to 485 nm (Table 3.1). Such functional divergence probably evolved as primary *sws2* gene sequences mutated and were positively conserved, which likely led to distinct conformations of the Sws2-associated chromophore, 11-*cis* retinal, in the dark state (Figure 3.2).

We used a similar protocol that was shown to be successful for the prediction of Rh1 and Rh2 λ_{max} values [73, 95], through building homology models using bovine rod opsin (RH1) as the template [102]. RH1 opsins are primarily present in vertebrate rods [71] and are the only mammalian visual pigment

class where the protein structure has been experimentally determined [99, 102]. Therefore, this, and in particular the bovine rod opsin protein structure, is the only template available for accurate homology modeling of vertebrate cone opsins. The sequence identity of the bovine RH1 template compared to teleost Rh1 rod opsins and Rh2 cone opsins ranged from 49-83% and 63-72%, respectively; which proved to be more than sufficient for the reliable modeling and accurate prediction of λ_{max} values [73, 95]. In the present study, we test whether this approach is also reliable for predicting λ_{max} values of more evolutionarily-distant Sws2 cone opsins (Figure 3.1) that have $\sim$48-51% sequence identity to the bovine RH1 template (Table 3.1). Furthermore, the selected Sws2 cone photopigments display a broader range and blue-wavelength-shifted λ_{max} values (i.e. 397-485 nm) compared to either Rh2 cone photopigments or rod opsins presented in prior teleost studies: 467-528 nm for Rh2 [95] and 444-519 nm for Rh1 [73] compared to the bovine RH1 template with a λ_{max} value at 498 nm [72]. Therefore, this present study also tests the capacity of our prior approach of combining homology modeling and MD simulations to accurately predict λ_{max} values for an opsin class that is difficult to predict from just the amino acid sequences, that are short-wavelength-shifted compared to the bovine template, and cover 88 nm of the electromagnetic spectrum.

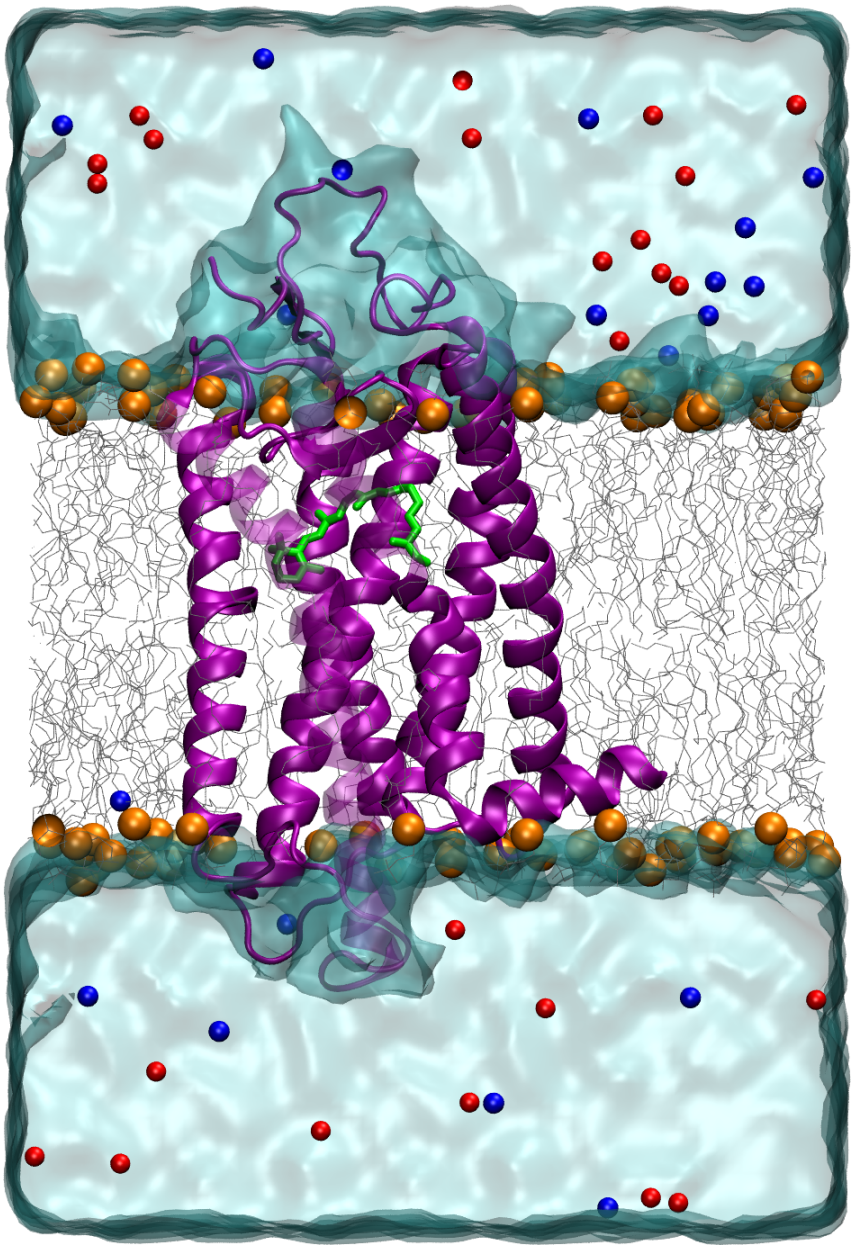

Figure 3.2: A representative 3D structure of Sws2 cone opsin (λ_{max} values <430 nm) homology structure (violet) with the chromophore (green) bound covalently to K296 of the opsin protein. It is inserted in a phospholipid bilayer (gray, carbon atoms; orange, phosphorus atoms) and surrounded by water molecules (light blue). Blue and red spheres indicate positive and negative counter ions, respectively.

Using the sequence information of 11 teleost Sws2 opsins, we built homology models based on the bovine rod photopigment (RH1 opsin + 11-*cis* retinal chromophore) structure as a template (Protein Data Bank (PDB) ID: 1U19) [102]. These homology structures with the chromophore attached covalently to K296 within the binding pocket were placed in lipid bilayers and water models (Figure 3.2). Each of

these systems were then subjected to 100 ns classical MD [103] simulations using the protocol described in the Methods section.

We analyzed the MD simulations for all 11 visual photopigments to understand the dynamics and to identify structural features associated with the chromophore and attached lysine residue (Figure 3.3A) that could potentially be used to explain differences in the $\lambda_{\max}$ values of the representative Sws2 photopigments used in this study. To understand the dynamics of the chromophore within the opsin binding pocket, we visualized the conformations of the chromophore seen in violet- ($\lambda_{\max}$ values <430 nm) vs. blue-sensitive ($\lambda_{\max}$ values >430 nm) photopigments (Figure 3.3B), where we observed a relatively compact cluster of chromophore conformations for blue-sensitive photopigments compared to violet-sensitive photopigments. This difference was more evident for the β-ionone ring and two methyl groups present at positions C9 and C13. In our previous study, the area under the curve (AUC) of root mean square fluctuations (RMSF) served as an additional feature of the chromophore that helped to predict the $\lambda_{\max}$ values of Rh2 cone visual photopigments [95]. We, therefore, first calculated RMSF values of all the heavy atoms of the chromophore and the linked lysine residue (LYS+RET) (Figure 3.3C) for each Sws2 photopigment as follows:

$$\mathrm{RMSF}_{(V)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(v_t - \bar{v})} \tag{3.1}$$

where T is the total number of molecular dynamics trajectory frames (V) and then calculated AUC of RMSF values (AUC $\mathrm{RMSF}_{(\mathrm{LYS+RET})}$) for each Sws2 photopigment. The atoms within the blue-sensitive photopigments clearly show lower RMSF values compared to the violet-sensitive photopigments (Figure 3.3C), suggesting that this chromophore feature may also be useful in predicting of $\lambda_{\max}$ values. Interestingly, this outcome is the opposite to what one would anticipate based upon our prior study, in which lower values of AUC $\mathrm{RMSF}_{(\mathrm{LYS+RET})}$ were associated with photopigments with shorter wavelength $\lambda_{\max}$ values [95].

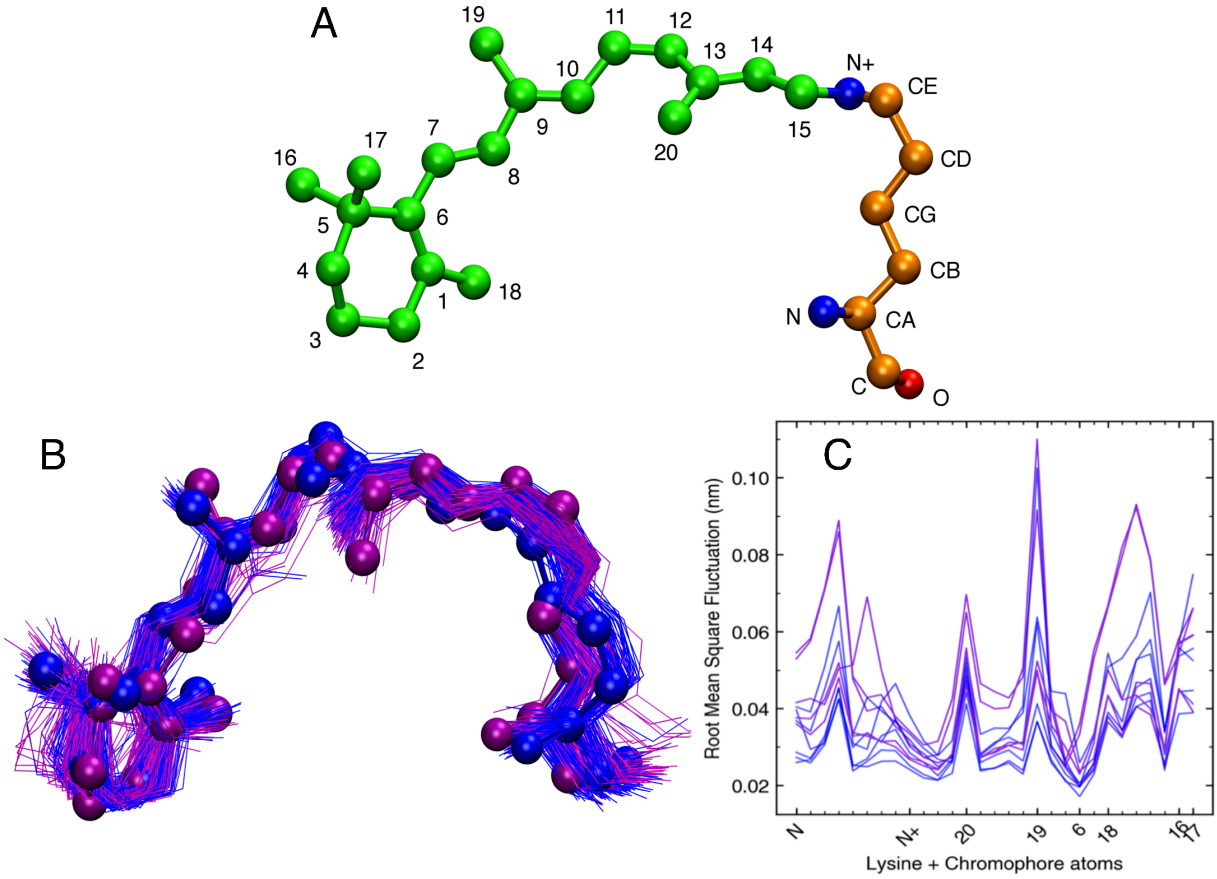

Figure 3.3: Conformations and fluctuations of 11-*cis* retinal chromophore and attached lysine in Sws2 photopigments.A) 3D orientation of 11-*cis* retinal linked to K296 of the opsin protein. B) Superposition of 11-*cis* retinal conformations from MD simulations trajectories. C) Root mean square fluctuation (RMSF) of 11-*cis* retinal linked to K296 (LYS+RET). The horizontal axis represents atoms of the LYS+RET. Blue- vs. violet-colored ball-and-stick conformations are those associated with Sws2 photopigments with λ_{max} values >430 nm vs. λ_{max} values <430 nm, respectively.

The dynamic nature of the chromophore during MD simulations suggests that we may use the geometric angles, dihedrals and AUC $RMSF_{(LYS+RET)}$ parameters to differentiate between blue-sensitive and violet-sensitive Sws2 photopigments. For each photopigment, we examined a total of 19 angles (15 Torsion Angles and 4 Geometric Angles) (Figure 3.3A) formed by the heavy atoms of the lysine residue at position 296 of the opsin covalently linked to 11-*cis* retinal. The model parameters that showed a reasonable correlation to experimental λ_{max} values were the median values of Torsions 3, 9, 10, 11, 12, as well as Geometric Angles 1 and 3, from a total of 19 examined angles. The standard model selection procedure was then used to determine the simplest linear regression model that best fitted to the shortlist of parameters showing a reasonable correlation to experimental λ_{max} values. From our model selection procedure, we found the simplest model for the 11 Sws2 photopigments examined contained three terms:

the median values of Torsion 3 (C15–C14–C13–C20), Torsion 12 (C19–C9–C8–C7), and Angle 3 (C3–C7–C8) (Figure 3.4). AUC $\text{RMSF}_{(\text{LYS+RET})}$ values were not identified by the model selection procedure as being predictively useful.

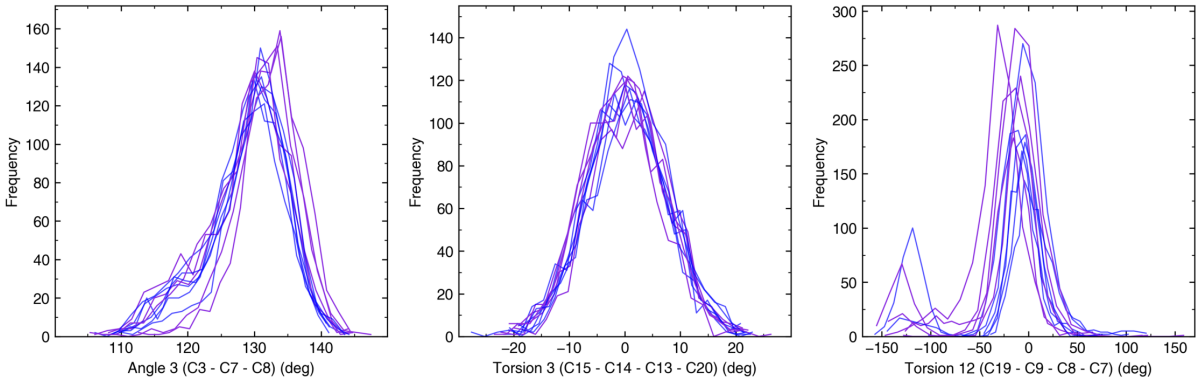

Figure 3.4: Frequency distribution of Angle 3, Torsion 3 and Torsion 12 observed in each opsin simulation. Blue and violet colors correspond to Sws2 photopigments with λ_{max} values >430 nm vs. λ_{max} values <430 nm, respectively.

The full model is explicitly given by:

$$\lambda_{max}(pred.) = 2677.5348 - (17.052 \times Angle\,3) + (5.1634 \times Torsion\,3) + (2.3642 \times Torsion\,12) \quad (3.2)$$

The larger values of Angle 3 predicted a spectral shift towards shorter wavelengths (i.e. a violet shift), while larger values of Torsion 3 and Torsion 12 predicted a shift towards longer wavelengths (i.e. a blue shift). Figure 3.5 shows empirically determined λ_{max} values vs. the model predicted values for each Sws2 photopigment analyzed, where our full model highly correlates with experimental data ($R^2 = 0.95$). The error in prediction
(error = | pred – exp |) ranged from 1.21 nm to 8.35 nm with an average error of 5.44 nm. To further test our statistical model, we carried out a "leave-one-out" analysis, where each Sws2 photopigment was removed from the regression analysis to obtain the coefficients for a model using Angle 3, Torsion 3 and Torsion 12 parameters, and then the λ_{max} value of the removed photopigment was predicted based upon the new linear model. The correlation of the individual predictions based upon only 10 pigments was reduced, but it was still highly acceptable ($R^2 = 0.86$) (Figure 3.5). The lower correlation coefficient derived from the "leave-one-out" approach is largely explained by the less accurate prediction of the λ_{max} value for the zebrafish Sws2 photopigment (i.e. experimental λ_{max} value at 416 nm vs. "leave-one-out" predicted λ_{max} value at 400 nm). This less accurate prediction is likely due to Torsion 3, which has a relatively higher median value compared to Torsion 12 with a relatively lower median value for

zebrafish compared to the median values of Torsion 3 and Torsion 12 of Sws2 pigments from other species. Nevertheless, our results indicate that the statistical model derived from MD simulations of predicted Sws2 visual photopigment structures has the power to predict their λ_{max} values with reasonable accuracy.

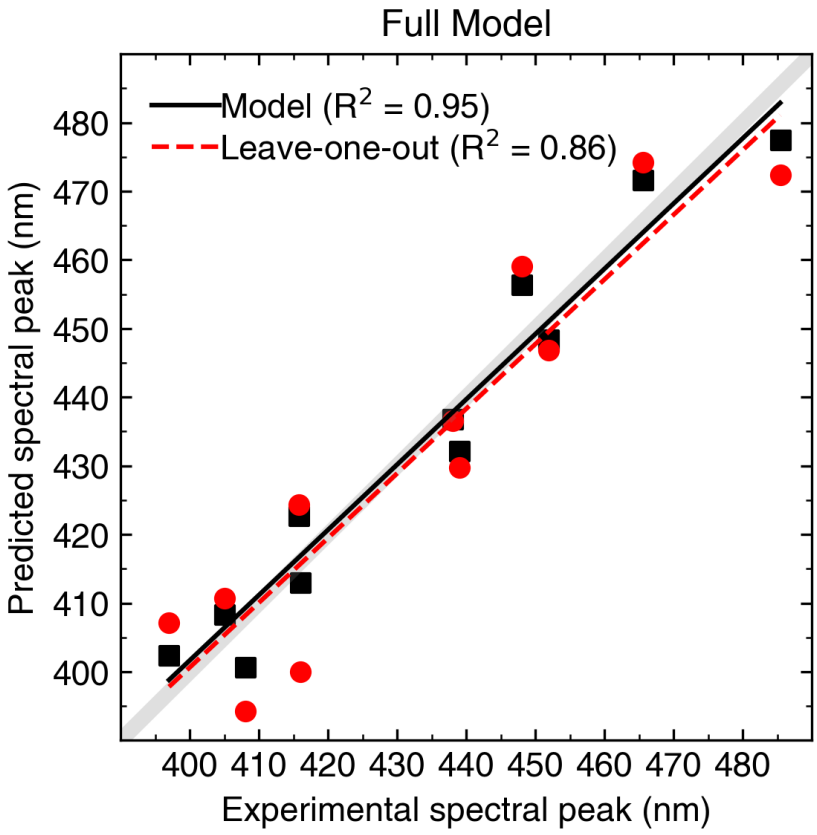

Figure 3.5: Experimental spectral peaks of absorbance (λ_{max}) compared to predicted λ_{max} values by the full model equation 1 outlined in the main text for all 11 Sws2 photopigments analyzed.Gray lines indicate a perfect (100%) correlation. Solid black lines and black symbols represent the linear relationships between model-predicted and the experimental λ_{max} values, whereas dashed red lines and red symbols show linear relationships between "leave-one-out" predictions and experimental λ_{max} values. Corresponding correlation coefficients for both approaches are indicated.

To further test this approach and to simulate performance at predicting unknown λ_{max} values for Sws2 photopigments, we performed a "leave-one-out" approach at a species level. For each species where both Sws2a and Sws2b opsins are represented in the dataset (Table 3.1), a subset was generated where their Sws2 chromophore parameters were removed. A new model based on the newly generated subset was chosen using the full model selection procedure detailed in the Methods section. This new best-fit model was then used to predict the spectral peaks of absorbance for the omitted opsins. In this study, there are four species with both Sws2a and Sws2b photopigments (Table 3.1), namely the medaka, the

bluefin killifish, the guppy, and the spotted flounder. For these four species, the following models were determined:

(i) With the medaka Sws2 photopigment removed:

$$\lambda_{\max} = 2761.5524 + 2.405 \times Torsion\ 12 - 17.64 \times Angle\ 3 \tag{3.3}$$

(ii) With the spotted flounder Sws2 photopigment removed:

$$\lambda_{\max} = 4201.6725 - 1.4975 \times Torsion\ 10 - 24.7764 \times Angle\ 1 - 3.2922 \times Angle\ 3 \tag{3.4}$$

(iii) With the bluefin killifish Sws2 photopigment removed:

$$\lambda_{\max} = 2706.2512 - 3.1214 \times Torsion\ 10 - 13.8215 \times Angle\ 3 \tag{3.5}$$

(iv) With the guppy Sws2 photopigment removed:

$$\begin{aligned} \lambda_{\max} = 5624.1521 + 3.6078 \times Torsion\ 9 - 1.8923 \times Torsion\ 10 - 11.0413 \times Torsion\ 11 \\ + 16.0527 \times Torsion\ 12 - 20.9014 \times Angle\ 3 \end{aligned} \tag{3.6}$$

Performance of the "leave-one-species-out" models varied with R^2 values of 0.84 to 0.93 for the full model and 0.76 to 0.94 for the "leave-one-photopigment-out" approach. This indicates that our approach generates models with good predictive power for new Sws2 sequences with unknown $\lambda_{\max}$ values that are not included in the model selection process. However, the drop in the performance when leaving out certain members can be attributed phylogenetic relationship among the chosen sequences and to the fact that we are leaving out two sequences from a relatively small dataset of 11 Sws2 sequences and leaving only nine opsin sequences to build the model using the shortlisted seven structural parameters. Angle 3 was present in all models (equations 2-5), as well as in the full model (equation 1), and had the largest weighting within each approach, with the exception of the model where the spotted flounder Sws2 modeling experiment was omitted. In this latter model, Angle 1 was also present and had the largest weighting. In general, these findings imply that Angle 3 (C3–C7–C8) is a significant predictor of $\lambda_{\max}$ values for Sws2 photopigments.

To explain the mechanistic role of 11 putative spectral tuning sites at positions 41, 47, 94, 97, 99, 109, 116, 168, 183, 269, 299 revealed by comparative analyses of Sws2a vs. Sws2b sequences [84, 2] on the conformation of the chromophore, as predicted by Cortesi et al. (2015) [84], we carried out MD trajectory analyses using visual molecular dynamics (VMD) via visual inspection. It should be noted that except for positions 94 and 269, the other nine putative tuning sites are located far from the chromophore binding pocket. Specifically, position 94 is located in close proximity to the heavy atoms (CE, N+) (Figure 3.3A) of the K296 residue that is covalently bound to the chromophore, so this site was analyzed in detail. In teleost Sws2b photopigments, position 94 is occupied by a cysteine (C) residue and its sidechain invades the space next to atoms (CE, N+) of K296, resulting in reduced space available for fluctuations of dihedrals/angles of the chromophore. By contrast, position 94 in Sws2a photopigments is occupied by threonine (T), alanine (A) or glycine (G). The T94 residue interacts with S186 and affects the atoms (CE, N+) of the chromophore bound to K296, while A94 and G94 lack large sidechains, thus providing space for the fluctuation of these atoms (CE, N+). To understand the influence of the spectral tuning site at position 94 on the fluctuation of these heavy atoms, we analyzed the dihedrals and geometric bond angles formed by (CE, N+) atoms of K296. We found Angle 2 to be clearly distinguishable for Sws2a compared to Sws2b photopigments, such that shorter wavelength-shifted (i.e. violet) Sws2b photopigments have wider distributions of Angle 2 compared to longer wavelength-shifted (i.e. blue) Sws2a photopigments.

The spectral tuning site at position 269, as described by Kasagi et al. (2018) [2], is occupied by alanine in all of the Sws2 opsin proteins analyzed in this study, except for *V. variegatus* Sws2a, in which this site is occupied by threonine that results in a shift of the $\lambda_{\max}$ value to longer wavelengths. Analysis of MD trajectories reveals that position 269 is located within the chromophore binding pocket and directly interacts with the β-ionone ring of the chromophore. The methyl group of the A269 side chain forms favorable hydrophobic contacts with the β-ionone ring of the chromophore, but with an A269T substitution, the hydroxyl group of threonine results in the β-ionone ring of the chromophore being positioned more distantly. Among all the 19 structural parameters analyzed that are associated with the chromophore, only Torsions 11 and 12 display spatial relationships such that they could be affected by known tuning sites. Interestingly, Torsion 12 is affected by A269T and was indeed useful for predicting $\lambda_{\max}$ values as can be seen in the combined full model.

3.3 Discussion

We have developed a new model for the prediction of the spectral peaks of absorbance for teleost Sws2 cone photopigments, with high accuracy over a wide range of $\lambda_{\max}$ values (i.e. 397-485 nm). Our approach is based on a predictive model as described in our previous studies [73, 95]. In the

present study, this approach required Sws2 opsin protein sequence data as input and a known template photopigment structure to build the homology models. MD simulations were then performed on these opsin homology models, with parameters that describe the conformational change in both opsin structure and the covalently-bound chromophore being extracted. From this, a statistical model was built. Similar to our previous studies, our current approach revealed that the structural features of the chromophore and its lysine attachment site play important roles in the prediction of $\lambda_{\max}$ values. In fact, the final predictive model consisted of three terms associated with chromophore conformation. This simple first-order regression model was found to be sufficient to estimate the $\lambda_{\max}$ values of Sws2 photopigments with high accuracy. This study further highlights the versatility of our approach in reliably predicting the $\lambda_{\max}$ values of evolutionarily more distant Sws2 cone opsin sequences, with ∼48-51% sequence identity to the bovine RH1 template.

A number of molecular and evolutionary approaches have been used in the field of visual neuroscience to predict visual photopigment $\lambda_{\max}$ values. One common strategy is the application of site-directed mutagenesis to opsin sequences derived directly found to be expressed in a particular extant species, where amino acid substitutions are made followed by measuring the spectral characteristics to identify potential contributions of specific amino acid residues to any observed spectral shifts [65, 2, 86, 87, 88, 91]. A similar strategy is to infer the amino acid sequence of the ancestral opsin sequence within a clade, followed by the same technique to experimentally determine $\lambda_{\max}$ values [72, 86]. Whereas the latter approach generally involves investigating multiple amino acid substitutions that may or may not be directly related to spectral tuning, the former technique frequently only studies single residue differences. Nonetheless, these methods are frequently used together and have complemented any comparative analyses that preliminarily identify residues that are likely to influence the $\lambda_{\max}$ value of a particular photopigment (e.g. Cortesi et al. (2015) [84]; reviewed by Shichida and Matsuyama (2009) [104]). For example, such an approach identified that a spectral shift of the $\lambda_{\max}$ value of teleost Rh2 photopigments, specifically with sensitivity in the green region of the visible spectrum to blue, was largely due to an E122Q substitution [85, 86]. Apart from some vertebrate LWS photopigments [93] and UV-sensitive SWS1 photopigments [87, 94], the prediction of other cone opsin $\lambda_{\max}$ values (i.e. SWS2 and RH2 photopigment classes) by manual interrogation of the amino acid sequences alone is extremely difficult and often inaccurate [64, 65]. Experimental spectral analyzes by MSP and/or *in vitro* regeneration of photopigments are not always viable options due to the financial limitations and the demand for specialist technical expertise. As such, alternative methods of accurately predicting $\lambda_{\max}$ values to understand photopigment function and ecological adaptation are critical. Our alternative molecular modeling-based approach is, by contrast, simple, accurate and efficient, and does not require site by site substitutions followed by *in vitro* experimentation or complex

quantum calculations. Previously, we successfully used a similar approach presented in this study to accurately predicted a broad range of λ_{max} values (467-528 nm) for teleost Rh2 cone photopigments [95] and rod (Rh1) photopigments (444-519 nm) [73]. With this current investigation, our approach now provides accurate predictions of λ_{max} values for Sws2 photopigments from 397-485 nm (i.e. those sensitive to violet vs. blue wavelengths). Not only is our approach useful for accurately predicting λ_{max} values using known spectral tuning sites, it may be used to identify and assess the spectral effects of putative unknown residues on the spectral properties of photopigments. It should be noted, however, that the predictions made using our modeling approach results in λ_{max} values that are based on rhodopsins that utilize a vitamin-A_1-derived chromophore. This is also the case for experimental approaches using *in vitro* regeneration protocols [65, 2, 86, 87, 88, 91]. In some vertebrates (e.g. lampreys [92, 105, 106], many freshwater teleosts [107], lungfishes [108, 109], the green anole lizard *Anolis carolinensis* [110, 111]; reviewed in [64, 65, 66, 112], the visual system is based on porphyropsins that incorporate a vitamin-A_2-derived chromophore or a combination of both rhodopsins and porphyropsins. Therefore, within the context of biological relevance, the predicted λ_{max} values that result using our approach may have to be converted, where appropriate, to longer wavelengths to account for the possession of a vitamin-A_2-derived chromophore in native photopigments. This is easily conducted by using a number of rhodopsin-to-porphyropsin transformation algorithms, such as those by Loew and Dartnell [113], Harosi [114], and Whitmore and Bowmaker [77].

One of the key elements of our predictive model is Angle 3 (C3–C7–C8), which is the dominant parameter in predicting the λ_{max} values of Sws2 photopigments. In contrast, the dominant parameters for the λ_{max} predictive models in Rh2 and RH1 class photopigments were AUC $RMSF_{(LYS+RET)}$ and presence and absence of disulfide bridge, respectively [73, 95]. This suggests that the structural features associated with spectral tuning are visual photopigment class specific. Specifically, our results show that larger values of Angle 3 lead to greater shorter wavelength shifts of λ_{max} values to the violet region of the visible spectrum. Furthermore, Angle 3 is an important parameter when the "leave-one-out" approach was applied at a species level for predicting the spectral peaks of absorbance for unknown opsins, suggesting this parameter is broadly predictive. Other parameters affecting the prediction of λ_{max} values are Torsion 3 and Torsion 12, but the magnitude of their effects is reduced in comparison with that of Angle 3. Nonetheless, like Angle 3, larger values of Torsion 3 and Torsion 12 also short-wavelength-shifted the λ_{max} value. Interestingly, Torsion 12 was the only element of the "three-term" model that is likely to be directly influenced by a known or suspected tuning site (i.e. residue 269). These results explain the possible mechanism that causes the observed long-wavelength shift in the λ_{max} value of the spotted flounder Sws2a photopigment. With the exception of this specific example, other

residues affecting Angle 3, Torsion 3, and Torsion 12 cannot be pinpointed to any one particular known or suspected tuning site. Instead, it is likely that multiple tuning sites collectively, either directly or indirectly, influence these chromophore structural features. This finding underscores the power of the homology modeling/MD approach, which takes into consideration the collective influence of the entire amino acid sequence to predict $\lambda_{\max}$ values rather than the sole specific contributions of individual sites. Although predictive and useful for explaining the influence of the known tuning sites, our genome to phenome approach does not provide chemical-physics mechanism to explain the shift in the spectral peaks of absorbance and identification of opsin-charge determinants affecting the chromophore. Such questions would be best answered via thorough QM (quantum mechanics)/MM (molecular mechanics) investigation for each Sws2 sequence [115]. Nonetheless, we believe our study can provide a reasonable starting geometry to carry out such QM/MM simulations.

Within the chromophore of the *D. rerio* (zebrafish) Sws2 photopigment, Torsion 3 has a relatively higher median value compared to the median value range of Torsion 3 of Sws2 photopigments found in other species, while Torsion 12 has a relatively lower median value compared to the median value range of Torsion 12 of other Sws2 photopigments. These "out-of-range" median values of Torsion 3 and Torsion 12 lead to a less accurate prediction of the $\lambda_{\max}$ value for the *D. rerio* Sws2 photopigment. Based on the evolutionary relationships of teleost Sws2 opsin protein sequences examined in this study, it appears that *D. rerio* (and maybe other cyprinids) diverged from the main *sws2* opsin clade before *sws2* duplicated into *sws2a* and *sws2b* subclasses. It is possible, therefore, that the zebrafish Sws2 opsin protein holds the retinal chromophore in a distinct conformation vs. the other Sws2 proteins investigated in this study, but which also results in a $\lambda_{\max}$ value that is similar to that exhibited by the Sws2b photopigment subclass. Thus, teleost Sws2 photopigments may have independently evolved more than one opsin-chromophore conformation strategy for attaining short-wavelength shifts of the $\lambda_{\max}$ value to the violet region of the visible spectrum. Such knowledge means that the model presented here and its predictive power might be improved in the future if a more diverse set of Sws2 photopigments is used to develop a more all-encompassing Sws2 model and/or to develop distinctive models for some selected phylogenetic groups.

Comparative analyzes of Sws2a vs. Sws2b photopigments have revealed candidate spectral tuning sites that could potentially explain the sensitivity to violet vs. blue wavelengths for Sws2b and Sws2s subgroup, respectively [84]. Thus, MD simulations can serve as a valuable, complementary tool to understand the contributions of candidate tuning sites to spectral shifts in the $\lambda_{\max}$ value. From MD trajectory analysis, Angle 2 was identified as a potential parameter that is affected by the presence of different residues at spectral tuning site 94. However, we note that as Angle 2 did not display a significant linear correlation with the actual $\lambda_{\max}$ value, Angle 2 was not considered as a candidate for the model selection procedure;

as such, Angle 2 did not appear in the predictive "three-term" model.

In conclusion, we have successfully tested our previously studied molecular modeling approach combining homology modeling, MD simulations and structural information of chromophore conformations and to accurately predicted the λ_{max} of Sws2 opsins. In future studies, we plan to consider additional features of each parameter (e.g. narrow vs. broad distribution) other than simple linear correlations, and other structural features such as distance between certain atoms and the functional group of the chromophore or the opsin tuning sites, which may further improve the predictive power of the resulting models. We will also expand our approach to extensively study the more phylogenetically distant classes of opsins, with large number of opsins included in the dataset. Finally, we aim to develop model(s) using a dataset of individual distinct classes of opsins to generate an online web platform for accurately predicting the λ_{max} values for any unknown vertebrate photopigment.

3.4 Methods

3.4.1 Phylogenetic analysis

The alignment and phylogenetic reconstructions were performed using the function "build" of ETE3 v3.1.1 [116] as implemented on the GenomeNet (https://www.genome.jp/tools/ete/). The multiple sequence alignment was provided as input file. ML tree was inferred using PhyML v20160115 ran with model and parameters: --alpha e --pinv e -f m -o tlr --bootstrap -2 --nclasses 4 [117]. Branch supports are the Chi2-based parametric values return by the approximate likelihood ratio test.

3.4.2 Homology modeling

Eleven teleost Sws2 cone opsin amino acid sequences (Table 3.1), namely *V. variegatus* (spotted flounder) [2], *M. achne* (slime flounder) [2], *P. olivaceus* (olive flounder) [2], *P. reticulata* (guppy) [100], *O. latipes* (Japanese rice fish; medaka) [82], *L. goodei* (bluefin killifish) [101], and *D. rerio* (zebrafish) [64], were downloaded from the UniProt database (https://www.uniprot.org/). These were selected because the corresponding spectral peaks of absorbance for their Sws2 photopigments (when reconstituted with a 11-*cis* retinal chromophore) have been experimentally measured. Collectively, these photopigments exhibit a wide range of λ_{max} values from 397-485 nm. An experimental 3D structure of a Sws2 cone photopigment is not available; hence, to build a homology model of 11 Sws2 photopigments, a template search was carried out using SWISS-MODEL (https://swissmodel.expasy.org/). The closest homologue ($\sim$50% sequence identity) with a high-quality 3D structure was found to be that of the bovine rod opsin (RH1). A high-resolution crystallographic structure of the bovine RH1 photopigment (PDB ID 1U19, 2.2 Å) [99], which lacks mutations and has an 11-*cis* retinal chromophore covalently bound within its binding

pocket, was downloaded from the Protein Data Bank. 3D coordinates of the bovine rod opsin structure were then used to build the homology models of 11 teleost Sws2 cone opsin sequences. The structure prediction wizard from the PRIME module of the Schrödinger suite was used for building a homology model for each protein sequence [118, 119]. The non-templated loops were refined using the refine loops module of PRIME and a generated model structure was validated by generating a Ramachandran plot by analyzing acceptable phi-psi regions of residues. The final homology model was modified to remove the intracellular unstructured coil (∼25 residues) region towards the carboxyl-terminus to prevent it from crossing the periodic boundaries during the molecular dynamics (MD) simulation.

3.4.3 Molecular dynamics (MD) simulation

All 11 Sws2 photopigment homology models were subjected to atomistic MD simulations using our protocol for input file generation and the system setup for MD simulations reported in our previous studies [73, 95]. Briefly, the homology models of each Sws2 opsin sequence with the chromophore bound covalently to the lysine residue in the binding pocket (K296) were uploaded to the CHARMM-GUI webserver (http://charmm-gui.org). Each system was placed in lipid bilayers and hydrated using a hexagonal solvent box with a 15 Å TIP3P water layer. The charge of the system was neutralized with 150 mM NaCl. The CHARMM36m forcefield [93] parameters were selected for all the components of the systems. After minimization and short equilibration simulations with harmonic restraints, the systems were subjected to 100 ns atomistic MD simulations. The production simulations were performed under an NPT ensemble for 100 ns using a Parrinello-Rahman barostat [120] with semi-isotropic pressure coupling and a Nosé-Hoover thermostat [121]. During the production MD simulations, snapshots were saved every 10 ps. GROMACS-2018.3 [122] was used for all 11 MD simulations. The visualization and analysis of MD trajectories were carried out using Visual Molecular Dynamics package [123]. GROMACS trajectory analysis tools were used to analyze Root Mean Square Fluctuations (RMSF) and 19 different internal degrees of freedom of the chromophore (i.e. torsion angles and geometric bond angles).

3.4.4 Quantification and statistical analysis

To determine the best linear regression model for predicting Sws2 λ_{max} values, and to avoid overfitting (i.e. a model with the same number of terms as the number of datapoints it is being fit to) we chose a shortlist of parameters to use in model selection based on the linear correlation of each of the 19 median angles and the experimental λ_{max}. These were then ranked by r^2 (i.e. coefficient of determination) values and the top seven performing angles were used, exception being taken with angles which did not show true linear correlation through visual inspection. The resulting model selection parameters included the

median values of Torsions 3, 9, 10, 11, 12, and Geometric Angles 1 and 3. From our list of 19 parameters, a shortlist was generated that composed of the medians of Torsion 3, Torsion 9, Torsion 10, Torsion 11, Torsion 12, and geometric Angles 1 and 3. A model selection procedure was then performed using the regsubsets function of the "leaps" R library (https://cran.r-project.org/web/packages/leaps/leaps.pdf). This evaluated all possible model subsets using the shortlisted angles and ranked them according to their Bayesian Information Criterion (BIC) value [124], which quantifies the explanatory power of a model with a penalty for the number of terms included. The resulting best-fit model was further evaluated via a "leave-one-out" procedure by reweighting the parameters of the best-fit model after removing each photopigment iteratively. For each Sws2 photopigment stimulation, the parameters of the best-fit model were reweighted with the given Sws2 photopigment data removed. This reweighted model was subsequently used to predict the spectral peak of absorbance for the omitted Sws2 opsin sequence.

Additionally, to further validate our approach and to simulate performance at predicting the spectral peaks of absorbance for Sws2 photopigments with unknown $\lambda_{\max}$ values, a systemic "leave-one-species-out" process was also performed. For a given species, both Sws2a and Sws2b opsin sequences were removed from the dataset. A new model was chosen using the full model selection procedure detailed above. Since removing two opsins results in a smaller sample of data, overfitting (i.e. a model with the same number of terms as the number of datapoints it is being fit to) is a possible risk factor. To mitigate this, the generated model space was explored within +2 BIC value of the best BIC ranked value. The resulting model was then used to predict the spectral peak of absorbance for the removed Sws2a and Sws2b photopigments.

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

Tawny R. Gonzalez,[2] Kyle P. Martin,[1,2] Jonathan E. Barnes,[1,2] 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

Published in PLOS ONE. As a contributing author, I performed the following:

- Final generation of all figures for publication.
- Validated all data.
- Used BindProfX to generate predictions to include standalone data.
- Reviewed literature and helped write manuscript.

Writing and editing was done with google docs and a word document 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 published by PLOS open access, with a creative commons license, and can be used freely in this dissertation. Final publication is available through PLOS: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240573 .

4.1 Introduction

Protein-protein binding is an essential physiological event that governs a large number of biological processes in the cell [125]. Amino acid mutations of these proteins can introduce diversity into genomes, and disrupt or modulate protein-protein interactions by changing the underlying binding free energy (ΔG, i.e. binding affinity), the amount of energy required to form protein complexes [126]. The binding free energy associated with a protein-protein complex determines the stability of the complex formation and the conditions for protein-protein association. Accurate prediction of binding free energies allows us to understand how these affinities can be modified, and leads to a more comprehensive understanding of protein interactions in living organisms [127].

Experimental biophysical methods can quantitatively measure change in the protein-protein binding free energy due to a mutation (i.e. relative binding affinity, $\Delta\Delta G$), but these methods are typically costly, laborious, and time-consuming since all mutant proteins must be expressed and purified. Many researchers have developed and utilized computational methods to predict $\Delta\Delta G$ values for single- or multiple-amino acid mutations (see e.g. [128, 129, 130]). Historically, the most promising in terms of accuracy are rigorous methods based on statistical mechanics that use molecular dynamics (MD) simulations and thus automatically address conformational flexibility and entropic effects [131, 132]. However, these methods are computationally expensive since they employ rigorous sampling and utilize classical mechanics [133] or quantum mechanics [134] approximations of intermolecular interactions, and require a large number of calculations per time-step. Because of the expense, rigorous methods are not well-suited to studying large sets of mutations or large proteins thus necessitating less expensive, non-rigorous methods.

Non-rigorous high-throughput methods attempt to lower the computational cost, as compared to rigorous methods, while still providing accurate $\Delta\Delta G$ predictions. They accomplish this by including precalculated physico-chemical structural information in combination with predictive algorithms. The core mechanics that drive these methods fall under numerous classification umbrellas which have been covered by review articles [135, 136]. These review articles provide a broad overview but do not provide an unbiased, rigorous, comparative analysis outside of what the original developers provide. The developers of any given method tend to provide comparisons with other methods of the same general class to define where their method fits in the current landscape. BindProfX, for example, is available as a web server and standalone and utilizes structure-based interface profiles with pseudo counts. Upon release, it was most notably compared to FoldX (a semi-empirical trained method [137]) and DCOMPLEX (a physics-based method [138]) [139, 140]. iSEE, a statistically trained method based on 31 structure, evolution, and energy-based terms was tested against FoldX, BindProfX, and BeAtMuSiC (a machine learning-based approach [141]). Mutabind [142] and some other methods not explored in this work follow a similar testing methodology [143, 144, 145]. While these comparisons are beneficial in providing context for how a given model fits in the existing research landscape, they are not very robust, since only a narrow subset of methodologies are included. Conversely for folding stability, Kroncke et al. compared a large number of available software methods on a small dataset of transmembrane proteins providing a general overview of performance [130]. Despite the narrow dataset, this study provides a diverse, useful collection of evaluation metrics between multiple classes of methods. Our intent in this study is to provide a similar robust comparison of methods for non-rigorous binding affinity estimation.

In this work, we evaluate the ability of eight non-rigorous methods to predict relative binding affinities due to single amino acid mutations. We restrict our study to cases where both an experimental structure

of the complex, and experimentally determined binding affinity values are available. To investigate the trade-off between speed and accuracy, we chose 16 protein-protein test complexes with empirical $\Delta\Delta G$ values for observed mutations. We calculated the $\Delta\Delta G$ values for each mutation using all eight methods and compared the results against empirical $\Delta\Delta G$ values. The goal of this study was to determine whether software methods that use (most costly) energy functions with a wider variety of physico-chemical structural features would provide more accurate binding affinity and interface destabilization predictions compared to those that rely on a single descriptive (less costly) energy function. We have determined scenarios in which some of these methods may be better or worse than traditional computational methods in predicting $\Delta\Delta G$ values.

4.2 Methods

4.2.1 Compilation of Experimental ΔΔG Values

To assess the performance of a range of protein-protein binding affinity prediction methods, we first assembled a dataset containing single amino acid mutations with known experimental $\Delta\Delta G$ values. This list was assembled from Structural Kinetic and Energetic database of Mutant Protein Interaction (SKEMPI) version 2.0 [146]. SKEMPI uses data from a variety of different biophysical measurement techniques; these are converted to $\Delta\Delta G$ values if not explicitly reported. Overall, the error associated with experimental $\Delta\Delta G$ values reported in the SKEMPI dataset is thought to range from 0.25 to 1 kcal/mol [146]. While generating this list, we considered four aspects: (i) type of 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 4.1 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.

4.2.2 Selection of Protein-Protein Binding Affinity Methods

Binding affinity prediction methods were chosen to have both a distinct approach to binding affinity calculation that utilized 3-D structural information and had functional standalone software in September 2020, available either online or upon request to the author. Table 4.2 summarizes the methods selected in this study, their approaches, and their type of scoring functions. For simplicity, we categorized scoring functions (mathematical functions to calculate $\Delta\Delta G$ values) as semi-empirical, statistical, or physics-

Non-Ab			Ab		
PDB	# Mutations	# Residues	PDB	# Mutations	# Residues
1a4y [147]	32 [148, 149, 150, 151, 152, 153]	583	1bj1 [154]	10 [155, 156]	547
1brs[157]	30 [158, 159, 160, 161]	199	1jrh [162]	42 [163, 164]	540
1cbw [165]	31 [166, 167]	299	1mlc[168]	11 [169]	561
1iar [170]	36 [171]	336	1vfb [172]	48 [169, 173, 174]	352
1jtg [175]	37 [176, 177, 178, 179, 180]	428	1yy9 [181]	16 [169, 182]	1058
1lfd [183]	19 [184, 185]	254	2jel [186]	43 [187]	520
1ppf [188]	190 [189, 190]	274	3hfm [191]	71 [192, 193, 194, 195, 196]	558
2wpt [197]	26 [198, 199, 200]	220	4i77 [201]	12 [202]	549

Table 4.1: Dataset used in our study containing 16 protein complexes. For both non-Ab (left) and Ab (right) categories, columns show PDB IDs, total number of residues in a complex, and number of experimental mutants per complex.

based. Semi-empirical methods replace as many calculations as possible with pre-calculated data and are trained using existing crystal structures and known binding affinity measurements for mutations [203]. Statistical methods use pre-calculated data and consider changes in coarse structural features such as the change in overall volume [204]. Physics-based methods use molecular mechanics based-energy functions to estimate enthalpic binding contributions [138]. In general, statistical or semi-empirical scoring functions involve a training step where existing datasets are leveraged to determine the weight of input parameters. MD, JayZ, and EasyE were not developed by training the methods against experimental data designed to improve predictive power while all other methods utilized this step.

4.2.3 Calculation and Comparison of Computational Speed

The methods in Table 4.2 were used to predict $\Delta\Delta G$ values for each mutation on our experimental list shown in Table 4.1. Runtimes were determined by using a representative protein complex from each category: 1ppf, a non-Ab complex with 274 total amino acids, and 1yy9, an Ab complex with 1058 total amino acids (see Table 4.2). These runtimes are estimates provided to give a point of comparison between the speeds of different methods. Specific runtimes will be determined by hardware specifications, method parameters, the number of mutations being computed, and overall protein size. For MD+FoldX, computational runtime was the length of time of the MD simulation plus the FoldX runtime for a single mutation. Reporting runtime in this fashion highlights the large CPUh requirement needed in order to add the sampling of MD into FoldX calculations. We note that, in contrast to the other methods tested here, the MD simulations that must be performed for MD+FoldX can be completed very quickly on modern GPUs, significantly offsetting the high initial cost of the MD+FoldX method. For all other

Name	Brief Description	Scoring Function	Runtime (CPU hours)
BindProfX [139, 140]	Interface profile score based on conservation of homologous interfaces	Semi-Empirical	1ppf = 0.57 CPUh 1yy9 = 0.73 CPUh
BindProfX(BPX) + FoldX v4 [139, 140]	Profile score weighted and combined with FoldX energy potential	Semi-Empirical	1ppf = 0.62 CPUh 1yy9 = 0.71 CPUh
iSEE [205]	Random forest model using structural, evolutionary, and energy-based features	Statistical	1ppf < 0.01 CPUh 1yy9 < 0.01 CPUh
DCOMPLEX v2 [138]	Structural ideal-gas reference state potential	Physics-Based	1ppf = 0.013 CPUh 1yy9 = 0.001 CPUh
EasyE v1.0 [204, 206]	GMEC-based method utilizing the Rosetta [207, 208] energy function	Statistical	1ppf = 0.48 CPUh 1yy9 = 0.09 CPUh
JayZ v1.0 [204, 206]	Partition-function method utilizing Rosetta energy function	Statistical	1ppf = 0.14 CPUh 1yy9 = 0.21 CPUh
FoldX v4 [137, 203]	Empirical energy score based on various energy parameters (e.g. van der Waals, solvation, electrostatics, hydrogen bonding)	Semi-Empirical	1ppf = 0.42 CPUh 1yy9 = 0.16 CPUh
MD+FoldX v4 [209, 210, 211]	Molecular dynamics used to explore conformation space and generate snapshots; FoldX score calculated for each snapshot and averaged	Semi-Empirical	1ppf = 941 CPUh 1yy9 = 4093 CPUh

Table 4.2: Methods used for comparison in study with a short summary of their approach and scoring function. Columns (left to right) indicate the method, a brief description of the method, the type of scoring function used, and runtimes. Runtimes are the amount of CPU hours for estimating the $\Delta\Delta G$ for a representative protein complex for Ab (1yy9, 1058 residues) and Non-Ab (1ppf, 274 residues) categories. Although 1yy9 is roughly four times bigger than 1ppf, the total runtime may or may not be affected depending on the method used.

Correlation	Brief Description	Type
Concordance	The concordance correlation coefficient (ρ_c) measures the degree to which the predicted $\Delta\Delta G$ value equals the actual experimental value (0 indicates no agreement and 1 perfect agreement).	Linear
Pearson	The Pearson correlation coefficient (r) measures the degree to which a uniform linear transformation of the predicted $\Delta\Delta G$ values (i.e., a shift and scale change) would yield the actual experimental values (0 indicates no agreement after transformation, 1 perfect agreement, and -1 perfect inverse agreement).	Linear
Kendall and Spearman	The rank correlation coefficient measures the degree to which the rank ordering of the predicted $\Delta\Delta G$ values matches the rank ordering of the actual experimental values (0 indicates no agreement after transformation, 1 perfect agreement, and -1 perfect inverse agreement). In a normal case, the Kendall correlation (τ) is considered more robust than the Spearman correlation (ρ) because of a smaller gross error sensitivity and more efficient due to a smaller asymptotic variance [212].	Rank
AUC and ROC	The receiver operating characteristic (ROC) curve tests several cutoff values for binning mutations as neutral or destabilizing between the most negative calculated $\Delta\Delta G$ value and the most positive calculated $\Delta\Delta G$ value, with true positive rates (sensitivity) calculated at each point. As the true positive rate is calculated, the classifier is moved to less extreme values; this yields the ROC curve. The area under curve (AUC) is a summary statistic that approximates how well the predictor actually discriminates between the two classifications.	N/A

Table 4.3: Statistical measures used to test the performance of each method in predicting $\Delta\Delta G$ values

methods, the algorithms rely either on various pre-calculated data or limited conformational sampling to calculate $\Delta\Delta G$ values rapidly.

4.2.4 Comparing Experimental and Predicted ΔΔG Values

To carry out statistical analysis of our results we built an in-house Python script (see S2 File) that uses a combination of libraries including matplotlib, numpy, pandas, statistics, scipy, and sklearn. Using this script, we compared predicted values to experimental $\Delta\Delta G$ values for each method.

To evaluate the predictive ability of each method tested, we compared the following correlation coefficients using our script: concordance (ρ_c), Pearson (r), Kendall (τ), and Spearman (ρ) (see Table 4.3). We distinguish between methods that were trained 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 since for optimal performance we expect a regression line that passes through the coordinate origin and has a slope of 1, leading to a correlation coefficient equal to 1.

To compare the discriminating power of the methods, we generated receiver operating characteristic (ROC) curves (see Table 4.3). 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 (AUC, 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: wild type amino acid type, mutant amino acid type, protein-protein interacting versus antibody-antigen, secondary structure classification of the mutation [213, 214], coordination number [215], Sneath index [216], mostly α-helical proteins versus mostly β-sheet proteins versus a mix of both α-helical and β-sheet 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 [217, 218, 219]. The script uses data from S3 File as an input and 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

The purpose of our study was to assess the ability of eight different relative binding affinity calculation methods (see Table 4.2) to estimate $\Delta\Delta G$ values. We selected 16 different protein complexes (eight Ab, eight non-Ab, see Table 4.1) with a total of 654 single amino acid mutations. Each method was then used to estimate $\Delta\Delta G$ values of 654 mutations and a variety of statistical measures were employed to assess their predictive ability. We also examined the computational speed of each method in the context of accuracy to determine its efficiency.

4.3.1 Non-Antibody-Antigen (non-Ab) Results

Our dataset of eight non-Ab test protein complexes contains 401 total mutations and are mainly classified as protein-protein systems formed by inhibitors and receptors that range from 199 to 583 residues in size. The distribution and our classification of experimental $\Delta\Delta G$ values for all non-Ab test complexes is as follows: 13% of point mutations resulted in $\Delta\Delta G$ values less than -0.5 kcal/mol (classified as destabilizing); 31% between -0.5 and 0.5 kcal/mol (neutral); and 56% greater than 0.5 kcal/mol (stabilizing).

Figures 4.1 (blue data points and values) and 4.2 show various performance metrics for each method to assess their ability to predict the non-Ab $\Delta\Delta G$ values. Overall, EasyE has the highest correlation coefficient, $r = 0.62$, and iSEE has the lowest, $r = 0.17$ (see Figures 4.1 and 4.2). JayZ and EasyE, both of which utilize Rosetta's conformational sampling algorithms, consistently have the best r values for non-Ab mutations.

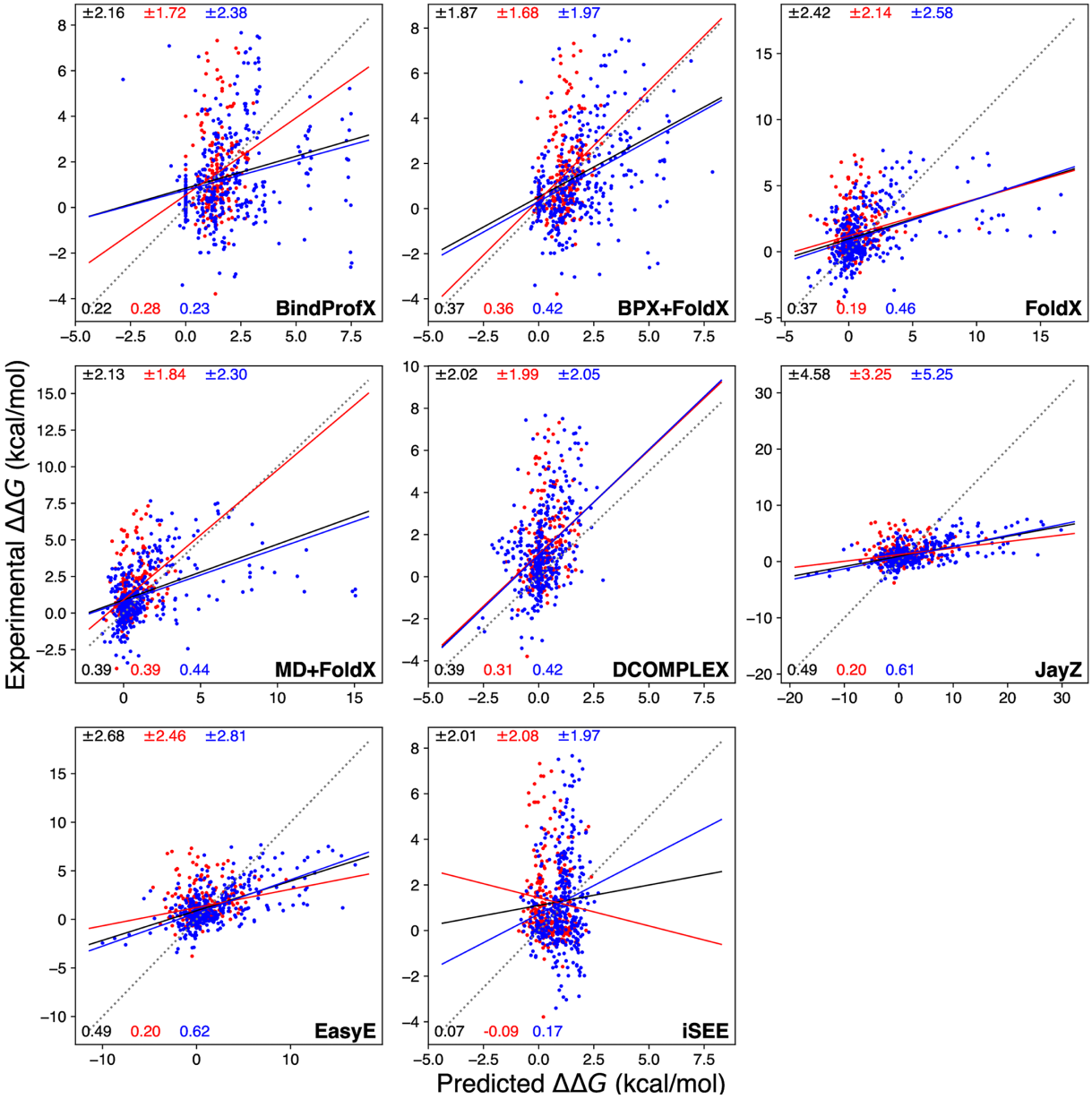

Figure 4.1: Calculated $\Delta\Delta G$ values (x-axis) compared to experimental $\Delta\Delta G$ values (y-axis) for each method tested in this study.Black, red, and blue lines are simple linear regressions from which r are derived. The red points are a scatter for Ab complexes and the blue points are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

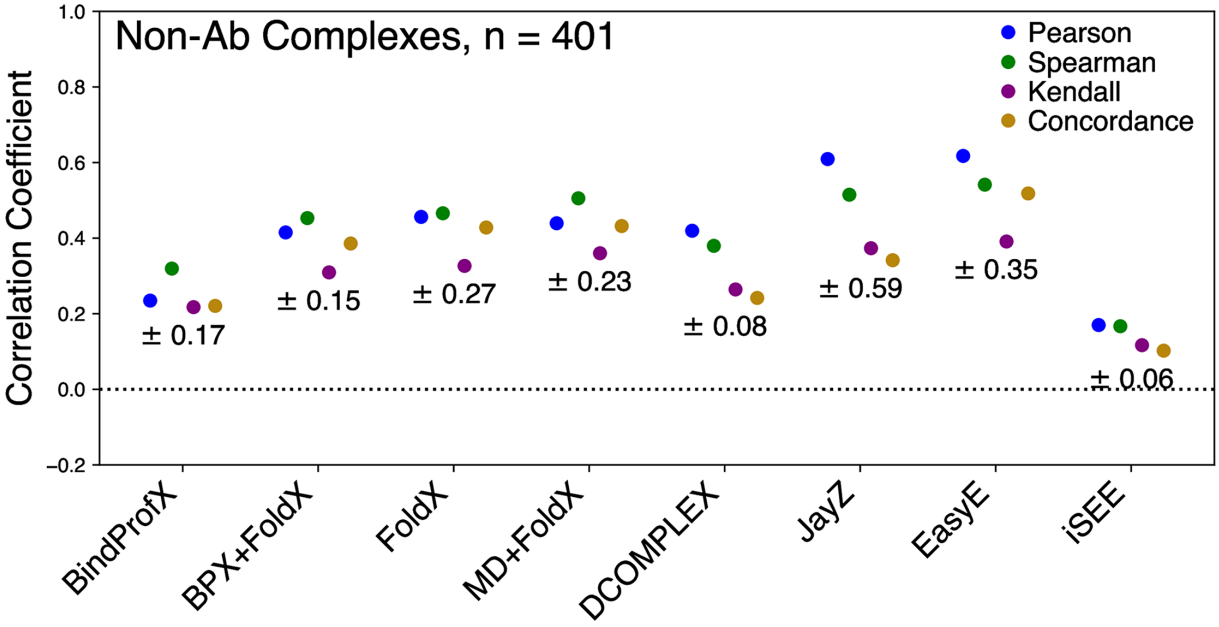

Figure 4.2: Performance of each method for non-Ab complexes (401 total mutations) in predicting true $\Delta\Delta G$ values (ρ_c), linearly correlated $\Delta\Delta G$ values (r), and rank order (ρ and τ). The error for each method is reported under the correlation points.

Figure 4.3 shows the ROC plot for all the tested methods. These ROC plots highlight how well a method can discriminate between stabilizing and destabilizing mutations. JayZ (0.84), EasyE (0.83), DCOMPLEX (0.82), FoldX (0.79), and MD+FoldX (0.76) have the highest AUC. Combined with the results from Figures 4.1 and 4.2, for the systems studied here, JayZ and EasyE methods are the best overall performers in terms of accuracy, discriminating stabilizing mutations from destabilizing, and ranking mutations based on their experimental $\Delta\Delta G$ values.

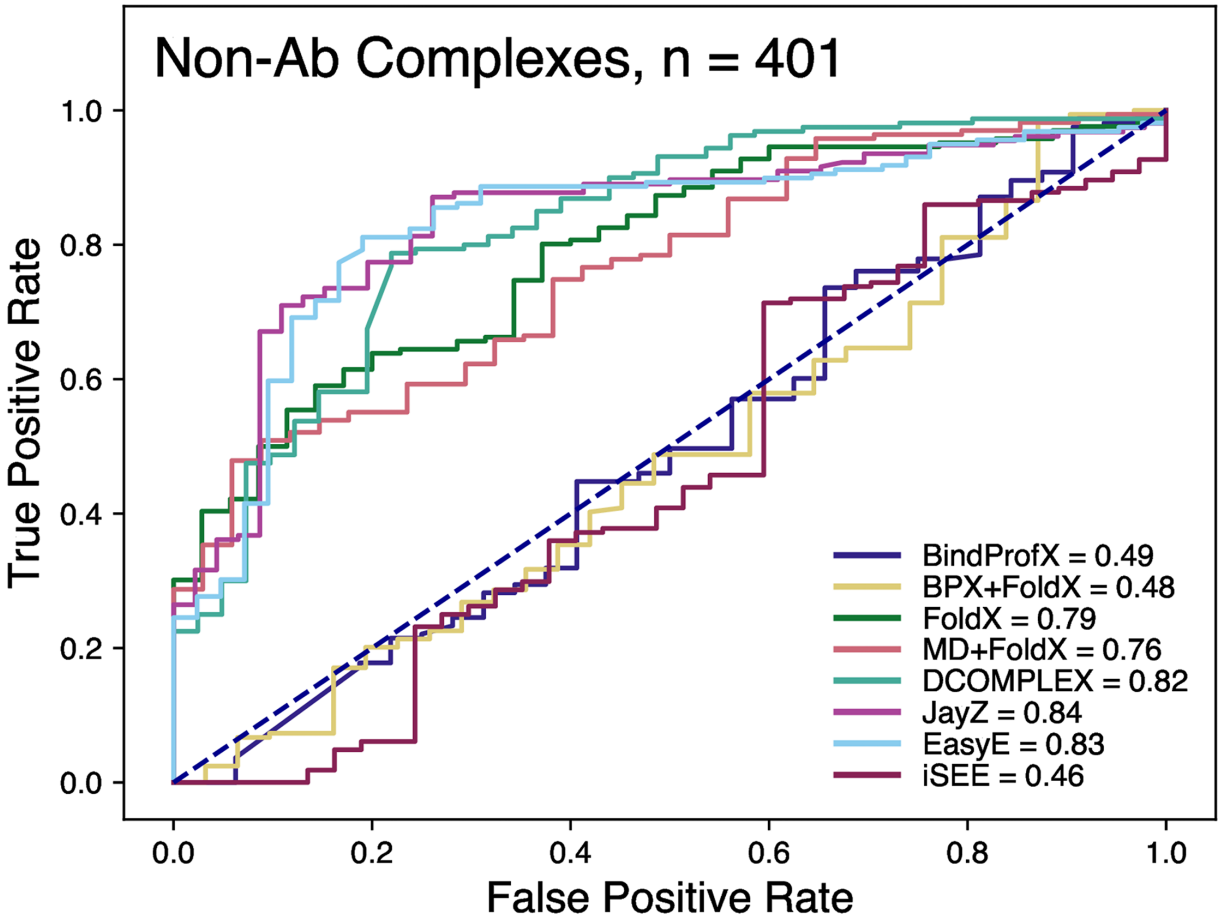

Figure 4.3: Receiver operating characteristic (ROC) curves for non-Ab complexes of the classification of variants as stabilizing ($\Delta\Delta G$ < -0.5 kcal/mol) or destabilizing ($\Delta\Delta G$ > 0.5 kcal/mol). The values in the legend represent the area-under-curve (AUC). The higher the value, the better method is at discriminating between destabilizing and destabilizing mutations.

Table 4.2 reports CPUh required (i.e. runtimes) for each method to calculate $\Delta\Delta G$ for the entire list of mutations for a representative non-Ab protein complex. BindProfX, BindProfX(BPX)+FoldX, JayZ, and EasyE allow users to specify a list of mutations that the method is then able to calculate in one setting. This list can be optimized based on the available hardware to achieve efficiency. iSEE requires significant preparatory work (see File S1) prior to calculation, but once completed, it calculates the $\Delta\Delta G$ values for the entire list of mutations nearly instantly. DCOMPLEX is not as flexible out of the box but can handle large numbers of mutations through an automated script. For MD+FoldX, 1yy9 (roughly four times larger than 1ppf) requires considerably more CPUh to calculate. All other methods calculate 1yy9 in a shorter time frame than 1ppf. This may seem counterintuitive. However, MD must statistically sample the conformational energy of the entire complex, while all other methods use algorithms that are likely impacted more by the number of residues involved in the interaction rather than the protein

size. Overall, DCOMPLEX has a much faster runtime compared to other methods, and if the goal is to determine stabilizing and destabilizing non-Ab mutations, it offers similar discriminating power to JayZ and EasyE, at a fraction of the computational cost. JayZ estimates $\Delta\Delta G$ value of one mutation in ∼2.7 s, EasyE in ∼9.1 s, but DCOMPLEX requires just ∼0.25 s. Overall, EasyE appears to be the best option for balancing accuracy and speed and DCOMPLEX is recommended for discriminating between stability and destabilizing mutations.

A method might not be a good overall performer in predicting $\Delta\Delta G$ values but could still perform well for mutations with certain physico-chemical and structural features. Therefore, we calculated various statistical measures to assess each method on unique subsets of mutations (see Table 4.4. This table shows eight different data subsets with two r per method. EasyE has the highest r for non-Ab for five out of eight subsets (wild type non-gly or non-pro, alpha helix, beta sheet, surface exposure, and large volume changes). Where this method did not have the highest r, it had either the second or third highest r. JayZ mirrors the performance of EasyE in all the same categories and performs better than Easy in the neutral charge subset. These results further highlight the versatility of EasyE's and JayZ's performance in estimating the effects of non-Ab mutations compared to the other methods tested in this study. All methods apart from iSEE and BindProfX perform surprisingly well in the WT Gly or Pro subset. iSEE's performance in this subset, while still mediocre compared to the other tested methods, is substantially better than in all other subsets.

4.3.2 Antibody-Antigen (Ab) Results

Our dataset of eight Ab test protein complexes contains 253 mutations and the proteins range in size from 352 to 1058 residues. The distribution and our classification of experimental $\Delta\Delta G$ values for all Ab test complexes is as follows: 5% of point mutations resulted in $\Delta\Delta G$ values less than -0.5 kcal/mol (classified as destabilizing); 40% between -0.5 and 0.5 kcal/mol (neutral); and 55% greater than 0.5 kcal/mol (stabilizing).

Figures 4.1 (data points and values in red), 4.4, and 4.5 show the performance of each method in predicting the $\Delta\Delta G$ values of Ab mutations. Overall, the highest correlation is for MD+FoldX with r = 0.39 and the lowest is iSEE with r = -0.09 (see Figures 4.1 and 4.4). An interesting trend is that the methods with the highest r values for non-Ab complexes do not have the highest r for Ab complexes.

Figure 4.5 shows the ROC plot for all the tested Ab methods. These ROC plots highlight how well a method is actually able to discriminate between stabilizing and destabilizing mutations. Compared to non-Ab complexes, all methods performed better for antibody-antigen complexes except for FoldX and DCOMPLEX which were marginally worse. JayZ (0.97), EasyE (0.98), FoldX (0.85), and MD+FoldX

Method	WT Gly or Pro	WT Non-Gly or Non-Pro	Alpha Helix	Beta Sheet	Surface Exposure	Neutral Charge	Hydrophobic to Polar	Large Vol Changes
BindProfX	Non-Ab: 0.08 Ab: -0.03	Non-Ab: 0.34 Ab: 0.33	Non-Ab: 0.29 Ab: 0.16	Non-Ab: 0.34 **Ab: 0.48**	Non-Ab: 0.22 Ab: 0.31	**Non-Ab: 0.37** Ab: 0.45	Non-Ab: 0.33 Ab: 0.29	Non-Ab: 0.13 Ab: 0.38
BPX+FoldX	**Non-Ab: 0.78** Ab: 0.09	Non-Ab: 0.46 Ab: 0.43	Non-Ab: 0.43 Ab: 0.38	Non-Ab: 0.35 Ab: 0.52	Non-Ab: 0.32 Ab: 0.40	Non-Ab: 0.52 **Ab: 0.56**	Non-Ab: 0.41 Ab: 0.35	Non-Ab: 0.71 Ab: 0.43
FoldX	**Non-Ab: 0.83** Ab: -0.11	Non-Ab: 0.45 Ab: 0.25	Non-Ab: 0.39 Ab: 0.25	Non-Ab: -0.05 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: 0.84** **Ab: 0.71**	Non-Ab: 0.49 Ab: 0.42	Non-Ab: 0.44 Ab: 0.54	Non-Ab: 0.08 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: 0.22 Ab: 0.30	Non-Ab: 0.52 Ab: 0.27	Non-Ab: 0.36 Ab: 0.56	Non-Ab: 0.38 Ab: 0.16	Non-Ab: 0.62 Ab: 0.28
JayZ	Non-Ab: 0.77 **Ab: 0.54**	Non-Ab: 0.49 Ab: 0.24	Non-Ab: 0.44 Ab: -0.06	Non-Ab: 0.30 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: 0.78 Ab: 0.29	Non-Ab: 0.52 Ab: 0.22	Non-Ab: 0.51 Ab: 0.06	Non-Ab: 0.36 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: 0.32 **Ab: 0.43**	Non-Ab: 0.29 Ab: -0.16	Non-Ab: 0.05 Ab: -0.04	Non-Ab: 0.14 Ab: -0.24	**Non-Ab: 0.38** Ab: 0.11	Non-Ab: 0.15 Ab: -0.11	Non-Ab: 0.14 Ab: -0.05	Non-Ab: 0.24 Ab: -0.44

Table 4.4: All methods r with respect to certain subsets. "WT Gly or Pro" are wild type amino acids that are either glycine or proline. "WT Non-Gly or Non-Pro" are wild type 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 have relative solvent accessibility values between 0 and 10%. "Neutral Charge" is a neutrally charged wild type amino acid mutating to a neutrally charged mutant amino acid. "Hydrophobic to Polar" is a hydrophobic or polar wild type 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 wild type amino acid. Values that are bolded are the highest r for each method and protein type. Values that are red or blue are the highest r for each subset, blue for non-Ab and red for Ab.

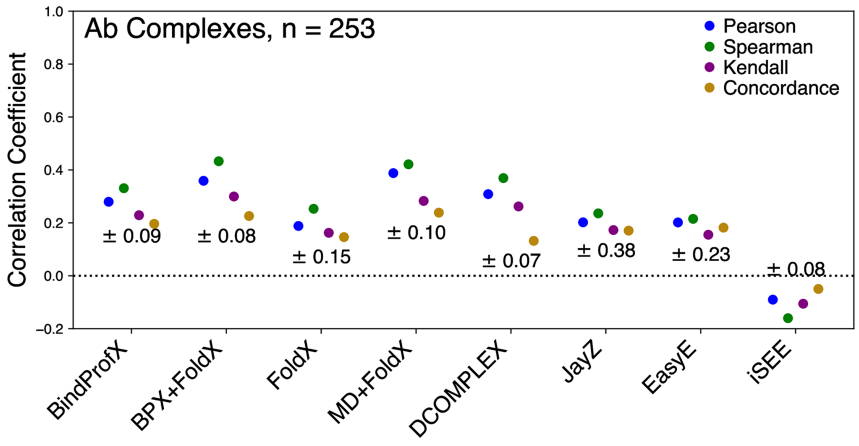

Figure 4.4: Performance of each evaluated method for Ab complexes (253 total mutations) in predicting true $\Delta\Delta G$ values (ρ_c), linearly correlated $\Delta\Delta G$ values (r), and rank order (ρ and τ). The error for each method is reported under the correlation points.

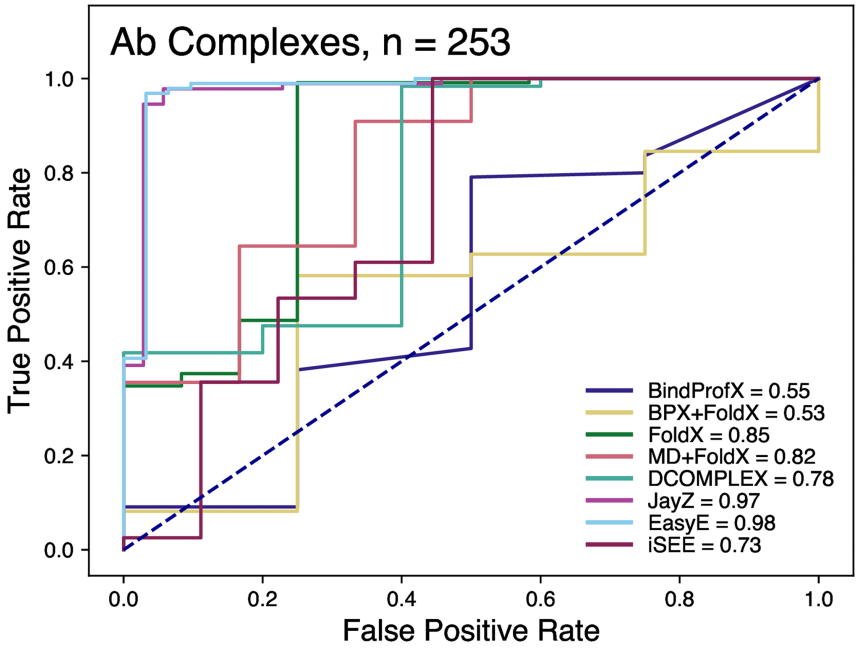

Figure 4.5: Receiver operating characteristic curves of the classification of variants that are more destabilized or less destabilized than 0.5 kcal/mol. The values in the legend represent the area-under-curve (AUC). The higher the value, the better the prediction capability of the method.

(0.82) had the highest AUC values. Combined with the results from Figures 4.1 and 4.4, at least for the systems studied here, it appears that the MD+FoldX method is the best overall performer in terms of accuracy, discriminating stabilizing mutations from destabilizing, and ranking mutations based on their experimental $\Delta\Delta G$ values.

Compared to other methods, EasyE has a much faster runtime and is recommended if the goal is to discriminate between stabilizing and destabilizing ($\Delta\Delta G$ for one mutation takes ~21 s, see Table 4.2). By comparison, MD+FoldX cost ~941 CPUh for one mutation of 1yy9. DCOMPLEX provides a slightly lower r (0.31) and computational cost (~0.35 s) for one mutation of 1yy9. Overall, MD+FoldX appears to be the best option for accuracy and EasyE or JayZ are the best options for discriminating between destabilizing and stabilizing mutations.

Table 4.4 summarizes the ability of each method to predict $\Delta\Delta G$ values for subsets of Ab mutations. Most methods had mediocre r values less than 0.60. The exceptions to this are MD+FoldX and DCOMPLEX within the WT Gly or Pro subset with $r = 0.71$ and 0.89, respectively. BPX+FoldX has the highest r for Ab complexes for five of the eight subsets (WT nonGly or nonPro, beta sheet, surface exposure, neutral charge, hydrophobic to polar, and large volume changes) and performs equally well for the neutral charge subset as DCOMPLEX, which also has the highest r for WT Gly or Pro subset. For the beta sheet subset, MD+FoldX had the second highest r. In the surface exposure subset, JayZ and EasyE

both had nearly identical r (0.36 and 0.35 respectively), the highest for this subset, but substantially worse than they did for non-Ab complexes.

4.4 Discussion

We assessed the performance of eight distinct protein-protein binding affinity calculation methods that use 3-D structural information. To test the performance of these methods, we selected 16 different protein complexes (see Table 4.1) with a total of 654 single amino acid mutations: eight antigen-antibody complexes (Ab, 253 mutations) and eight non-antigen-antibody (Non-Ab, 401 mutations) complexes. Each method was used to estimate $\Delta\Delta G$ values of the 654 mutations, a variety of statistical measures, CPU cost, and physico-chemical structural features to assess the performance. Our results suggest each method has both strengths and weaknesses depending on the properties of the protein system. Most methods did not perform well when applied to mutations in Ab complexes compared to non-Ab complexes. Rosetta-based JayZ and EasyE were able to classify mutations as destabilizing ($\Delta\Delta G$ < -0.5 kcal/mol) with high (83-98%) accuracy at relatively low computational cost. Some of the best results for Ab systems came from combining MD simulations with FoldX with a r coefficient of 0.39, but at the highest computational cost of all the tested methods.

Figure 4.1 summarizes the performance of each method in terms of its ability to estimate $\Delta\Delta G$ values for all (non-Ab + Ab) single mutations. None of the test methods show a very high r between experimental and predicted $\Delta\Delta G$ values. Two of the best performing methods, JayZ and EasyE, both have an r of 0.49 for all mutations, with a higher r of 0.61 and 0.62 respectively for non-Ab complexes. These results agree with published results from the authors of JayZ and EasyE. Our results agree moderately with published results from iSEE (they obtained $r = 0.25$, we obtained $r = 0.17$) and BindProfX (they used a much larger dataset). Published results for DCOMPLEX show a very good correlation of $r = 0.87$; much larger than what we obtained here. This difference is very likely due to the dataset size and compilation; DCOMPLEX was originally tested against 69 experimental data points, compared to the 654 values used here. MD+FoldX has an r of 0.39 for Ab complexes and appears to perform well for larger systems, which could indicate the importance of conformational sampling for antibody-antigen systems. Other methods used in this study have little to no conformational sampling which could explain their poor performance on Ab complexes. By contrast, these same methods perform well for non-Ab complexes, suggesting that conformational sampling is not the limiting factor to achieve accurate results for these protein complexes. For example, FoldX has a trained scoring function derived using a dataset of mostly non-Ab complexes and performs poorly for Ab complexes when using a single structure (see Table 4.2). However, when used with snapshots from an MD simulation, this same method

outperforms all other methods selected in this study. This highlights the need for conformational sampling for reliable and efficient predictions of binding affinity for some systems. In our previous study, we combined coarse-grained forcefield with umbrella sampling to calculate $\Delta\Delta G$ values for eight mutations of 3hfm Ab complex (one of the test systems in this study) and obtained better predictions than FoldX and MD+FoldX [95]. This study further emphasizes the need for better conformational strategies for some systems. A rigorous endpoint free energy method could potentially be employed to overcome the conformational sampling problem. Endpoint methods typically use molecular mechanics force fields to generate conformational ensembles at the two states of interest. These ensembles are then evaluated with implicit solvent models such as molecular mechanics generalized Born surface area (MM/GBSA) and molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) [220, 221, 222]. These methods are computationally less expensive than other rigorous approaches since simulations are only performed for two states, however their accuracy is system-dependent and sensitive to simulation protocols such as sampling strategy and entropy calculation. MM/PBSA and MM/GBSA have been successfully used by several groups to estimate $\Delta\Delta G$ values for a small number of protein complexes and recently reviewed by Wang E et al [221] and Wang C et al [222]. These studies obtained consistently higher overall correlation to experimental $\Delta\Delta G$ values, albeit for a small subset of mutations, compared to the methods tested in our study, but at the expense of significantly higher computational costs.

Statistical measures used to analyze performance are listed and defined in Table 4.3. For Ab, BPX+FoldX, MD+FoldX, and DCOMPLEX have the highest r values of the methods in our study (see Figure 4.4). MD+FoldX appears to be the most accurate method for Ab complexes. BindProfX, FoldX, JayZ, EasyE, and iSEE have low r and ρ_c indicating that affinities estimated using these methods do not correlate well with experimental $\Delta\Delta G$ values using a linear transformation. Also, the τ and ρ were lower compared to MD+FoldX, indicating these methods do poorly at calculating a rank order that matches experimental data.

The ROC curves allow us to determine each method's ability to classify mutations as either destabilizing or neutral/stabilizing (Figures 4.3 and 4.5). For non-Ab complexes, JayZ (0.84 AUC) and EasyE (0.83 AUC) have the best true positive rate followed by DCOMPLEX (0.82 AUC). For Ab complexes, JayZ (0.97 AUC) and EasyE (0.98 AUC) have better true positive rates than MD+FoldX, the method with the highest r value. If classification of destabilizing vs stabilizing is the primary need, then JayZ or EasyE are both recommended over the other methods tested here due to their high accuracy and fast runtime.

While accuracy is generally the main reason for choosing a particular method, computational efficiency is also an important consideration, especially when predicting the effects of a large number of mutations.

Here, we discuss the performance of each method in terms of its trade-off between speed and accuracy for predicting $\Delta\Delta G$ values. For all single mutations and our non-Ab subset, EasyE and JayZ performed well; JayZ is the faster method of the two with EasyE at a similar speed to FoldX. DCOMPLEX is more accurate than FoldX for all single mutations and has similar accuracy as FoldX for non-Ab mutations, but at much lower cost. MD+FoldX has similar accuracy to DCOMPLEX for all single mutations and has similar accuracy to FoldX in non-Ab mutations but is by far the most computationally expensive method we tested. Although a synergistic combination of BPX+FoldX implements several structural and physico-chemical interaction terms in its algorithm, computation time was longer than all but MD+FoldX without a concomitant improvement in r. We note that this method is perhaps the most accessible of those tested, due to the easy-to-use online server interface and accuracy that is similar to FoldX for most systems. 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 (not including the non-trivial preparation time). For Ab complexes, MD+FoldX, the slowest of all the methods, had the highest accuracy, followed by DCOMPLEX. iSEE is again the fastest of all methods but also the least accurate. BindProfX utilizes several pre-calculated physico-chemical structural data in its scoring function while, JayZ and EasyE each layer an additional predictive calculating feature on top of Rosetta's backbone sampling, adding complexity to the predictive algorithms. However, all three have similar r yet they do not achieve the accuracy of MD+FoldX. Overall, for non-Ab complexes, EasyE and JayZ appear to have the best balance between speed and accuracy of the methods we tested. For Ab complexes, DCOMPLEX appears to have the best balance.

We have demonstrated that all the tested methods have specific strengths and weaknesses; some perform better in specific contexts (Table 4.4), and some have longer runtimes to obtain similar predictive power to comparably faster methods. This highlights the complexity of the physico-chemical properties and structural features that drive, and limit, these predictive models. Moreover, our study highlights the need to separately evaluate the performance of future $\Delta\Delta G$ predictors for both Ab and non-Ab complexes. There is also a need for a much larger training dataset of experimentally measured binding affinities for both types of complexes. New binding affinity calculation approaches are also needed to properly account for the contribution of bridging water molecules that are often present at the protein-protein interface. Our results can be used to make informed decisions for methods that may be preferable for a particular study or system. Table 4.4 suggests that if the goal is to estimate only the order of magnitude or sign of relative binding affinities, then the preferred method will likely be very different than if the goal is to obtain the best possible accuracy for antibody-antigen systems. To improve accessibility, we have generated an

in-house Python script that can be used to parse any of the parameters used in this study and provide tailored information. This information in combination with the runtime and other details provided in this study can be used to inform users on methods that can provide the best accuracy and efficiency for a given protein-protein complex type, set of physico-chemical features or structural parameters, and set of mutations. Additionally, the script can be extended to other methods and feature-sets, potentially elucidating specific problems or areas of improvement to existing and future methods.

4.5 Conclusions

In this study, we have assessed the accuracy and efficiency of eight computational methods on predicting binding affinity changes due to single amino acid mutations. Methods were tested on 16 different protein complexes: eight antigen-antibody (Ab) and eight non-antigen-antibody (Non-Ab) complexes. While some methods perform consistently better than others, how well each performs depends on the physico-chemical and structural components of each complex. EasyE was the most accurate for non-Ab complexes, and MD+FoldX was most accurate for Ab complexes. JayZ and EasyE were better able to distinguish between destabilizing ($\Delta\Delta G >$ 0.5 kcal/mol) and stabilizing ($\Delta\Delta G <$ -0.5 kcal/mol) as compared to any other method. Future work could include more systems or different methods, including those that are solely web server-based in order to expand and better refine our conclusions on their predictive capability.

Chapter 5: Work In Progress

5.1 The Effect of Mutations on Binding Interactions Between the SARS-CoV-2 Receptor Binding Domain and Neutralizing Antibodies

Jonathan E. Barnes,[1,2] Peik K. Lund,[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

This work is currently in progress and expected to be submitted by September 2021. As first author, I contributed the following:

- Molecular dynamics simulations for all tested complexes.
- Performed FoldX analysis for relevant mutations on all complexes.
- Generated all figures.
- Currently writing manuscript.

5.1.1 Introduction

First discovered in 2019, SARS-CoV-2 is a beta-genus coronavirus responsible for COVID-19 and the current pandemic. Despite sharing similarities to SARS-CoV-1, responsible for an epidemic in 2003, the 2019 variant is far more infectious and deadly. The severity of COVID-19 necessitates an understanding of how it could evolve to escape potential treatments as well as ways to strengthen treatments against it. While there has been work devoted to understanding the impact of possible mutations in the spike S protein and its ability to bind to angiotensin-converting enzyme 2 (first step in infection process), there has not been such an effort to study interactions with antibodies. Here, we used a computational pipeline that was previously designed in our lab and applied it to the SARS-CoV-2 S protein receptor binding domain (RBD) bound to two neutralizing antibodies(Ab). Molecular dynamics simulations were used to generate trajectory snapshots. These snapshots were used as inputs for FoldX, a fast semi-empirical method for estimating folding and binding free energies. These free energy calculations were then averaged to get final estimates. We mutated sites within both Ab and the RBD that were within 10Å of the RBD-Ab binding interface. We found a large number of potential antibody escape mutations

in the RBD (i.e., those predicted to destabilize RBD-Ab interactions), some of which agree with other studies. We also found a smaller number of potential antibody strengthening mutations in Ab (i.e., those predicted to stabilize RBD-Ab interactions) that could be used to improve the therapeutic value of Ab. These results provide a basis for further studies on the effects of mutations in the RBD and antibodies and provide a starting point for building a list of potential escape mutations for antibodies.

5.1.2 Methods

To predict the effects of a large number of mutations we utilized our previously developed process [209] that consists of molecular dynamics simulations on the wildtype structure to generate conformations to then use as inputs for methods that can quickly perform mutations and predict their effect on folding stability and binding affinity. We performed 100 ns molecular dynamics simulations on the wildtype structure of two Ab-RBD complexes . We then analyzed the energy minimized structure to determine sites on the RBD that are 10 Å away from any other atom of the antibody chains and vice versa to perform a deep mutational scan. We then used the resulting snapshots as inputs for both FoldX and Rosetta. All possible mutations to the other 19 amino acids were performed on these sites and free energies of folding and binding were calculated. Since we are interested in antibody escape mutations, we evaluated them on the following criteria. A mutation was considered destabilizing if the binding affinity change was greater than 2.0 kcal/mol:

$$\Delta\Delta G_{\text{bind}} \geq 2.0 \text{ kcal/mol} \tag{5.1}$$

and functional, as-in the chain in question would fold, if the folding stability change was between -3.0 and +3.0 kcal/mol:

$$-3.0 \text{ kcal/mol} \leq \Delta\Delta G_{\text{fold}} \leq +3.0 \text{ kcal/mol} \tag{5.2}$$

The resulting sites and mutations were then compared across both methods and further evaluated.

5.1.3 Results

Our preliminary results indicate a small number of sites, for example in the case of antibody B38 (PDBID 7BZ5) there are only 8 sites (with 63 total mutations) that meet this criteria (Figure 5.1). More detailed comparison shows a small subset of mutations that meet the criteria for both FoldX and Rosetta predictions leading to three mutations at one site that are relevant for antibody CB6 (PDBID 7C01) as shown by Table 5.1.

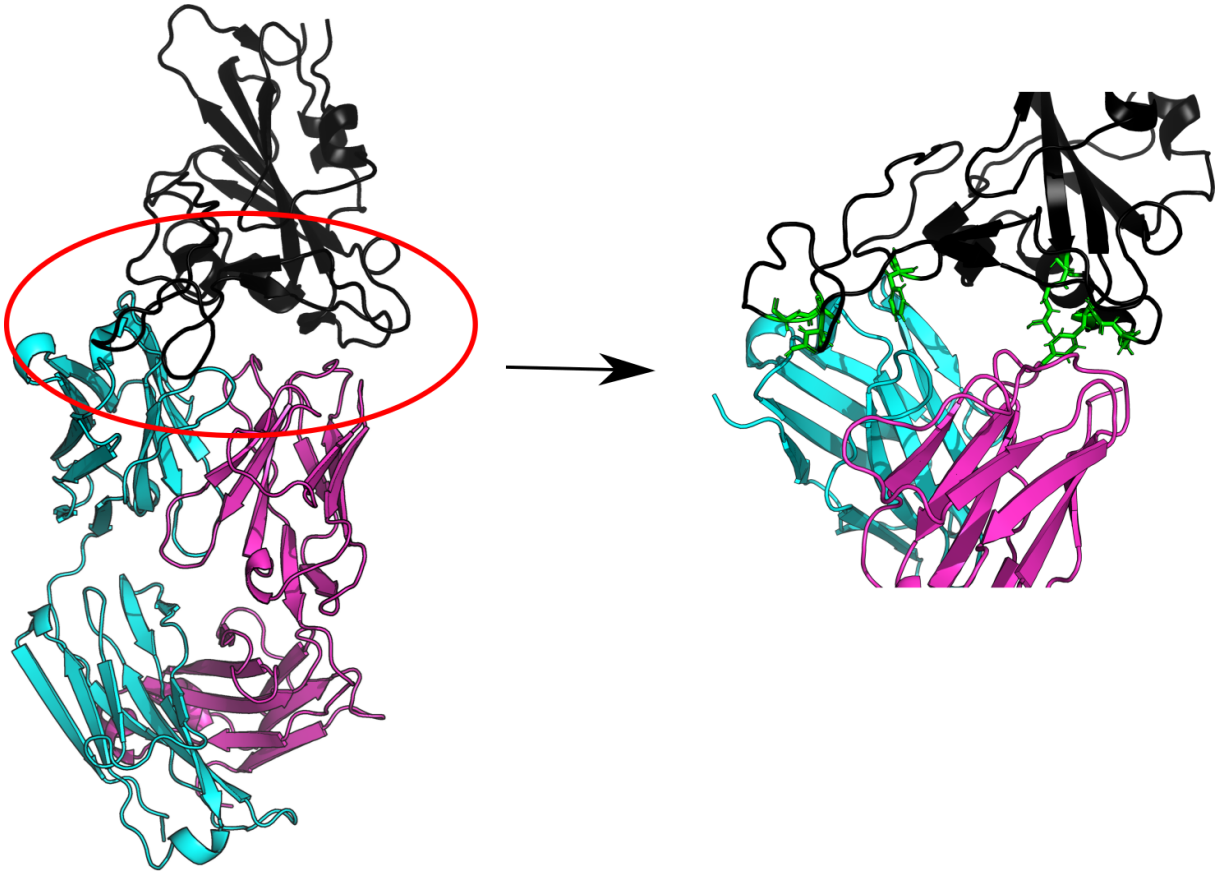

Figure 5.1: Structure of antibody B38 (green and blue chains) bound to the SARS-CoV-2-S RBD (black). Sites indicated as escape mutation sites by FoldX are indicated in green on the right. The red circle is the approximate region of interest where mutations were applied.

(B38 Ab) 7BZ5				
Mutation	**FoldX Bind**	**Rosetta Bind**	**FoldX Fold**	**Rosetta Fold**
NA501W	7.855410	3.518855	0.827161	2.300893
AA475R	5.247150	9.360632	-0.239804	2.709709
GA502P	5.081110	4.699627	-1.638115	-1.164836
AA475K	3.557506	9.958868	-0.364579	2.212432
AA475Q	3.035420	2.319998	-0.178825	2.826245
NA487H	2.953438	19.506420	0.297826	1.836539
NA487L	2.913032	34.758094	-0.286146	0.778064
NA487Q	2.832056	5.142841	0.046827	0.258826
GA502W	2.813440	4.090323	-0.115811	2.687098
NA487R	2.786487	14.451188	-0.257515	1.578110
GA502Y	2.739992	2.739401	-0.497031	0.947838
NA487E	2.714440	5.407409	0.071277	0.077108
NA487K	2.646834	17.978997	-0.308940	1.047558
YA505L	2.557022	2.119078	-0.333273	-0.943696
NA487M	2.499836	17.157780	-0.423787	2.416557
GA502H	2.274105	3.145832	0.054993	-0.309804
GA502F	2.052142	2.763998	-0.587860	1.275437
(CB6 Ab) 7C01				
Mutation	**FoldX Bind**	**Rosetta Bind**	**FoldX Fold**	**Rosetta Fold**
GA476N	3.018729	8.646407	1.105709	1.597133
GA476D	2.911374	9.527501	0.878663	-0.262864
GA476T	2.418111	10.526777	2.596795	2.288247

Table 5.1: Overlapping results for mutations in the RBD between FoldX and Rosetta. These are mutations that both methods flagged as meeting the aforementioned criteria (Equations 5.1 and 5.2)

5.1.4 Conclusions

This work builds a watchlist of potential antibody escape mutations for SARS-CoV-2 for two antibodies, with overlapping sites that could apply to other antibodies not tested here. Our results could also help inform better treatments and help design more effective antibodies for treating infection.

5.2 An EvoDevo study of visual opsin dynamics and spectral modeling in salmonids

Mariann Eilertsen[1] , Wayne I L Davies [2,3] , Dharmeshkumar Patel [4] , Jonathan E Barnes [5] , Deborah L Stenkamp [6,7], Jagdish S Patel [4,6], Jessica K Mountford [8] , Rita Karlsen [1], David W P Dolan [9] and Jon Vidar Helvik [1]

[1] Department of Biological Sciences, University of Bergen, Bergen, Norway [2] Umeå Centre for Molecular Medicine, Umeå University, Umeå, Sweden [3] School of Life Sciences, College of Science, Health and Engineering, La Trobe University, Melbourne Campus, Melbourne, Victoria, VIC 3086, Australia [4] Institute for Modeling Collaboration and Innovation (IMCI), University of Idaho, Moscow, ID, United States of America [5] Department of Physics, University of Idaho, Moscow, ID, United States of America [6] Department of Biological Sciences, University of Idaho, Moscow, ID, United States of America [7] Institute for Bioinformatics and Evolutionary Biology, University of Idaho, Moscow, ID, United States of America [8] Lions Eye Institute, University of Western Australia, Perth, Australia [9] Department of Informatics, University of Bergen, Bergen, Norway

This work is currently in progress and expected to be submitted by September 2021. As a contributing author, I performed the following:

- Statistical analysis for spectral peak prediction of two species.
- Currently assisting with manuscript preparation.

5.2.1 abstract

Salmonids are interesting models for visual neuroscientists as many species follow a distinct developmental program from demersal eggs and a large yolk sac to hatching at an advanced developmental stage. Furthermore, these economically important teleost fishes inhabit both marine and freshwater habitats, and, as such, experience diverse light environments during their prolonged life histories. At a genome level, salmonids have undergone a species-specific whole genome duplication event(i.e. Ss4R) compared to other teleosts that are themselves already far more genetically diverse compared to many non-teleost vertebrates. Thus, salmonids display phenotypically plastic visual systems that appear to be closely related to their river-marine-river migration patterns. This is most likely due to a complex interplay between their larger, more gene-rich genomes and broad spectrally-enriched habitats; however, the molecular basis (and functional consequences) for such diversity is not fully understood. This detailed study used recent

genome sequencing advances to identify the complete repertoire and genome organization of visual opsin genes from seven different salmonid species, namely the Atlantic salmon (*Salmo salar*), the Arctic char (*Salvelinus alpinus*), the brown trout (*Salmo trutta*), the Chinook salmon (*Oncorhynchus tshawytcha*), the Coho salmon (*Oncorhynchus kisutch*), the rainbow trout (Oncorhynchus mykiss) and the sockeye salmon (Oncorhynchus nerka), compared to those of the northern pike (*Esox lucius*), a closely-related non-salmonid species. Results showed that several opsin genes were not retained after the Ss4R genome duplication event, which is consistent with the concept of salmonid rediploidization. Developmentally, in-depth transcriptomic analyses of S. salar revealed differential expression within each opsin class, with two of the long-wavelength-sensitive (*lws*) opsins not being expressed before the start of feeding. Also, early opsin expression in the retina was located centrally, expanding dorsally and ventrally as eye development progressed, with rod opsin (*rh1*) being the dominant visual opsin post-hatching. Of the visual photopigment genes that are conserved across salmonids, molecular modeling predicted the greatest variation in spectral sensitivity to be within the rh2 class, with a 40 nm difference in the λ_{max} values between these four medium-wavelength-sensitive photopigments. Overall, it appears that opsin duplication and expression, and their respective spectral tuning profiles, evolved to maximize specialist color vision throughout an anadromous lifecycle, with some visual opsin genes being lost to tailor marine-based vision.

Chapter 6: Conclusion

6.1 Summary

In our study on epistasis (Chapter 2), we showed that statistical modeling can be used to detemrine mechanisms involved in protein interactions. We demonstrated that while we are working with small datasets and simple static features we can still show with statistical likelihood that some biophysical features are more important, and more significant contributors to the phenomenon than others.

In our study on Short Wavelength Sensitive 2 (*Sws2*) opsins (Chapter 3), we showed that molecular modeling can be used to as a reference to build statistical models to predict phenotypes, in this case color vision, with high accuracy. Molecular dynamics were performed on the dark state of the opsin proteins for 11 species of opsin in deep sea fish. The resulting trajectories were used to inform linear statistical models with a resulting 3-term model which could predict the spectral sensitivity of *Sws2* opsins with an R^2 of 0.95.

In our study of binding affinity software (see Chapter 4), we showed the accuracy of eight different methods at predicting the effects of mutations on the binding affinity for 16 protein-protein complexes. While we don't have a be-all end-all best method our results indicate the some methods are better than others for given contexts and demonstrates there are still gaps in our ability to predict mutational affects.

This body of work is built on the common theme of using modeling, both statistical and molecular, to map protein genotype to resulting phenotypes. In the case of epistasis (Chapter 2), we studied how physical attributes of the protein can explain the non-additive nature of multiple missense mutations on protein folding and binding. In the case of SWS2 opsins (Chapter 3), we used aspects of protein and chromophore structure to predict the color vision sensitivity phenotype. For the binding affinity methods analysis of Chapter 4, we investigated the performance of fast methods for predicting the effects of protein genotype on their binding affinity phenotype, determining the contexts they each perform best.

6.2 Future Research

Our study on epistasis (Chapter 2) demonstrated that we can explain some (∼30%) of the observed epistasis in the available binding and folding data. The larger question is what could account for the remaining 70%? Using molecular dynamics and more broad statistical methods could add in dynamical properties and more complex paramaters to attempt at building a more complete picture of epistasis.

Future work for the Sws2 Opsin project (Chapter 3) could include building models for other classes of opsins to further prove the capabilities of the approach. Additionally, with a complete set of structures

and models for the full gamut of opsin proteins could allow for the development of a singular model or mapping to go directly from sequence to spectral sensitivity, negating the need for the molecular dynamics step for novel opsins.

Our work on binding affinity software (Chapter 4) can be readily expanded on. More methods could be tested to further build a more complete library of methods and their strengths and weaknesses. More data can be used to improve on our existing recommendations. Additionally, by investigating the contexts no method performed well at making predictions better models could potentially be developed.

References

[1] J.-J. Lin, F.-Y. Wang, W.-H. Li, and T.-Y. Wang. The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. *Scientific Reports*, 7:15568, 2017.

[2] S. Kasagi, K. Mizusawa, and A. Takahashi. Green-shifting of sws2a opsin sensitivity and loss of function of rh2-a opsin in flounders, genus verasper. *Ecology and Evolution*, 8:1399–1410, 2018.

[3] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002.

[4] Jason H Moore. A global view of epistasis. *Nature Genetics*, 37(1):13–14, January 2005.

[5] Trudy FC Mackay and Jason H Moore. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(6):125, 2014.

[6] R. Sanjuan and S. F. Elena. Epistasis correlates to genomic complexity. *Proceedings of the National Academy of Sciences*, 103(39):14402–14405, September 2006.

[7] Michael C. Whitlock, Patrick C. Phillips, Francisco B.-G. Moore, and Stephen J. Tonsor. Multiple fitness peaks and epistasis. *Annual Review of Ecology and Systematics*, 26(1):601–629, 1995.

[8] C. Natarajan, N. Inoguchi, R. E. Weber, A. Fago, H. Moriyama, and J. F. Storz. Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. *Science*, 340(6138):1324–1327, June 2013.

[9] Merijn L. M. Salverda, Eynat Dellus, Florien A. Gorter, Alfons J. M. Debets, John van der Oost, Rolf F. Hoekstra, Dan S. Tawfik, and J. Arjan G. M. de Visser. Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLoS Genetics*, 7(3):e1001321, March 2011.

[10] Jeremy A. Draghi and Joshua B. Plotkin. SELECTION BIASES THE PREVALENCE AND TYPE OF EPISTASIS ALONG ADAPTIVE TRAJECTORIES: SELECTION BIASES EPISTASIS ALONG ADAPTIVE TRAJECTORIES. *Evolution*, 67(11):3120–3131, November 2013.

[11] Roger D. Kouyos, Olin K. Silander, and Sebastian Bonhoeffer. Epistasis between deleterious mutations and the evolution of recombination. *Trends in Ecology & Evolution*, 22(6):308–315, June 2007.

[12] Darin R. Rokyta, Paul Joyce, S. Brian Caudle, Craig Miller, Craig J. Beisel, and Holly A. Wichman. Epistasis between Beneficial Mutations and the Phenotype-to-Fitness Map for a ssDNA Virus. *PLoS Genetics*, 7(6):e1002075, June 2011.

[13] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, May 2013.

[14] J. D. Bloom, L. I. Gong, and D. Baltimore. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science*, 328(5983):1272–1275, June 2010.

[15] E. R. Lozovsky, T. Chookajorn, K. M. Brown, M. Imwong, P. J. Shaw, S. Kamchonwongpaisan, D. E. Neafsey, D. M. Weinreich, and D. L. Hartl. Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proceedings of the National Academy of Sciences*, 106(29):12025–12030, July 2009.

[16] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*, 461(7263):515–519, September 2009.

[17] E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, and J. W. Thornton. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science*, 317(5844):1544–1548, September 2007.

[18] D. M. Weinreich. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312(5770):111–114, April 2006.

[19] Michael S. Breen, Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, October 2012.

[20] Daniel J. Kvitek and Gavin Sherlock. Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. *PLoS Genetics*, 7(4):e1002056, April 2011.

[21] H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segre, and C. J. Marx. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*, 332(6034):1190–1192, June 2011.

[22] Xinzhu Wei and Jianzhi Zhang. Patterns and Mechanisms of Diminishing Returns from Beneficial Mutations. *Molecular Biology and Evolution*, 36(5):1008–1021, May 2019.

[23] S. Kryazhimskiy, D. P. Rice, E. R. Jerison, and M. M. Desai. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522, June 2014.

[24] A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper. Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science*, 332(6034):1193–1196, June 2011.

[25] Beth Shapiro, Andrew Rambaut, Oliver G. Pybus, and Edward C. Holmes. A Phylogenetic Method for Detecting Positive Epistasis in Gene Sequences and Its Application to RNA Virus Evolution. *Molecular Biology and Evolution*, 23(9):1724–1730, September 2006.

[26] Rafael Sanjuán, José M. Cuevas, Andrés Moya, and Santiago F. Elena. Epistasis and the Adaptability of an RNA Virus. *Genetics*, 170(3):1001–1008, July 2005.

[27] Christina L. Burch and Lin Chao. Epistasis and Its Relationship to Canalization in the RNA Virus 6. *Genetics*, 167(2):559–567, June 2004.

[28] Y. Michalakis. EVOLUTION: Epistasis in RNA Viruses. *Science*, 306(5701):1492–1493, November 2004.

[29] Jack da Silva, Mia Coetzer, Rebecca Nedellec, Cristina Pastore, and Donald E. Mosier. Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics*, 185(1):293–303, May 2010.

[30] R. Sanjuan, A. Moya, and S. F. Elena. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proceedings of the National Academy of Sciences*, 101(43):15376–15379, October 2004.

[31] S. Bonhoeffer. Evidence for Positive Epistasis in HIV-1. *Science*, 306(5701):1547–1550, November 2004.

[32] Sandra Trindade, Ana Sousa, Karina Bivar Xavier, Francisco Dionisio, Miguel Godinho Ferreira, and Isabel Gordo. Positive Epistasis Drives the Acquisition of Multidrug Resistance. *PLoS Genetics*, 5(7):e1000578, July 2009.

[33] Jason H. Moore. The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. *Human Heredity*, 56(1-3):73–82, 2003.

[34] Liskin Swint-Kruse. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophysical Journal*, 111(1):10–18, July 2016.

[35] C. Melero, N. Ollikainen, I. Harwood, J. Karpiak, and T. Kortemme. Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. *Proceedings of the National Academy of Sciences*, 111(43):15426–15431, October 2014.

[36] Charlotte M. Miton and Nobuhiko Tokuriki. How mutational epistasis impairs predictability in protein evolution and design: How Epistasis Impairs Predictability in Enzyme Evolution. *Protein Science*, 25(7):1260–1272, July 2016.

[37] Manfred T. Reetz. The Importance of Additive and Non-Additive Mutational Effects in Protein Engineering. *Angewandte Chemie International Edition*, 52(10):2658–2666, March 2013.

[38] James A. Wells. Additivity of mutational effects in proteins. *Biochemistry*, 29(37):8509–8517, September 1990.

[39] Eynat Dellus-Gur, Mikael Elias, Emilia Caselli, Fabio Prati, Merijn L.M. Salverda, J. Arjan G.M. de Visser, James S. Fraser, and Dan S. Tawfik. Negative Epistasis and Evolvability in TEM-1 β-Lactamase—The Thin Line between an Enzyme's Conformational Freedom and Disorder. *Journal of Molecular Biology*, 427(14):2396–2409, July 2015.

[40] Courtney E. Gonzalez and Marc Ostermeier. Pervasive Pairwise Intragenic Epistasis among Sequential Mutations in TEM-1 β-Lactamase. *Journal of Molecular Biology*, 431(10):1981–1992, May 2019.

[41] C. Anders Olson, Nicholas C. Wu, and Ren Sun. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22):2643–2651, November 2014.

[42] Andrei Y. Istomin, M. Michael Gromiha, Oleg K. Vorov, Donald J. Jacobs, and Dennis R. Livesay. New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins: Structure, Function, and Bioinformatics*, 70(3):915–924, February 2008.

[43] Haoran Yu and Paul A. Dalby. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proceedings of the National Academy of Sciences*, 115(47):E11043–E11052, November 2018.

[44] Sherlyn Jemimah and M. Michael Gromiha. Exploring additivity effects of double mutations on the binding affinity of protein-protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 86(5):536–547, May 2018.

[45] Justina Jankauskaite, 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*, 3:462–469, February 2019.

[46] Akinori Sarai, Hatsuho Uedaira, K. Abdulla Bava, Koji Kitajima, and M. Michael Gromiha. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research*, 32(suppl_1):D120–D121, January 2004.

[47] H.M Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Databank. *Nucliec Acids Research*, 28:235–242, 2000.

[48] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.

[49] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017.

[50] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl_2):W382–W388, July 2005.

[51] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[52] 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), November 2013.

[53] Kenneth P. Burnham, David Raymond Anderson, and Kenneth P. Burnham. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, New York, 2nd ed edition, 2002. OCLC: ocm48557578.

[54] R Core Team. R: A Language and Environment for Statistical Computing. 2020.

[55] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002.

[56] Marc J. Mazerolle. AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c), 2020. R package version 2.3-1.

[57] Sophie E Jacksod and Alan R Fersht. Contribution of Residues in the Reactive Site Loop of Chymotrypsin Inhibitor 2 to Protein Stability and Activityt. page 8.

[58] Jaume Pons, Arvind Rajpal, and Jack F. Kirsch. Energetic analysis of an antigen/antibody interface: Alanine scanning mutagenesis and double mutant cycles on the hyhel-10/lysozyme interaction. *Protein Science*, 8(5):958–968, 1999.

[59] Yili Li, Hongmin Li, Sandra J. Smith-Gill, and Roy A. Mariuzza. Three-Dimensional Structures of the Free and Antigen-Bound Fab from Monoclonal Antilysozyme Antibody HyHEL-63 † , ‡. *Biochemistry*, 39(21):6296–6309, May 2000.

[60] Gideon Schreiber and Alan R. Fersht. Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology*, 248(2):478–486, January 1995.

[61] Ellen R. Goldman, William Dall'Acqua, Bradford C. Braden, and Roy A. Mariuzza. Analysis of Binding Interactions in an Idiotope-Antiidiotope Protein-Protein Complex by Double Mutant Cycles †. *Biochemistry*, 36(1):49–56, January 1997.

[62] Gary J. Pielak and Xuming Wang. Interactions between Yeast Iso-1-cytochrome *c* and Its Peroxidase †. *Biochemistry*, 40(2):422–428, January 2001.

[63] Liang-Tsung Huang and M. Michael Gromiha. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. *Bioinformatics*, 25(17):2181–2187, September 2009.

[64] S. Yokoyama, N.S. Blow, and F.B. Radlwimmer. Molecular evolution of color vision of zebra finch. *Gene*, 259:17–24, 2000.

[65] W.I.L. Davies, S.E. Wilkie, J.A. Cowing, M.W. Hankins, and D.M. Hunt. Anion sensitivity and spectral tuning of middle- and long-wavelength-sensitive (MWS/LWS) visual pigments. *Cell Mol Life Sci*, 69:2455–2464, 2012.

[66] J.K. Bowmaker. Evolution of vertebrate visual pigments. *Vision Research*, 48:2022–2041, 2008.

[67] F.I. Hárosi. An analysis of two spectral properties of vertebrate visual pigments. *Vision Research*, 34:1359–1367, 1994.

[68] J.W.L. Parry and J.K. Bowmaker. Visual pigment reconstitution in intact goldfish retina using synthetic retinaldehyde isomers. *Vision Research*, 40:2241–2247, 2000.

[69] W. Wang, J.H. Geiger, and B. Borhan. The photochemical determinants of color vision. *BioEssays*, 36:65–74, 2014.

[70] J.M. Enright, M.B. Toomey, S. Sato, S.E. Temple, Fujiwara Allen, JR, and R. Cyp27c1 red-shifts the spectral sensitivity of photoreceptors by converting vitamin a1 into a2. *Current Biology*, 25:3048–3057, 2015.

[71] D Lagman, D Ocampo Daza, J Widmark, XM Abalo, G Sundström, and D Larhammar. The vertebrate ancestral repertoire of visual opsins, transducin alpha subunits and oxytocin/vasopressin receptors was established by duplication of their shared genomic region in the two rounds of early vertebrate genome duplications. *BMC Evolutionary Biology*, 13:238, 2013.

[72] S. Yokoyama, T. Tada, H. Zhang, and L. Britt. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. In *Proceedings of the National Academy of Sciences*, volume 105, page 13480–13485, 2008.

[73] Z. Musilova, F. Cortesi, M. Matschiner, W.I.L. Davies, J.S. Patel, and S.M. Stieb. Vision using multiple distinct rod opsins in deep-sea fishes. *Science*, 364:588–592, 2019.

[74] D.M. Hunt, K.S. Dulai, J.C. Partridge, P. Cottrill, and J.K. Bowmaker. The molecular basis for spectral tuning of rod visual pigments in deep-sea fish. *J Exp Biol*, 204:3333–3344, 2001.

[75] J.N. Lythgoe. *The Ecology of Vision*. Clarendon Press, 1979.

[76] G. Wald. The porphyropsin visual system. *J Gen Physiol*, 22:775–794, 1939.

[77] A.V. Whitmore and J.K. Bowmaker. Seasonal variation in cone sensitivity and short-wave absorbing visual pigments in the rudd scardinius erythrophthalmus. *J Comp Physiol A*, 166:103–115, 1989.

[78] D.A. Marques, J.S. Taylor, F.C. Jones, F.D. Palma, D.M. Kingsley, and T.E. Reimchen. Convergent evolution of sws2 opsin facilitates adaptive radiation of threespine stickleback into different light environments. *PLOS Biology*, 15:2001627, 2017.

[79] N.I. Bloch. Evolution of opsin expression in birds driven by sexual selection and habitat. In *Proceedings of the Royal Society B: Biological Sciences*, volume 282, page 20142321, 2015.

[80] Meyer A. Härer A, Torres-Dowdall J. Rapid adaptation to a novel light environment: The importance of ontogeny and phenotypic plasticity in shaping the visual system of nicaraguan midas cichlid fish (amphilophus citrinellus spp. *Molecular Ecology*, 26:5582–5593, 2017.

[81] A. Chinen, T. Hamaoka, Y. Yamada, and S. Kawamura. Gene duplication and spectral diversification of cone visual pigments of zebrafish. *Genetics*, 163:663–675, 2003.

[82] Y. Matsumoto, S. Fukamachi, H. Mitani, and S. Kawamura. Functional characterization of visual opsin repertoire in medaka (oryzias latipes. *Gene*, 371:268–278, 2006.

[83] C.M. Hofmann and K.L. Carleton. Gene duplication and differential gene expression play an important role in the diversification of visual pigments in fish. *Integr Comp Biol*, 49:630–643, 2009.

[84] F Cortesi, Z Musilová, SM Stieb, NS Hart, UE Siebeck, and et al Malmstrøm, M. Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *Proc Natl Acad Sci USA*, 112:1493–1498, 2015.

[85] Lychakov Govardovskii, VI and D.V. Some quantitative characteristics of the frog retinal rod outer segments. *Tsitologiia*, 27:524–529, 1975.

[86] A. Chinen, Y. Matsumoto, and S. Kawamura. Reconstitution of ancestral green visual pigments of zebrafish and molecular mechanism of their spectral differentiation. *Mol Biol Evol*, 22:1001–1010, 2005.

[87] J.A. Cowing, S. Poopalasundaram, S.E. Wilkie, P.R. Robinson, J.K. Bowmaker, and D.M. Hunt. The molecular mechanism for the spectral shifts between vertebrate ultraviolet- and violet-sensitive cone visual pigments. *Biochem J*, 367:129–135, 2002.

[88] Y. Matsumoto, C. Hiramatsu, Y. Matsushita, N. Ozawa, R. Ashino, and M. Nakata. Evolutionary renovation of L/M opsin polymorphism confers a fruit discrimination advantage to ateline new world monkeys. *Molecular Ecology*, 23:1799–1812, 2014.

[89] W.L. Davies, J.A. Cowing, J.K. Bowmaker, L.S. Carvalho, D.J. Gower, and D.M. Hunt. Shedding light on serpent sight: The visual pigments of henophidian snakes. *J Neurosci*, 29:7519–7525, 2009.

[90] W.L. Davies, L.S. Carvalho, J.A. Cowing, L.D. Beazley, D.M. Hunt, and C.A. Arrese. Visual pigments of the platypus: A novel route to mammalian colour vision. *Current Biology*, 17:161– 163, 2007.

[91] W.L. Davies, L.S. Carvalho, B.-H. Tay, S. Brenner, D.M. Hunt, and B. Venkatesh. Into the blue: Gene duplication and loss underlie color vision adaptations in a deep-sea chimaera, the elephant shark callorhinchus milii. *Genome Res*, 2009-07-16.

[92] W.L. Davies, J.A. Cowing, L.S. Carvalho, I.C. Potter, A.E.O. Trezise, and D.M. Hunt. Functional characterization, tuning, and regulation of visual pigment gene expression in an anadromous lamprey. *The FASEB Journal*, 21:2713–2724, 2007.

[93] S. Yokoyama and F.B. Radlwimmer. The molecular genetics and evolution of red and green color vision in vertebrates. *Genetics*, 158:1697–1710, 2001.

[94] S. Yokoyama, F.B. Radlwimmer, and S. Kawamura. Regeneration of ultraviolet pigments of vertebrates. *FEBS Lett*, 423:155–158, 1998.

[95] J.S. Patel, C.J. Brown, F.M. Ytreberg, and D.L. Stenkamp. Predicting peak spectral sensitivities of vertebrate cone visual pigments using atomistic molecular simulations. *PLOS Computational Biology*, 14:1005974, 2018.

[96] J. Rajput, D.B. Rahbek, L.H. Andersen, A. Hirshfeld, M. Sheves, and P. Altoè. Probing and modeling the absorption of retinal protein chromophores in vacuo. *Angewandte Chemie International Edition*, 49:1790–1793, 2010.

[97] S. Sekharan, M. Sugihara, and V. Buss. Origin of spectral tuning in rhodopsin—it is not the binding pocket. *Angewandte Chemie International Edition*, 46:269–271, 2007.

[98] E. Kloppmann, T. Becker, and G.M. Ullmann. Electrostatic potential at the retinal of three archaeal rhodopsins: Implications for their different absorption spectra. *Proteins: Structure, Function, and Bioinformatics*, 61:953–965, 2005.

[99] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, and B.A. Fox. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science*, 289:739–745, 2000.

[100] S. Kawamura, S. Kasagi, D. Kasai, A. Tezuka, A. Shoji, and A. Takahashi. Spectral sensitivity of guppy visual pigments reconstituted in vitro to resolve association of opsins with cone cell types. *Vision Research*, 127:67–73, 2016.

[101] S. Yokoyama, N. Takenaka, and N. Blow. A novel spectral tuning in the short wavelength-sensitive (sws1 and sws2) pigments of bluefin killifish (lucania goodei. *Gene*, 396:196–202, 2007.

[102] T. Okada, M. Sugihara, A.-N. Bondar, M. Elstner, P. Entel, and V. Buss. The retinal conformation and its environment in rhodopsin in light of a new 2.2Å crystal structure††this paper is dedicated to dr yoshimasa kyogoku. *Journal of Molecular Biology*, 342:571–583, 2004.

[103] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, 9:646–652, 2002.

[104] Y. Shichida and T. Matsuyama. Evolution of opsins and phototransduction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:2881–2895, 2009.

[105] S.P. Collin, N.S. Hart, J. Shand, and I.C. Potter. Morphology and spectral absorption characteristics of retinal photoreceptors in the southern hemisphere lamprey (geotria australis. *Vis Neurosci*, 20:119–130, 2003.

[106] F.I. Hárosi and J. Kleinschmidt. Visual pigments in the sea lamprey, petromyzon marinus. *Visual Neuroscience*, 10:711–715, 1993.

[107] M. Toyama, M. Hironaka, Y. Yamahama, H. Horiguchi, O. Tsukada, and N. Uto. Presence of rhodopsin and porphyropsin in the eyes of 164 fishes, representing marine, diadromous, coastal and freshwater species–a qualitative and comparative study. *Photochem Photobiol*, 84:996–1002, 2008.

[108] Morphology, characterization, and distribution of retinal photoreceptors in the australian lungfish neoceratodus forsteri (krefft, 1870) - bailes - 2006. *Journal of Comparative Neurology - Wiley Online Library*, 2020-07-19. Available:.

[109] H.J. Bailes, W.L. Davies, A.E. Trezise, and S.P. Collin. Visual pigments in a living fossil, the australian lungfish neoceratodus forsteri. *BMC Evolutionary Biology*, 7:200, 2007.

[110] I. Provencio, E.R. Loew, and R.G. Foster. Vitamin a2-based visual pigments in fully terrestrial vertebrates. *Vision Res*, 32:2201–2208, 1992.

[111] E.R. Loew, L.J. Fleishman, R.G. Foster, and I. Provencio. Visual pigments and oil droplets in diurnal lizards: a comparative study of caribbean anoles. *Journal of Experimental Biology*, 205:927–938, 2002.

[112] C. Katti, M. Stacey-Solis, N.A. Coronel-Rojas, and W.I.L. Davies. The diversity and adaptive evolution of visual photopigments in reptiles. *Front Ecol Evol*, 7, 2019.

[113] E.R. Loew and H.J.A. Dartnall. Vitamin a1/a2-based visual pigment mixtures in cones of the rudd. *Vision Research*, 16:891–896, 1976.

[114] F.I. Harosi. Ultraviolet-and violet-absorbing vertebrate visual pigments: dichroic and bleaching properties. *The visual system*, 1985:41–55.

[115] Ramkumar Rajamani, Yen-lin Lin, and Jiali Gao. The Opsin Shift and Mechanism of Spectral Tuning of Rhodopsin. *Journal of computational chemistry*, 32(5):854–865, April 2011.

[116] J. Huerta-Cepas, F. Serra, and P. Bork. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*, 33:1635–1638, 2016.

[117] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Syst Biol*, 59:307–321, 2010.

[118] M.P. Jacobson, D.L. Pincus, C.S. Rapp, T.J.F. Day, B. Honig, and D.E. Shaw. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55:351–367, 2004.

[119] M.P. Jacobson, R.A. Friesner, Z. Xiang, and B. Honig. On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology*, 320:597–608, 2002.

[120] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52:7182–7190, 1981.

[121] D.J. Evans and B.L. Holian. The nose–hoover thermostat. *The Journal of Chemical Physics*, 83:4069, 1998.

[122] H. Berendsen, D. Spoel, E. Lindahl, B. Hess, G. Groenhof, and A. Mark. Gromacs: fast, flexible, and free. *J Comput Chem*, 26:1701–1718, 2005.

[123] W. Humphrey, A. Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[124] G. Schwarz. Estimating the dimension of a model. *Ann Statist*, 6:461–464, 1978.

[125] S. Jones and J.M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[126] C.M. Yates and M.J.E. Sternberg. The effects of non-synonymous single nucleotide polymorphisms (nssnps) on protein–protein interactions. *Journal of Molecular Biology*, 425(21):3949–63, 2013.

[127] M. Baaden and S.J. Marrink. Coarse-grain modelling of protein-protein interactions. *Curr Opin Struct Biol*, 23(6):24172141, 2013.

[128] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M.L. Tress. Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics*, 10(3):233–46, 2009. PubMed PMID: 19346321.

[129] 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):2216–25, 2010. PubMed PMID: 20329755.

[130] B.M. Kroncke, A.M. Duran, J.L. Mendenhall, J. Meiler, J.D. Blume, and C.R. Sanders. Documentation of an imperative to improve methods for predicting membrane protein stability. *Biochemistry*, 55(36):5002–9, 2016.

[131] R.C. Bernardi, M.C. Melo, and K. Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta*, 1850(5):872–7, 2015.

[132] V. Spiwok, Z. Sucur, and P. Hosek. Enhanced sampling techniques in biomolecular simulations. *Biotechnol Adv*, 33(6, 2015. Pt 2):1130-40. PubMed PMID: 25482668.

[133] 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):3789–98, 2013.

[134] P. Pokorná, H. Kruse, M. Krepl, and J. Šponer. QM/MM calculations on protein–rna complexes: Understanding limitations of classical md simulations and search for reliable cost-effective qm methods. *J Chem Theory Comput*, 14(10):5419–33, 2018.

[135] C. Geng, L.C. Xue, J. Roel-Touris, and A.M.J.J. Bonvin. Finding the $\delta\delta$g spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Computational Molecular Science*, 9(5), 2019.

[136] M.M. Gromiha, K. Yugandhar, and S. Jemimah. Protein-protein interactions: scoring schemes and binding affinity. *Curr Opin Struct Biol*, 44(31-8):27866112, 2016.

[137] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The foldx web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue), 2005.

[138] S. Liu, C. Zhang, H. Zhou, and Y. 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, 2004.

[139] P. Xiong, C. Zhang, W. Zheng, and Y. Zhang. Bindprofx: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J Mol Biol*, 429(3):426–34, 2017.

[140] Zhang Brender, JR and Y. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS Comp Biol*, 11(10), 2015.

[141] Y. Dehouck, J.M. Kwasigroch, M. Rooman, and Gilis D. BeAtMuSiC. Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res*, 41(Web Server issue), 2013.

[142] M. Li, F.L. Simonetti, A. Goncearenco, and A.R. Panchenko. Mutabind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic Acids Res*, 44(W1), 2016.

[143] T. Vreven, H. Hwang, B.G. Pierce, and Z. Weng. Prediction of protein–protein binding free energies. *Protein Sci*, 21(3):396–404, 2012.

[144] C.H.M. Rodrigues, Y. Myung, D.E.V. Pires, and D.B. Ascher. mcsm-ppi2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res*, 47(W1), 2019.

[145] S. Jemimah, M. Sekijima, and M.M. Gromiha. Proaffimuseq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. *Bioinformatics*, 36(6):1725–30, 2019.

[146] J. Jankauskaitė, B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio, and I.H. Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–9, 2018.

[147] A.C. Papageorgiou, R. Shapiro, and K.R. 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):5162–77, 1997.

[148] C.Z. Chen and R. Shapiro. Superadditive and subadditive effects of "hot spot" mutations within the interfaces of placental ribonuclease inhibitor with angiogenin and ribonuclease a. *Biochemistry*, 38(29):10413501, 1999.

[149] R. Shapiro, M. Ruiz-Gutierrez, and C.Z. Chen. Analysis of the interactions of human ribonuclease inhibitor with angiogenin and ribonuclease a by mutagenesis: importance of inhibitor residues inside versus outside the c-terminal "hot spot". *J Mol Biol*, 302(2):497–519, 2000. PubMed PMID: 10970748.

[150] R. Shapiro and B.L. Vallee. Identification of functional arginines in human angiogenin by site-directed mutagenesis. *Biochemistry*, 31(49):1281426, 1992.

[151] R. Shapiro and B.L. Vallee. Site-directed mutagenesis of histidine-13 and histidine-114 of human angiogenin. alanine derivatives inhibit angiogenin-induced angiogenesis. *Biochemistry*, 28(18):7401–8, 1989. PubMed PMID: 2479414.

[152] F.S. Lee and B.L. Vallee. Binding of placental ribonuclease inhibitor to the active site of angiogenin. *Biochemistry*, 28(8):2742853, 1989.

[153] C.Z. Chen and R. Shapiro. Site-specific mutagenesis reveals differences in the structural bases for tight binding of rnase inhibitor to angiogenin and rnase a. *Proc Natl Acad Sci U S A*, 94(5):1761–6, 1997.

[154] Y.A. Muller, Y. Chen, H.W. Christinger, B. Li, B.C. Cunningham, and H.B. Lowman. Vegf and the fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 a resolution and mutational analysis of the interface. *Structure*, 6(9):1153–67, 1998. PubMed PMID: 9753694.

[155] Y. Chen, C. Wiesmann, G. Fuh, B. Li, H.W. Christinger, and P. McKay. Selection and analysis of an optimized anti-vegf antibody: crystal structure of an affinity-matured fab in complex with antigen. *J Mol Biol*, 293(4):865–81, 1999. PubMed PMID: 10543973.

[156] Y.A. Muller, Y. Chen, H.W. Christinger, B. Li, B.C. Cunningham, and H.B. Lowman. Vegf and the fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 xe5; resolution and mutational analysis of the interface. *Structure*, 6(9):1153–67, 1998.

[157] 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):8878–89, 1994. PubMed PMID: 8043575.

[158] R.W. Hartley. Directed mutagenesis and barnase-barstar recognition. *Biochemistry*, 32(23):5978–84, 1993.

[159] G. Schreiber and A.R. Fersht. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J Mol Biol*, 248(2):478–86, 1995. PubMed PMID: 7739054.

[160] G. Schreiber and A.R. Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry*, 32(19):5145–50, 1993. PubMed PMID: 8494892.

[161] C. Frisch, G. Schreiber, C.M. Johnson, and A.R. Fersht. Thermodynamics of the interaction of barnase and barstar: changes in free energy versus changes in enthalpy on mutation. *J Mol Biol*, 267(3):696–706, 1997. PubMed PMID: 9126847.

[162] S. Sogabe, F. Stuart, C. Henke, A. Bridges, G. Williams, and A. Birch. 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. J Mol Biol*, 273(4):882–97, 1997.

[163] S. Lang, J. Xu, F. Stuart, R.M. Thomas, J.W. Vrijbloed, and J.A. Robinson. Analysis of antibody a6 binding to the extracellular interferon gamma receptor alpha-chain by alanine-scanning mutagenesis and random mutagenesis with phage display. *Biochemistry*, 39(51):11123892, 2000.

[164] S. Sogabe, F. Stuart, C. Henke, A. Bridges, G. Williams, and A. Birch. Neutralizing epitopes on the extracellular interferon gamma receptor (ifngammar) alpha-chain characterized by homolog scanning mutagenesis and x-ray crystal structure of the a6 fab-ifngammar1-108 complex. *J Mol Biol*, 273(4):882–97, 1997. PubMed PMID: 9367779.

[165] A.J. Scheidig, T.R. Hynes, L.A. Pelletier, J.A. Wells, and A.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 Sci*, 6(9):1806–24, 1997.

[166] D. Krowarsch, M. Dadlez, O. Buczek, I. Krokoszynska, A.O. Smalas, and J. Otlewski. Interscaffolding additivity: binding of p1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *J Mol Biol*, 289(1):10339415, 1999.

[167] M.J. Castro and S. Anderson. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to beta-trypsin and alpha-chymotrypsin. *Biochemistry*, 35(35):8784199, 1996.

[168] B.C. Braden, H. Souchon, J.-L. Eiselé, G.A. Bentley, T.N. Bhat, and J. Navaza. Three-dimensional structures of the free and the antigen-complexed fab from monoclonal anti-lysozyme antibody d44.1. *J Mol Biol*, 243(4):767–81, 1994.

[169] S.M. Lippow, K.D. Wittrup, and B. Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol*, 25(10):1171–6, 2007.

[170] T. Hage, W. Sebald, and P. Reinemer. Crystal structure of the interleukin-4/receptor x3b1; chain complex reveals a mosaic binding interface. *Cell*, 97(2):271–81, 1999.

[171] Y. Wang, B.J. Shen, and W. Sebald. A mixed-charge pair in human interleukin 4 dominates high-affinity interaction with the receptor alpha chain. *Proc Natl Acad Sci U S A*, 94(5):1657–62, 1997.

[172] T.N. Bhat, G.A. Bentley, G. Boulot, M.I. Greene, D. Tello, and W. Dall'Acqua. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences*, 91(3):1089–93, 1994.

[173] W. Dall'Acqua, E.R. Goldman, E. Eisenstein, and R.A. Mariuzza. A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry*, 35(30):9667–76, 1996. PubMed PMID: 8703938.

[174] W. Dall'Acqua, E.R. Goldman, W. Lin, C. Teng, D. Tsuchiya, and H. Li. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, 37(22):7981–91, 1998. PubMed PMID: 9609690.

[175] D. Lim, H.U. Park, L. De Castro, S.G. Kang, H.S. Lee, and S. Jensen. Crystal structure and kinetic analysis of β-lactamase inhibitor protein-ii in complex with tem-1 β-lactamase. *Nat Struct Biol*, 8(10):848–52, 2001.

[176] S. Albeck, R. Unger, and G. Schreiber. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J Mol Biol*, 298(3):503–20, 2000.

[177] T. Selzer, S. Albeck, and G. Schreiber. Rational design of faster associating and tighter binding protein complexes. *Nat Struct Biol*, 7(7):537–41, 2000.

[178] Z. Zhang and T. Palzkill. Determinants of binding affinity and specificity for the interaction of tem-1 and sme-1 beta-lactamase with beta-lactamase inhibitory protein. *J Biol Chem*, 278(46):12933802, 2003.

[179] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein-protein binding interfaces. *Proc Natl Acad Sci U S A*, 102(1):57–62, 2005.

[180] D. Reichmann, M. Cohen, R. Abramovich, O. Dym, D. Lim, and N.C. Strynadka. Binding hot spots in the tem1-blip interface in light of its modular architecture. *J Mol Biol*, 365(3):17070843, 2007.

[181] S. Li, K.R. Schmitz, P.D. Jeffrey, J.J.W. Wiltzius, P. Kussie, and K.M. Ferguson. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell*, 7(4):301–11, 2005.

[182] J.N. Haidar, W. Zhu, J. Lypowy, B.G. Pierce, A. Bari, and K. Persaud. Backbone flexibility of cdr3 and immune recognition of antigens. *J Mol Biol*, 426(7):1583–99, 2014. PubMed PMID: 24380763.

[183] L. Huang, F. Hofer, G.S. Martin, and S.-H. Kim. Structural basis for the interaction of ras with raigds. *Nat Struct Biol*, 5(6):422–6, 1998.

[184] C. Kiel, T. Selzer, Y. Shaul, G. Schreiber, and C. Herrmann. Electrostatically optimized ras-binding ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proc Natl Acad Sci U S A*, 101(25):9223–8, 2004.

[185] C. Kiel, L. Serrano, and C. Herrmann. A detailed thermodynamic analysis of ras/effector complex interfaces. *J Mol Biol*, 340(5):15236966, 2004.

[186] L. Prasad, E.B. Waygood, J.S. Lee, and L.T.J. Delbaere. The 2.5 Å resolution structure of the jel42 fab fragment/hpr complex11edited by i. *A. Wilson. J Mol Biol*, 280(5):829–45, 1998.

[187] S. Sharma, F. Georges, L.T. Delbaere, J.S. Lee, R.E. Klevit, and E.B. Waygood. Epitope mapping by mutagenesis distinguishes between the two tertiary structures of the histidine-containing protein hpr. *Proc Natl Acad Sci U S A*, 88(11):4877–81, 1991.

[188] 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. *EMBO J*, 5(10):2453–8, 1986.

[189] S.M. Lu, W. Lu, M.A. Qasim, S. Anderson, I. Apostol, and W. Ardelt. Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proc Natl Acad Sci U S A*, 98(4):1410–5, 2001. PubMed PMID: 11171964; PubMed Central PMCID: PMCPMC29270.

[190] W. Lu, I. Apostol, M.A. Qasim, N. Warne, R. Wynn, and W.L. Zhang. Binding of amino acid side-chains to s1 cavities of serine proteinases. *J Mol Biol*, 266(2):441–61, 1997. PubMed PMID: 9047374.

[191] 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. *Proc Natl Acad Sci U S A*, 86(15):5938–42, 1989.

[192] J. Pons, A. Rajpal, and J.F. Kirsch. Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the hyhel-10/lysozyme interaction. *Protein Sci*, 8(5):958–68, 1999.

[193] K. Tsumoto, K. Ogasahara, Y. Ueda, K. Watanabe, K. Yutani, and I. Kumagai. Role of tyr residues in the contact region of anti-lysozyme monoclonal antibody hyhel10 for antigen binding. *J Biol Chem*, 270(31):18551–7, 1995. PubMed PMID: 7629185.

[194] L.N. Kam-Morgan, S.J. Smith-Gill, M.G. Taylor, L. Zhang, A.C. Wilson, and J.F. Kirsch. High-resolution mapping of the hyhel-10 epitope of chicken lysozyme by site-directed mutagenesis. *Proc Natl Acad Sci U S A*, 90(9):3958–62, 1993. PubMed PMID: 7683415; PubMed Central PMCID: PMCPMC46425.

[195] M.G. Taylor, A. Rajpal, and J.F. Kirsch. Kinetic epitope mapping of the chicken lysozyme.hyhel-10 fab complex: delineation of docking trajectories. *Protein Sci*, 7(9):1857–67, 1998.

[196] A. Rajpal, M.G. Taylor, and J.F. Kirsch. Quantitative evaluation of the chicken lysozyme epitope in the hyhel-10 fab complex: free energies and kinetics. *Protein Sci*, 7(9):1868–74, 1998.

[197] N.A.G. Meenan, A. Sharma, S.J. Fleishman, C.J. MacDonald, B. Morel, and R. Boetzel. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proceedings of the National Academy of Sciences*, 107(22):10080–5, 2010.

[198] A.H. Keeble, L.A. Joachimiak, M.J. Maté, N. Meenan, N. Kirkpatrick, and D. Baker. Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *J Mol Biol*, 379(4):745–59, 2008. PubMed PMID: 18471830.

[199] M. Sokolovski, J. Cveticanin, D. Hayoun, I. Korobko, M. Sharon, and A. Horovitz. Measuring inter-protein pairwise interaction energies from a single native mass spectrum by double-mutant cycle analysis. *Nature communications*, 8(1), 2017.

[200] W. Li, S.J. Hamill, A.M. Hemmings, G.R. Moore, R. James, and C. Kleanthous. Dual recognition and the role of specificity-determining residues in colicin e9 dnase-immunity protein interactions. *Biochemistry*, 37(34):11771–9, 1998. PubMed PMID: 9718299.

[201] M. Ultsch, J. Bevers, G. Nakamura, R. Vandlen, R.F. Kelley, and L.C. Wu. Structural basis of signaling blockade by anti-il-13 antibody lebrikizumab. *J Mol Biol*, 425(8):1330–9, 2013.

[202] M. Ultsch, J. Bevers, G. Nakamura, R. Vandlen, R.F. Kelley, and L.C. Wu. Structural basis of signaling blockade by anti-il-13 antibody lebrikizumab. *J Mol Biol*, 425(8):1330–9, 2013. PubMed PMID: 23357170.

[203] R. Guerois, J.E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol*, 320(2):369–87, 2002.

[204] C. Viricel, S. Givry, T. Schiex, and S. Barbe. Cost function network-based design of protein–protein interactions: predicting changes in binding affinity. *Bioinformatics*, 34(15):2581–9, 2018.

[205] C. Geng, A. Vangone, G.E. Folkers, L.C. Xue, and Bonvin A.M.J.J. iSEE. Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics*, 87(2):110–9, 2019.

[206] B. Hurley, B. O'Sullivan, D. Allouche, G. Katsirelos, T. Schiex, and M. Zytnicki. Multi-language evaluation of exact solvers in graphical model discrete optimization. *Constraints*, 21(3):413–34, 2016.

[207] R.F. Alford, A. Leaver-Fay, O.'Meara Jeliazkov, JR, DiMaio MJ, Park FP, and H. The rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*, 13(6):3031–48, 2017.

[208] H. Park, P. Bradley, P. Greisen, Y. Liu, V.K. Mulligan, and D.E. Kim. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J Chem Theory Comput*, 12(12):6201–12, 2016.

[209] C.R. Miller, E.L. Johnson, A.Z. Burke, K.P. Martin, T.A. Miura, and H.A. Wichman. Initiating a watch list for ebola virus antibody escape mutations. *PeerJ*, 4:e1674, 2016.

[210] J.S. Patel, C.J. Quates, E.L. Johnson, and F.M. Ytreberg. Expanding the watch list for potential ebola virus antibody escape mutations. *PloS one*, 14(3), 2019.

[211] J. Yang, N. Naik, J.S. Patel, C.S. Wylie, W. Gu, and J. Huang. Predicting the viability of beta-lactamase: How folding and binding free energies correlate with beta-lactamase fitness. *PloS one*, 15(5), 2020.

[212] C. Croux and C. Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical Methods Applications*, 19(4):497–515, 2010.

[213] M.Z. Tien, A.G. Meyer, D.K. Sydykova, S.J. Spielman, and C.O. Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, 8(11), 2013.

[214] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637, 1983.

[215] G.A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi. Plumed 2: New feathers for an old bird. *Comput Phys Commun*, 185(2):604–13, 2014.

[216] P.H.A. Sneath. Relations between chemical structure and biological activity in peptides. *J Theor Biol*, 12(2):157–95, 1966.

[217] Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, and Karaca E. The haddock2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J Mol Biol*, 428(4):720–5, 2016.

[218] C. Dominguez, R. Boelens, and A.M.J.J. Bonvin. Haddock: A Protein-Protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–7, 2003.

[219] T.A. Wassenaar, M. Dijk, N. Loureiro-Ferreira, G. Schot, S.J. Vries, and C. Schmitz. Wenmr: Structural biology on the grid. *Journal of Grid Computing*, 10(4):743–67, 2012.

[220] T. Siebenmorgen and M. Zacharias. Computational prediction of protein–protein binding affinities. *WIREs Computational Molecular Science*, 10(3), 2020.

[221] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, and J.Z.H. Zhang. End-point binding free energy calculation with mm/pbsa and mm/gbsa: Strategies and applications in drug design. *Chem Rev*, 119(16):9478–508, 2019. PubMed PMID: 31244000.

[222] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo. Recent developments and applications of the mmpbsa method. *Frontiers in molecular biosciences*, 4(87), 2017.