Computational Analysis, Design, and Experimental Validation of Antibody Binding Affinity Improvements Beyond *In Vivo* Maturation

by

Shaun Matthew Lippow
Bachelor of Science in Chemical Engineering
University of Wisconsin - Madison, 2001

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2007

Author .....
Department of Chemical Engineering
February 9, 2007

Certified by .....
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Certified by .....
K. Dane Wittrup
J. R. Mares Professor of Chemical Engineering and Biological Engineering
Thesis Supervisor

Accepted by .....
William M. Deen
Chairman, Department Committee on Graduate Students

Computational Analysis, Design, and Experimental Validation of Antibody Binding Affinity Improvements Beyond *In Vivo* Maturation

by

Shaun Matthew Lippow

Submitted to the Department of Chemical Engineering
on February 9, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

This thesis presents novel methods for the analysis and design of high-affinity protein interactions using a combination of high-resolution structural data and physics-based molecular models. First, computational analysis was used to investigate the molecular basis for the affinity improvement of over 1000-fold of the fluorescein-binding antibody variant 4M5.3, engineered previously from the antibody 4-4-20 using directed evolution. Electrostatic calculations revealed mechanistic hypotheses for the role of four mutations in a portion of the improvement, subsequently validated by separate biochemical experiments. Next, methods were developed to computationally redesign protein interactions in order to rationally improve binding affinity. In the anti-lysozyme model antibody D1.3, modest binding improvements were achieved, with the results indicating potentially increased sucesss using predictions that emphasize electrostatics, as well as the need to address the over-prediction of large amino acids. New methods, taking advantage of the computed electrostatics of binding, yielded robust and significant improvements for both model and therapeutic antibodies. The antibody D44.1 was improved 140-fold to 30 pM, and the FDA-approved antibody cetuximab (Erbitux) was improved 10-fold to 52 pM, with an experimental success rate of greater than 60% for single mutations designed to remove unsatisfied polar groups or improve misbalanced electrostatic interactions. Finally, a physics-based improvement to the calculation of the nonpolar component of solvation free energy was implemented and parameterized to address the over-prediction of large amino acids. These results demonstrate novel computational capabilities and indicate their applicability for enhancing and accelerating development of reagents and therapeutics.

Thesis Supervisor: Bruce Tidor
Title: Professor of Biological Engineering and Computer Science

Thesis Supervisor: K. Dane Wittrup
Title: J. R. Mares Professor of Chemical Engineering and Biological Engineering

Acknowledgments

I am indebted to by advisors for their support, ideas, and cooperation throughout the entire thesis process. I thank Dane, especially for his outlook on projects, acute research sense, and career advice from early on, and Bruce, especially for keeping an eye on the 30,000 foot view, emphasizing the rigorous way to do good science, and guiding me with effective scientific presentation.

I am grateful for the useful and timely advice from both Bernhardt Trout and Doug Lauffenburger. Their insights proved invaluable in shaping this work and hopefully increasing its impact.

This work would also not have been possible without the help of many past and present members of both the Tidor and Wittrup labs. I would like to thank everyone that I had the privilege of overlapping with: a phenomenal group of people that gave freely of their time to help and teach me, and to share with me their own research.

In particular, I thank: Karl Hanf for helping in the very beginning and laying part of the protein design groundwork; David Green for insightful research opinions and technical help in particular with continuum electrostatics; Michael Altman for exceedingly unselfish assistance in all areas of computational execution, as well as tailoring development of his dead-end elimination and A* software for my work; Alessandro Senes for contributions to the protein side chain rotamer libraries; Stefan Zajic for patience and skill in teaching me new molecular biology techniques; Wai Lau and Daša Lipovšek for additional wet lab instruction; Brian Joughin for valuable research discussions; and David, Michael, Brian, Mala Radhakrishnan, Kathryn Armstrong, Bracken King, and Dave Huggins for feedback on my protein design software.

I thank my family for years of love and support. This wouldn't have happened without them (don't worry, you don't have to read the whole thing).

And finally, I thank my fiancée, Meg, for her patience, support, and love.

Contents

List of Figures

List of Tables

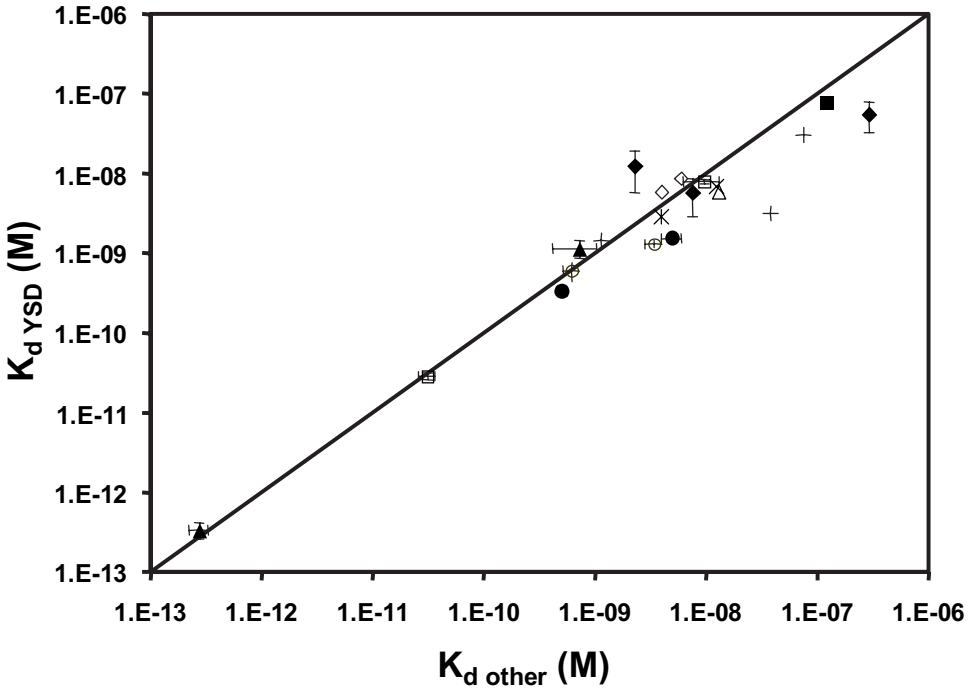
Chapter 1

Introduction

Antibodies are used extensively in diagnostics and as therapeutic agents. Achieving high-affinity binding of antibodies to their antigens is important for expanding detection limits, extending dissociation half-times, decreasing drug dosages, and increasing drug efficacy. However, antibodies produced from an *in vivo* immune response generally exhibit affinities in the 10 nM to 100 pM range [1], often necessitating further engineering. Directed evolution can be used to engineer molecular properties, but computational design holds the promise of far greater exploration of sequence space than possible experimentally, enabling rapid and inexpensive antibody improvement.

The improvement of protein–ligand interactions using different molecular display and directed evolution methods has been reviewed recently by Levin and Weiss [2]. Platforms for display include phage, bacteria, yeast, ribosome, and mRNA, each with their respective advantages and disadvantages. These engineering strategies require significantly more time than would be required to directly express and validate computationally-designed variants. In this work we take advantage of the yeast-display system for rapid and accurate characterization of rationally-designed antibody variants. Equilibrium binding affinities can be measured for 1–20 variants within 1–2 weeks time, and the yeast-display method yields measured binding affinities in quantitative agreement with measurements from other, off-yeast methods (Figure 1.1).

This thesis presents the development of computational protein design methods for the rational selection of mutations to improve binding affinity. Computational design depends critically on two capabilities: accurate energetic evaluation and thorough conformational search. Previous work has addressed many problems related to the design of improved protein–protein binding affinity, such as the design of stable protein folds [14–18], binding pockets for peptides and small molecules [19–22], altered



	System and yeast-display reference(s)	Off-yeast method
○	scFvs binding to hen egg-white lysozyme (HEL) [4]	Fluorescence quenching [3]
●	^{10}Fn3-based antibody mimics binding to HEL [5]	Equilibrium competition with purified antibody mimics [5]
+	Neutralizing scFvs binding to botulinum neurotoxin [6]	Surface plasmon resonance [6]
▲	scFvs binding to fluorescein [7]	Fluorescence quenching [8]
□	scFv binding to carcinoembryonic antigen (CEA) [9]	Mammalian cell-displayed CEA with soluble scFv [10]
■	scFv binding to xeroderma pigmentosum-complementing protein group A [11]	Surface plasmon resonance [11]
◇	scFvs binding to p53 peptides [11]	Surface plasmon resonance [11]
∗	scFv binding to epidermal growth factor (EGF) [11]	Surface plasmon resonance [11]
×	scFv binding to heparin-binding EGF [11]	Surface plasmon resonance [11]
◆	Soluble T-cell receptors binding to staphylococcal enterotoxin C3 [12]	Surface plasmon resonance [12]
△	Soluble T-cell receptor binding to toxic shock syndrome toxin-1 [13]	Surface plasmon resonance [13]

Figure 1.1: Comparison of binding affinities determined by yeast surface display ($K_{d\ YSD}$) and other methods ($K_{d\ other}$). scFv, single-chain antibody fragment.

protein–protein specificity [23–29], and altered enzymatic activity [30–35]. The design of improved antigen-binding affinity has met with limited success, however [36–39]. Challenges for protein–protein affinity design include conformational change upon binding, water molecules trapped between binding partners, polar and charged side chains, and the trade-off of protein–solvent with protein–protein interactions from the unbound to bound state. Fine energy discrimination for redesign from nanomolar to picomolar affinities is a particular challenge.

A robust design strategy should both produce a significant fraction of designs that are successful when tested experimentally and yield substantial improvements across multiple systems. Though there are potentially many mutations that confer improved binding affinity for a particular interaction, calculations need only identify a subset to be successful [40, 41]. Our approach utilizes thorough optimization techniques that exhaustively rank-order the best solutions in a discretized search space. Although some of these solutions are expected to be improved designs, others will be unsuccessful but may be useful in learning about deficiencies in the energy functions, search procedures, or other methodology.

The present work

In Chapter 2, we use computational tools to analyze a high-affinity variant of a fluorescein-binding antibody. The motivation for the work was to gain insight into the molecular mechanisms for the over 1000-fold binding affinity improvement. Previous biochemical and structural analyses had not elucidated the energetic role of any of the 14 mutations accumulated by the four rounds of directed evolution [8]. At the same time, we were motivated by an interest in applying similar computational tools for the rational design of affinity-enhancing mutations. Though the 14-mutation high-affinity variant represents just one of possibly many pathways for affinity maturation, if the calculations could explain or recapitulate even a subset of the mutations selected, the potential applicability of similar computational methods for design would be partially validated. Here we show that rigorous analysis us-

ing the Poisson–Boltzmann continuum electrostatic framework alone results in novel hypotheses for molecular mechanisms for four of the mutations involved [42], with subsequent, separate mutational analyses supporting our conclusions [7].

Chapter 3 introduces a two-stage hierarchical design procedure to address the limitations of energy function accuracy and conformational search thoroughness. Single mutations for improving the binding affinity of an anti-lysozyme model antibody were computationally designed and experimentally assayed. While the affinity improvements were marginal and the success rate was low, the systematic design approach facilitated two important conclusions. First, additional emphasis on the electrostatic term of the binding free energy may improve the predictive accuracy of the methods. Second, an energy function improvement was needed to address the over-prediction of unintuitive mutations to larger amino acids.

In Chapter 4, the hypothesis that prediction accuracy would improve using a method with greater attention to calculated electrostatics was explored. Moreover, several new antibody systems were used to investigate the dependence on system versus the change in methods, as our goal is the development of computational methods that are robust for both high experimental success rates and transferability to different proteins. Overall, we find significant improvements for both model and therapeutic antibodies, including 140-fold improvement to 30 pM and 10-fold improvement to 52 pM, respectively, validating the methods and indicating their applicability for enhancing and accelerating the developement of reagents and therapeutics.

Despite successful affinity maturation in Chapter 4 using computational protein design, deficiencies in the methods clearly remain and limit their further extension to more ambitious design challenges. The finding in Chapter 3 that predictions for improving affinity are dominated by mutations to larger amino acids indicated a fundamental problem in the underlying energy function. Future computational endeavors addressing problems involving the simultaneous design of greater numbers of protein positions would like by limited by this accuracy issue.

Chapter 5 investigates a new model [43] for the nonpolar component of the solvation free energy in order to address a hypothesized imbalance in the calculation

of protein–protein and protein–water interactions. The model is implemented and parameterized, and then analyzed for its effect on the design of single mutations for improved binding affinity. The new nonpolar model attenuates as expected the overall prediction of larger side chains, however, it does not greatly affect the few largest-magnitude unintuitive designs.

This thesis contributes novel and robust methods for the improvement of protein binding affinity using computationally-predicted side-chain mutations. We have advanced the field of computational protein design in two directions: one toward engineering, and one toward basic science. We developed prediction methods that focus on calculated improved electrostatic free energy of binding and yield significant affinity improvements in model and therapeutic antibodies. Also, we investigated an improved, physics-based calculation of the nonpolar component of solvation free energy to address a fundamental tendency of current design methods to favor mutation to larger amino acids. Our improvements to current design capabilities should enhance and accelerate the development of protein reagents and therapeutics.

Chapter 2

Computational analysis of a high-affinity mutant antibody[1]

Abstract

Computational analysis was used to study the molecular basis for affinity improvement in an ultra-high-affinity single-chain antibody. Previous biochemical and structural work resulted in little insight for the 14 mutations used for over 1000-fold improvement. Here, electrostatic calculations reveal several mechanistic hypotheses for the role of four mutations in a portion of the energetic improvement. Subsequent biochemical experiments validated many of these hypotheses, supporting the role for computation in molecular analysis. Overall, the binding affinity improvement appears to be the sum of many small changes. This work indicates potentially novel computational capabilities for using electrostatic calculations to design high-affinity protein interactions.

[1]Portions of this chapter have been previously published as:

Midelfort, K. S., Hernandez, H. H., Lippow, S. M., Tidor, B., Drennan, C. L. & Wittrup, K. D. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J. Mol. Biol.* **343**:685-701 (2004).

2.1 Introduction

Expanding the fundamental understanding of high-affinity molecular interactions is important for improving the ability to engineer molecules with enhanced affinity. Analysis of protein variants with diverse binding affinities for a shared target can reveal insights into the mechanistic, molecular basis of improved protein interactions. Although there has been much study of the transition from micromolar to nanomolar affinity [44–49], there are few examples of studies of improvement from nanomolar to picomolar and higher affinity.

Boder and co-workers previously engineered the 4M5.3 variant of the 4-4-20 single-chain antibody to bind its hapten, fluorescein, with over 1000-fold improved binding affinity from nanomolar to femtomolar [8]. Subsequently, Midelfort and co-workers compared the two antibodies using thermodynamic, kinetic, structural, and theoretical analyses [42]. A summary of their experimental findings is below, followed by the details of this theoretical contribution.

Thermodynamics of the 4-4-20 and 4M5.3 interactions were studied by direct equilibrium titration, equilibrium competition titration, and isothermal titration calorimetry. Improvement in the change in enthalpy of binding, -4.0 (± 0.1) kcal/mol, is the majority of the -4.5 (± 0.1) kcal/mol free energy difference between 4-4-20 and 4M5.3. Stopped-flow fluorescence binding kinetics demonstrate the equivalence of K_{d} and $k_{\mathrm{off}}/k_{\mathrm{on}}$, consistent with a two-state transition from the unbound to bound states. The kinetic data provide no evidence for an encounter complex.

The structure of 4M5.3 in complex with fluorescein was solved to 1.5 Å resolution by molecular replacement using the previously reported 4-4-20 structure [50]. Overall, there are no large differences between the two structures. The RSMD over backbone atoms is 0.60 Å. Of the 14 mutations in 4M5.3, three are in the first contact shell, four in the second, and three in the third. The final four side chains are further away and solvent-exposed. 4-4-20 buries 25 Å^2 more polar surface area and there is a negligible difference in shape complementarity.

The results of different biochemical experiments did not lead to any mechanistic

Table 2.1: Computed total electrostatic contributions to binding.[a]

	scFv Desolvation	Fluorescein Desolvation	Interaction	Net Binding
4-4-20	14.36	14.80	−19.97	9.20
4M5.3	14.68	16.64	−25.62	5.70
4M5.3 − 4-4-20	0.32	1.84	−5.65	−3.50

[a]All values in kcal/mol.

explanations for the over three orders of magnitude binding improvement of 4M5.3. With the crystal structure of the 4M5.3/fluorescein complex solved in collaboration with the Drennan lab, a collaboration was started between the Wittrup and Tidor labs to use computation to investigate the pair of antibodies.

2.2 Results

2.2.1 Contribution of electrostatics to binding

The role of improved electrostatic interactions was investigated through continuum electrostatic binding calculations on the two prepared, unminimized crystal structures. The overall electrostatic contribution to binding computed by this model is summarized in Table 2.1. An overall improvement of −3.5 kcal/mol for 4M5.3 relative to 4-4-20 was computed due to improved interactions with fluorescein (−5.7 kcal/mol) despite increased desolvation penalty for the fluorescein (+1.8 kcal/mol). Interestingly, the antibody desolvation cost was very similar for both complexes and in neither case was the full cost of desolvating the binding partners recovered in intermolecular interactions.

To further understand the computed difference in electrostatics, the −3.5 kcal/mol total electrostatic affinity difference was dissected into individual components. A separate component was defined for each backbone amino (Cα–NH), each backbone carbonyl (C=O), and each side chain group beyond Cα. The energetics for each component were divided into desolvation (changed interactions with solvent due to binding), indirect interactions (changed intramolecular interactions due to binding),

and direct interactions (changed intermolecular interactions between binding partners in the bound state). The components provide additive contributions to the computed difference in binding affinity. This type of dissection is possible in a strictly additive manner because of the superposition properties of the linearized Poisson–Boltzmann equation. After breaking the 4-4-20 and 4M5.3 energetics into their respective components, all corresponding components between the two antibody variants were differenced, so that non-zero values indicate components of potential interest. Table 2.2 shows the difference of component analyses between the two crystal structures. Three of the top four components and four of the top 20 components, ranked by the metric *SumAbs* (Equation 2.1), are 4M5.3 side chain mutations.

$$SumAbs_i = |Desolv_i^{4M5.3} - Desolv_i^{4-4-20}| + \sum_{j \neq i} |Inter_{i-j}^{4M5.3} - Inter_{i-j}^{4-4-20}| \quad (2.1)$$

SumAbs is a measure of how different each component is between the two analyses. $Desolv_i^{\mathrm{X}}$ represents the desolvation penalty for component i in complex X, and $Inter_{i-j}^{\mathrm{X}}$ represents the interaction free energy between the atoms in component i and the atoms in component j in complex X.

The direct role of the 14 mutations together was calculated by summing together the components involving the 14 mutated side chains, revealing a total of -1.1 kcal/mol (Table 2.3). Four side chains dominate this value: H31 Asp-to-His, H101 Ser-to-Ala, H102 Tyr-to-Ser, and H106 Asp-to-Glu. The remaining -2.4 kcal/mol in computed electrostatic improvement is the result of secondary effects from the mutated residues (e.g. altering the binding conformation of fluorescein in the site). Three components dominate this -2.4 kcal/mol: L31 His, L39 Arg, and L96 Ser, each of which exhibit slightly shortened hydrogen-bond distances in 4M5.3 relative to 4-4-20.

2.2.2 Effect of minimization on electrostatic contributions

Because the results of this analysis produced significant free energy differences that were the sum of quite small values from individual changed interactions, further calcu-

Table 2.2: Difference in electrostatic energy components.

Component	Desolv.[a]	Indirect[b]	Direct[c]	Contrib.[d]	Mut.[e]	$SumAbs^f$
Fluorescein	1.84	N.A.	−5.65	−0.98	−3.81	
L39 Arg	0.01	1.24	−1.45	−0.09	−0.20	6.57
H106 Asp-to-Glu*	−0.33	0.11	−0.13	−0.34	−0.35	4.50
H101 Ser-to-Ala*	−0.11	0.38	−0.79	−0.31	−0.52	4.01
H102 Tyr-to-Ser*	0.68	−0.51	−0.46	0.19	−0.29	3.59
H105 carbonyl	−0.32	0.28	−0.16	−0.26	−0.20	3.13
W615 water	0.04	0.42	−0.85	−0.18	−0.40	2.53
H31 carbonyl	0.21	−0.95	0.24	−0.15	−0.50	2.36
H102 amino	−0.21	0.13	0.15	−0.07	0.07	2.18
H101 carbonyl	−0.04	−0.17	0.14	−0.05	−0.07	2.15
L31 His	−0.15	−0.30	−0.47	−0.53	−0.91	1.85
H74 Arg	0.05	0.21	−0.18	0.07	0.08	1.68
H102 carbonyl	−0.07	0.15	0.10	0.06	0.19	1.64
L96 Ser	−0.02	0.15	−0.76	−0.32	−0.63	1.64
H104 amino	−0.04	0.01	0.28	0.11	0.26	1.55
H53 Asn	0.27	−0.46	0.11	0.10	−0.07	1.52
H52 Arg	−0.01	0.15	−0.12	0.00	0.01	1.46
H103 carbonyl	−0.01	−0.00	−0.23	−0.12	−0.24	1.31
H31 Asp-to-His*	−0.01	−0.04	−0.29	−0.18	−0.34	1.20
H104 carbonyl	−0.06	0.17	−0.27	−0.11	−0.16	1.19
H103 amino	−0.04	0.09	−0.04	−0.01	0.02	1.17

All values are the differences between analyses on 4M5.3 and 4-4-20, in kcal/mol. Negative (positive) values correspond to a contribution toward improvement (reduction) for 4M5.3 binding relative to 4-4-20.

*Site of 4M5.3 mutation

[a]Desolvation penalty.

[b]Sum of indirect interactions with all other scFv components.

[c]Direct interaction between scFv component(s) and fluorescein.

[d]Contribution = desolvation + (1/2)indirect + (1/2)direct.

[e]Mutation = desolvation + indirect + direct.

[f]*SumbAbs*, sum of the absolute value of desolvation and all interaction terms as given by Equation 2.1.

Table 2.3: Electrostatic impact at the mutated side-chain positions.

Position	Desolv.[a]	Ind. in[b]	Ind. out[c]	Direct[d]	Total
H1 Glu-to-Gly	0.00	−0.01	0.03	−0.06	−0.03
H31 Asp-to-His	−0.01	0.03	−0.09	−0.29	−0.37
H51 Ile-to-Phe	0.00	0.00	0.02	−0.03	−0.02
H101 Ser-to-Ala	−0.11	−0.20	0.78	−0.79	−0.31
H102 Tyr-to-Ser	0.68	0.03	−0.57	−0.46	−0.32
H106 Asp-to-Glu	−0.33	−0.20	0.50	−0.13	−0.15
H108 Trp-to-Leu	0.00	0.03	−0.13	0.18	0.08
Total	0.23	−0.32	0.54	−1.58	−1.13

All values are the differences between analyses on 4M5.3 and 4-4-20, in kcal/mol. For the mutations L60 Phe-to-Val, L81 Ser-to-Asn, H16 Arg-to-Gly, H17 Pro-to-Ala, H24 Ala-to-Thr, H30 Ser-to-Gly, and H93 Met-to-Thr, each of the individual five electrostatic values (in columns in Table 2.2) are ≤ 0.01 in magnitude, and these rows are left out of the Table. Negative (positive) values indicate an improvement (reduction) for 4M5.3.

[a]Desolvation penalty.

[b]Half sum of indirect interactions with the 13 other mutated side-chain components.

[c]Sum of indirect interactions with all but the 13 other mutated side-chain components.

[d]Direct interaction with fluorescein.

Table 2.4: Total electrostatic contributions after minimization.[a,b]

	scFv Desolvation		Fluorescein Desolvation		Interaction		Net Binding	
4-4-20	14.39	(+0.03)	14.28	(−0.52)	−21.57	(−1.60)	7.11	(−2.09)
4M5.3	14.66	(−0.02)	16.62	(−0.02)	−25.45	(+0.17)	5.84	(+0.14)
Difference	0.27	(−0.05)	2.34	(+0.50)	−3.87	(+1.78)	−1.27	(+2.23)

[a]All values in kcal/mol.

[b]Values in parentheses show change from before minimization.

Table 2.5: Components most affected by minimization.

Specific interaction	Change in strength (kcal/mol)
L39 Arg-fluorescein	+0.58
L96 Ser-fluorescein	+0.55
L31 His-fluorescein	+0.53

Table 2.6: Net electrostatic impact of the mutated side-chain components after minimization

	Desolv.	Ind. in	Ind. out	Direct	Total
Total[a]	0.29	−0.29	0.27	−1.42	−1.14

[a]Compared to before minimization in Table 2.3.

lations were performed to investigate sensitivities to precise crystal structure atomic locations. The electrostatic component analyses were repeated on 4-4-20 and 4M5.3 crystal structures each subjected to a constrained minimization. Table 2.4 summarizes the effects of minimization on electrostatic calculations. The net calculated 4M5.3 improvement decreased from −3.50 kcal/mol down to −1.27 kcal/mol, and the biggest change is the interaction between 4-4-20 and fluorescein computed to be more favorable after minimization than before.

Analysis of the effects of minimization on the component analyses reveals that three interactions are responsible for most of the change (Table 2.5). L31 His, L39 Arg, and L96 Ser, the same three residues identified above as having shortened hydrogen bond lengths and a large computed role in the 4M5.3 improvement, are no longer calculated to be as different between 4-4-20 and 4M5.3 when using the minimized structures. The small changes in the hydrogen-bond distances due to the minimization result in relatively large changes in the interaction differences. On the other hand, the contribution of the 14 mutated side chain components after minimization is predominantly unchanged, totaling −1.1 kcal/mol (Table 2.6). Thus, the calculations suggest that the subtle structural differences between the crystal structures ("secondary effects") appear to coalesce during minimization and may not be real, but the direct computed effects appear robust to small structural changes.

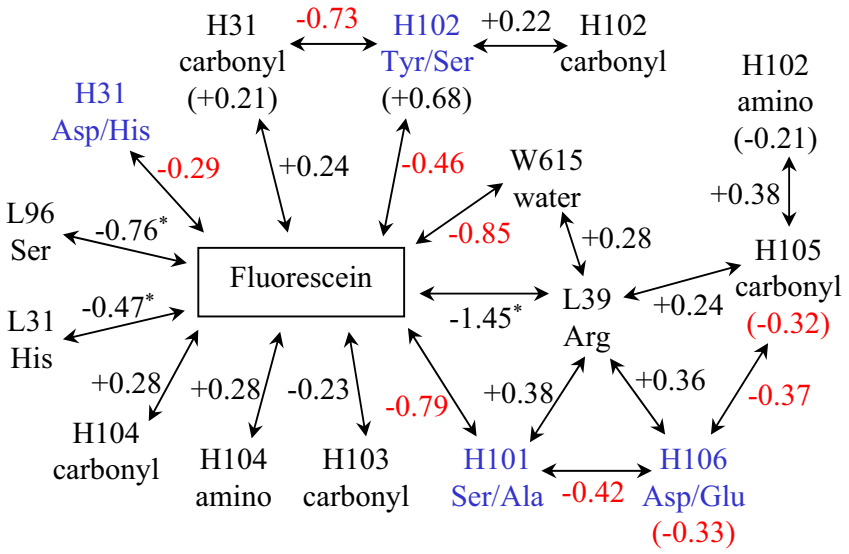


Figure 2.1: Diagram of all major (> 0.2 kcal/mol) electrostatic interaction and desolvation penalty differences (4M5.3 − 4-4-20) among components in Table 2.2. Arrows connecting to fluorescein correspond to changes in direct interactions, whereas arrows connecting scFv components correspond to changes in indirected interactions, and changes in desolvation penalties are shown in parentheses. The three interation differences marked with * are those which dramatically decrease (become more positive) with minimization. All other values < -0.25 kcal/mol are shown in red. Mutated side chains (four) are labeled in blue.

A diagram of the calculated differences in electrostatic components of binding between 4-4-20 and 4M5.3 is presented in Figure 2.1. Besides the three interaction differences that were shown to be highly sensitive to minimization, seven out of the nine significant energetic changes (values in red) are associated with one or more of the four mutated residues: H31, H101, H102, or H106. The H31 and H102 side chains are uncoupled from other side chains, whereas H101 and H106 are strongly coupled to each other.

Four of the 14 mutations account for the majority of the electrostatic differences in Table 2.3: H31 Asp-to-His, H101 Ser-to-Ala, H102 Tyr-to-Ser, and H106 Asp-to-Glu. Reexamination of the 4-4-20 and 4M5.3 structures with the electrostatic results in mind reveals several mechanistic hypotheses for the 4M5.3 improved binding affinity. These mechanisms are outlined in the Discussion section.

2.3 Discussion

Electrostatic components of the binding free energy were calculated using crystal structure snapshots and the framework of a rigid binding model with continuum solvent. Calculations on the two crystal structures revealed -3.5 kcal/mol in favor of 4M5.3, though subsequent analysis shows that hydrogen bonds from the non-mutated residues L39 Arg, L96 Ser, and L31 His with fluorescein account for much of the -3.5 kcal/mol improvement and are highly sensitive to precise atomic locations. It is possible, but still uncertain, that mutations in 4M5.3 are responsible for small structural changes in conserved residues. At the same time, however, four of the mutated residues dominated -1.1 kcal/mol of the binding affinity improvement, regardless of whether the structures used for analysis had been subjected to the constrained minimization. Several specific mechanisms for affinity improvement are revealed through these four residues.

Overall, improvements are found in all three terms of the electrostatic binding free energy: direct interactions (e.g. a hydrogen bond between the scFv and fluorescein), indirect interactions (e.g. a hydrogen bond within the scFv that is buried upon binding and hence strengthened due to loss of solvent-screening), and desolvation (i.e. the penalty paid by a polar group for leaving its unbound, aqueous environment and entering a buried environment upon binding). Binding affinity improvements can be the result of either the introduction/strengthening of a favorable interaction, the removal/weakening of an unfavorable interaction, or the reduction of a desolvation penalty.

Figures 2.2A and 2.2B show the local environment of the mutated sites H101, H106, and H108. H108 was not indicated by the electrostatic analysis for binding free energy, although the position is clearly part of an intramolecular hydrogen-bonding network that includes H101 and H106. Integrating the quantitative calculations in Table 2.3 with qualitative structure analysis reveals the following mechanisms for enhanced affinity. The H101 serine to alanine mutation eliminates the unsatisfied hydroxyl oxygen, removing a small desolvation penalty, yet loses the favorable indirect

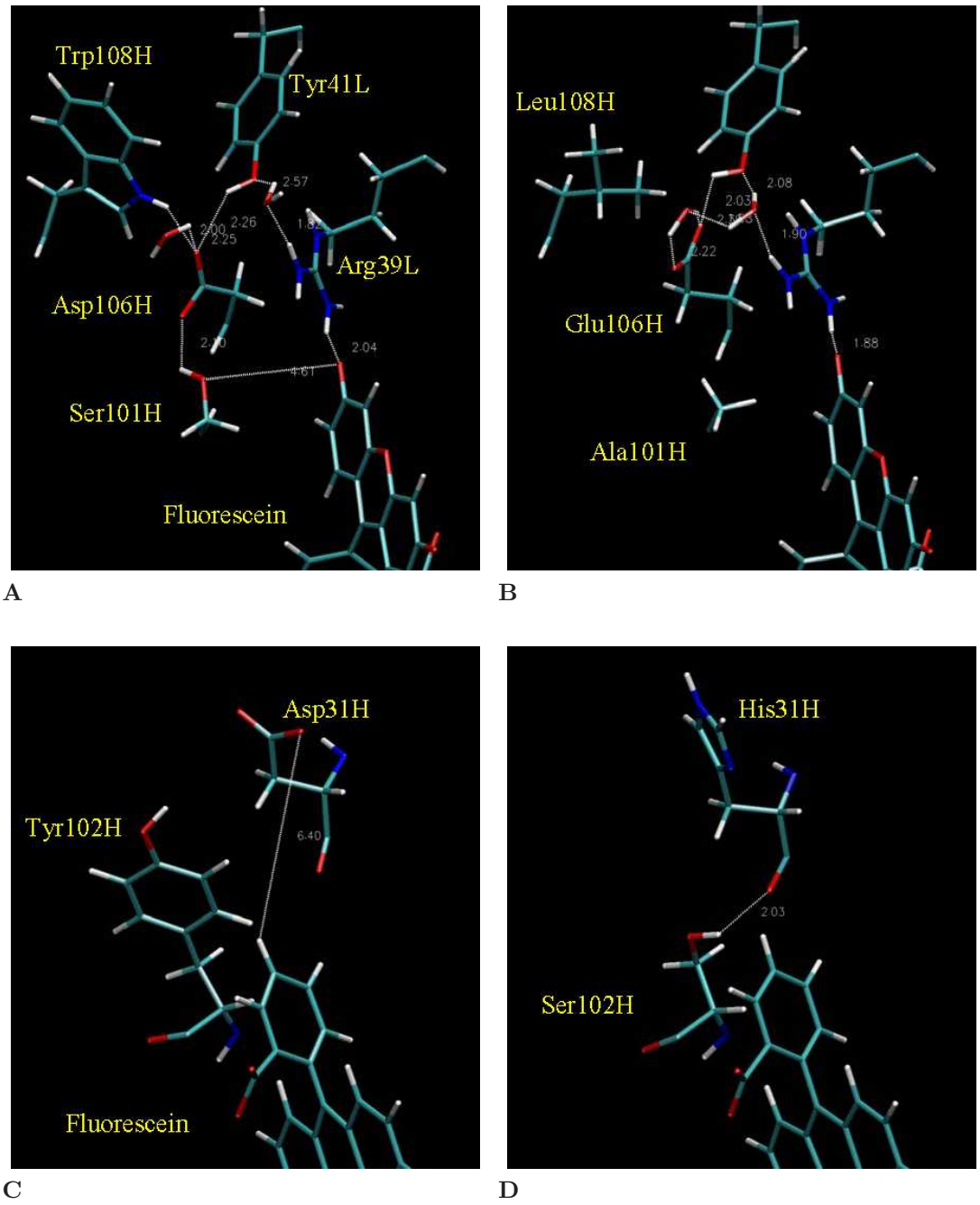


Figure 2.2: Molecular detail of mutated side chains, showing the two regions of the binding site that include the four residues identified by the electrostatic analysis. (A-B) Region of H101 and H106 together with neighboring hydrogen-bonding residues L39, L41, and H108; (A) 4-4-20, and (B) 4M5.3. (C-D) Region of H31 and H102 with fluorescein; (C) 4-4-20, and (D) 4M5.3.

interaction from the H106 aspartate hydrogen bond (mutated to glutamate), for a net favorable effect. This effect is captured in the context of the other 13 mutations, but this analysis does not preclude the possibility of H101 serine to alanine also being a beneficial single mutation (although some interaction with the mutation at H106 is likely). The H106 aspartate to glutamate mutation reduces the residue desolvation penalty by burying the negative charge further into the protein. More notably, the H101 and H106 mutations appear to mutually maintain a satisfied intramolecular hydrogen-bonding network and either mutation alone might disrupt antibody stability. The differences in calculated component energies shown in Figure 2.1 support possible cooperativity between positions H101 and H106. Likewise, the H108 tryptophan to leucine mutation in the bordering region with two bound water molecules may be coupled to the H101 and H106 mutations via hydrogen bonds.

Figures 2.2C and 2.2D show the local environments of H31 and H102. The mutation at H102 from tyrosine to serine creates a new intramolecular hydrogen bond to the backbone carbonyl at residue H31. This hydrogen bond is buried and strengthened upon binding, as shown by the -0.73 kcal/mol term in Figure 2.1. Though the hydrogen bond is to a mutated position (H31), it is independent of the side chain, making this a distinct mechanism from the actual side chain mutation at H31. The aspartate to histidine mutation at H31 removes a negative charge, which although solvent-exposed, was within 7 Å of the -2 charged fluorescein. This mutation is calculated to improve the direct interaction by eliminating a long-range electrostatic repulsion (-0.3 kcal/mol).

The electrostatic calculations indicate 4 of the 14 mutations for a role in the affinity improvement. Mechanisms discovered include the removal of an unsatisfied hydrogen-bonding group (H101 Ser-to-Ala), reduction of desolvation penalty (H106 Asp-to-Glu), creation of an intramolecular hydrogen bond for an indirect effect (H102 Tyr-to-Ser), and removal of long-range charge repulsion (H31 Asp-to-His). Other hypothesized mechanisms include the maintenance of a well-ordered intramolecular hydrogen-bonding network, though calculations of antibody stability were not performed.

The 4-4-20 and 4M5.3 structures used for computational analysis were prepared to minimize bias from inherent crystal structure differences; however, a few differences remain between the structures used for computational analysis and the single-chain antibodies used in the binding experiments. The N-terminal residue of the heavy chain in 4-4-20 is a glutamate in all experiments but the crystal structure contains an aspartate at this position [50]. This discrepancy is most likely minimal since both side chains are charged -1, solvent exposed, and about 20 Å from the fluorescein. Second, the fluorescein used in all binding experiments was biotinylated at the 5 carbon position, pointing out of the binding site, whereas both crystal structures were obtained with neat fluorescein. A concern is whether the 4M5.3 mutations enhance binding affinity through interaction with atoms only present in the biotinylation linker. Two of the mutated residues, H31 Asp-to-His and H102 Tyr-to-Ser are within at least 5 Å of the beginning of the thiourea linker with hydrogen-bonding capabilities. Previous experiments to indicate the extent of the importance of the linker in the binding interaction were not successful.

The experimental and computational results do not definitely elucidate the mechanism for the full 1,800-fold affinity improvement of 4M5.3. The binding improvement appears to be a result of a variety of many interactions and the sum of many small changes. Nevertheless, the electrostatic calculations reveal several mechanisms that may account for part of the improvement.

Subsequent work by Midelfort and Wittrup investigated the individual effects of seven of the 14 4M5.3 mutations, confirming many of the electrostatics-based hypotheses [7]. The single mutations H31 Asp-to-His, H101 Ser-to-Ala, and H102 Tyr-to-Ser are each improved as expected, with H31 and H102 energetically independent as expected. The mutations at H101, H106, and H108 show interactions, or non-additivity, in their effects on ligand affinity, consistent with the postulated role of intramolecular hydrogen bonding.

Two key conclusions are drawn from the complementary experimental and computational work. First, calculations can be used to generate hypotheses for structural mechanisms not found by biochemical techniques. Second, consistency of the

electrostatic analysis with energetics of improved binding support the use of electrostatics as an engineering design tool. This work indicates that it may be realistic for structure-based calculations to correctly identify mutations that improve binding affinity. Moreover, large improvements in affinity may be designed using many smaller, yet additive mutations.

2.4 Methods

2.4.1 Preparation of protein structures

The 1.85 Å 4-4-20 Fab structure (1FLR [50]) and the 1.5 Å structure of 4M5.3 (1X9Q [42]) were used as the basis for molecular modeling, continuum electrostatic calculations, and theoretical analysis. The two structures were prepared in parallel to minimize differences that might bias comparative analysis. Only residues with corresponding crystallographic data in both structures were used, prompting the removal of the 4-4-20 C_L and C_{H1} domains (residues after L112 and after H117) and a few N- and C-terminal residues corresponding to the flexible linker (4-4-20 V_L112 and 4M5.3 V_L0, V_H-2, V_H-1, V_H0), all of which are solvent exposed and at least 15 Å from the fluorescein binding pocket. Six water molecules were retained in each structure (4-4-20: 606, 608, 615, 618, 676, 689; 4M5.3: 6, 3, 1, 11, 26, 47); these six solvent molecules make corresponding interactions in the two structures. All other water molecules were removed from the structure files and were modeled implicitly. Side chain titration states, single conformations for all multiple-occupancy residues, and the crystallographic carbon/nitrogen/oxygen uncertainties in asparagine, glutamine, and histidine side chains were resolved based on examination of side chain local environments. The fluorescein was modeled in its net charge -2 state reflecting the pH of 8 in the binding experiments, and this choice is supported by the local environments of the fluorescein protonatable sites in each structure.

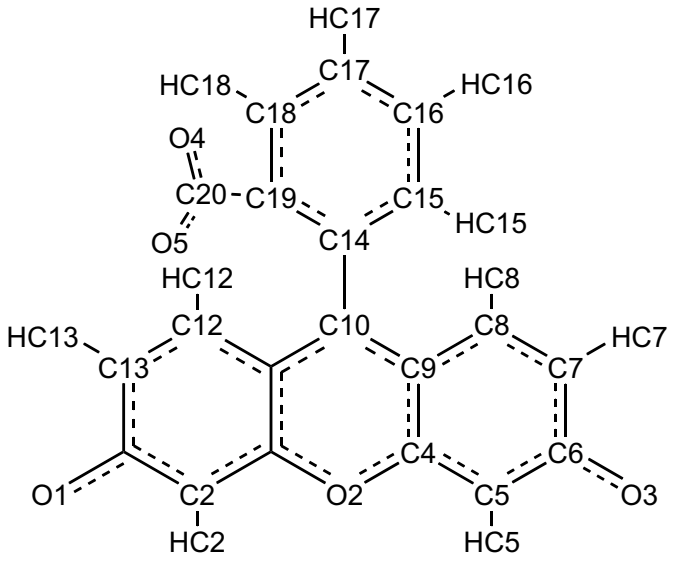
Molecular mechanics was used to prepare the protein structures. The HBUILD facility [51] in the program CHARMM [52] was used to build all hydrogen atoms onto

each structure. The CHARMm22 all-atom parameter set [53] with the CHARMM-adapted TIP3P water model was used for molecular mechanics calculations, assigning appropriate general atom types for the fluorescein (Figure 2.3). Fluorescein partial atomic charges (Figure 2.3) were obtained as described by Green and Tidor [54] by first using the program GAUSSIAN98 [55] with restricted Hartree–Fock and the 6-31G* basis set to optimize the geometry of fluorescein starting from that in the 4-4-20 structure, and subsequently fitting the electrostatic potential using the restrained-fitting methods (RESP) of Bayly *et al* [56].

A second set of 4-4-20 and 4M5.3 structures for comparison were prepared by subjecting each to a harmonically restrained minimization to convergence with 10 kcal/mol/Å^2 force constants on every atom in the system, no non-bonded cut-offs, and a 4r distance-dependent dielectric constant. The root-mean-square deviation for structures before and after minimization were 0.093 Å (0.094 Å) and 0.053 Å (0.053 Å) for 4-4-20 and 4M5.3, respectively, taken over all atoms (taken over side chain atoms only).

2.4.2 Electrostatic calculations

Electrostatic contributions of individual functional groups to the binding free energy were computed using a rigid binding model with continuum solvent following previous work [57]. A locally modified version of the DELPHI program [58–61] was used to solve the linearized Poisson–Boltzmann equation with finite-difference methods. PARSE parameters [62] were used for atomic radii and partial atomic charges of protein. Fluorescein partial atomic charges were obtained from the RESP fitting procedure described above, and radii were assigned based on the PARSE convention. PARSE parameters do not include aliphatic hydrogen atoms and thus the appropriate hydrogen atoms were removed from the system. Fluorescein does not contain any aliphatic hydrogens. The dielectric constant was assigned to 4 for protein, ligand, and explicit water, and 80 for the implicit solvent regions. A salt concentration of 0.145 M was used with a 2.0-Å Stern layer and a molecular surface generated with a 1.4-Å probe sphere. Each molecule was oriented to minimize the volume of the bounding



A

Atom Name	CHARMm22 Atom Type	Partial Charge	Atomic Radius	Atom Name	CHARMm22 Atom Type	Partial Charge	Atomic Radius
C1	C6R	0.69369	1.7	C19	C6R	0.15207	1.7
C2	C6R	-0.56851	1.7	C20	C	0.68940	1.7
C3	CR66	0.30935	1.7	O1	OC	-0.72934	1.4
C4	CR66	0.30921	1.7	O2	O6R	-0.27339	1.4
C5	C6R	-0.56775	1.7	O3	OC	-0.72934	1.4
C6	C6R	0.69370	1.7	O4	OC	-0.68458	1.4
C7	C6R	-0.39479	1.7	O5	OC	-0.78647	1.4
C8	C6R	-0.19359	1.7	HC2	HA	0.15375	1.0
C9	CR66	0.02389	1.7	HC5	HA	0.15338	1.0
C10	C6RP	0.03161	1.7	HC7	HA	0.12383	1.0
C11	CR66	0.02436	1.7	HC8	HA	0.18721	1.0
C12	C6R	-0.19481	1.7	HC12	HA	0.18765	1.0
C13	C6R	-0.39421	1.7	HC13	HA	0.12377	1.0
C14	C6RP	-0.09826	1.7	HC15	HA	0.10480	1.0
C15	C6R	-0.19170	1.7	HC16	HA	0.10643	1.0
C16	C6R	-0.09936	1.7	HC17	HA	0.12460	1.0
C17	C6R	-0.24327	1.7	HC18	HA	0.14604	1.0
C18	C6R	-0.18937	1.7				

B

Figure 2.3: Computational parameterization of fluorescein. The appropriate general atom types from the CHARMm22 all-atom parameter set were used for molecular mechanics calculations. The partial atomic charge distribution used in both molecular mechanics and continuum electrostatics calculations was obtained by quantum-mechanical geometry minimization followed by restrained fitting of the electrostatic potential. Radii for continuum electrostatic calculations were assigned following PARSE convention.

cube. A focusing procedure was used that includes a low grid spacing using 23% fill and Debye–Hückel boundary conditions followed sequentially by higher resolution calculations first at 92% and then at 184% fill centered on the specific functional group of interest. Ten translations relative to the grid were performed and averages were used. The standard error of the mean was on the order of 0.001 kcal/mol, much less than the 0.1-1.0 kcal/mol range of differences identified between 4-4-20 and 4M5.3 components. A $129 \times 129 \times 129$ grid was used, resulting in final grid spacings of 4.49 and 4.40 grid units per Å for 4-4-20 and 4M5.3, respectively. Convergence of free energies with respect to grid resolution (data not shown) indicate that a difference of 0.09 grid units per Å results in about 0.01 kcal/mol change in net binding free energy. Therefore, the comparison between 4-4-20 and 4M5.3 electrostatic calculations on slightly different grids remains valid. For binding calculations, all explicit water molecules remained associated with the antibody in the unbound state. Overall electrostatic contributions to binding were dissected into component contributions from individual chemical groups and physical sources based on the work of Hendsch and Tidor [57].

Chapter 3

Development of computational methods for the design of improved protein binding affinity

Abstract

The development of computational methods for the redesign of high-affinity protein interactions is an important problem that is both fundamental to the advancement of protein design and directly applicable to solving pressing biotechnology needs. Here we have developed structure-based computational methods for the redesign of proteins with improved binding affinity for their protein or small-molecule targets. These methods were applied to the redesign of the model anti-lysozyme antibody D1.3. The results of design calculations to improve binding affinity were unintuitively dominated by mutations to larger amino acids. Predictions of antibody single mutations to improve binding were tested experimentally, yielding a low success-rate and only marginal improvements. Nevertheless, the results validated predictions based on improved binding electrostatics, suggesting an altered design procedure that emphasizes electrostatic predictions (Chapter 4). The results also validated initial energy-function concerns, leading to a subsequent investigation of an improved nonpolar solvation model (Chapter 5).

3.1 Introduction

A fundamental challenge in computational protein design is balancing the accurate evaluation of atomic interactions and the efficient search of a combinatorially-complex space. Early progress in the field addressed the redesign of hydrophobic protein cores in which the physical interactions could be well-approximated by straight-forward pairwise-additive van der Waals interactions, allowing the adoption and development of dead-end elimination algorithms for finding the global minimum energy conformation (GMEC) in a rotamerized discrete side chain space [14, 16, 17, 63–67]. As more complex problems were addressed, such as binding affinity and specificity, enzymatic activity, and *de novo* structures, increasingly complex energy functions were introduced to more accurately evaluate interactions [15, 20–24, 26–29, 68–82]. However, the constraint of a pairwise-additive energy function has limited the development of accurate electrostatics in protein design [83–87].

A two-stage hierarchical design procedure was used to overcome the combined limitations of energy function accuracy and conformational search thoroughness [88]. In the first stage, approximations are made such that the problem is amenable to a suite of algorithms adept at pruning large combinatorial search spaces. In the second stage, with conformational space greatly reduced, low-energy structures are enumerated using more accurate, yet more computationally-demanding models. Together, these steps allow us to examine problems of significant conformational complexity without sacrificing our choice of energy function for final evaluation.

3.2 Methods

Conformational search is initially simplified by assuming a rigid protein backbone and allowing only discrete side chain rotamers. The physics-based energy function is pairwise-decomposable, permitting application of dead-end elimination and A* search algorithms. For each protein sequence, we find its global minimum energy conformation (GMEC), and if this energy is within a specified energy cut-off of the wild-type

GMEC energy, then a continued list of lowest-energy structures are found for that sequence. Second, we reevaluate the lowest-energy structures of each sequence using more accurate, yet more more computationally-demanding models, such as Poisson–Boltzmann continuum electrostatics, unbound state side chain conformation search, and minimization. Structures are reranked based on these latter calculations. Binding energy is initially predicted from the bound state conformation and a rigid binding model. The unbound state search is used to approximate flexible binding and estimate a deformation penalty which offsets binding. Changes to protein fold stability are approximated from the energetic difference between the folded state and isolated model compounds.

3.2.1 Structure preparation

The crystal structure of the complex between the Fv fragment (light chain "L", heavy chain "H", and lysozyme "C") of the antibody D1.3 and hen egg-white lysozyme was obtained from the Protein Data Bank (1VFB [89]). Most crystallographic water molecules were removed, except for 19 that bridge the binding interface or are buried away from bulk solvent. For calculations of two-state rigid binding, water molecules were either assigned to the antibody (water numbers 145, 149, 150, 173, 174, 177, 178, 179, 180, 223, 257, 748) or to lysozyme (water numbers 152, 155, 169, 181, 200, 222, 228). Titration states, multiple occupancies, and asparagine, glutamine, and histidine carbon/nitrogen/oxygen crystallographic uncertainties were resolved based on optimization of hydrogen-bonding in the side-chain local environments. All side chains remained in their default protonation state, with the histidine tautomers as follows. Proton on Nϵ: positions L90 and H86; proton on Nδ: C15 and L30. The terminal dihedral angle of the histidine, asparagine, or glutamine side chains at the following positions were rotated 180°: H56, H77, L37, L90, C19, C37, C59, C93. Hydrogen-atom positions were assigned using the HBUILD facility [51] in the computer program package CHARMM [52] with the PARAM22 all-atom parameter set [90] and the CHARMM-adapted TIP3P water model.

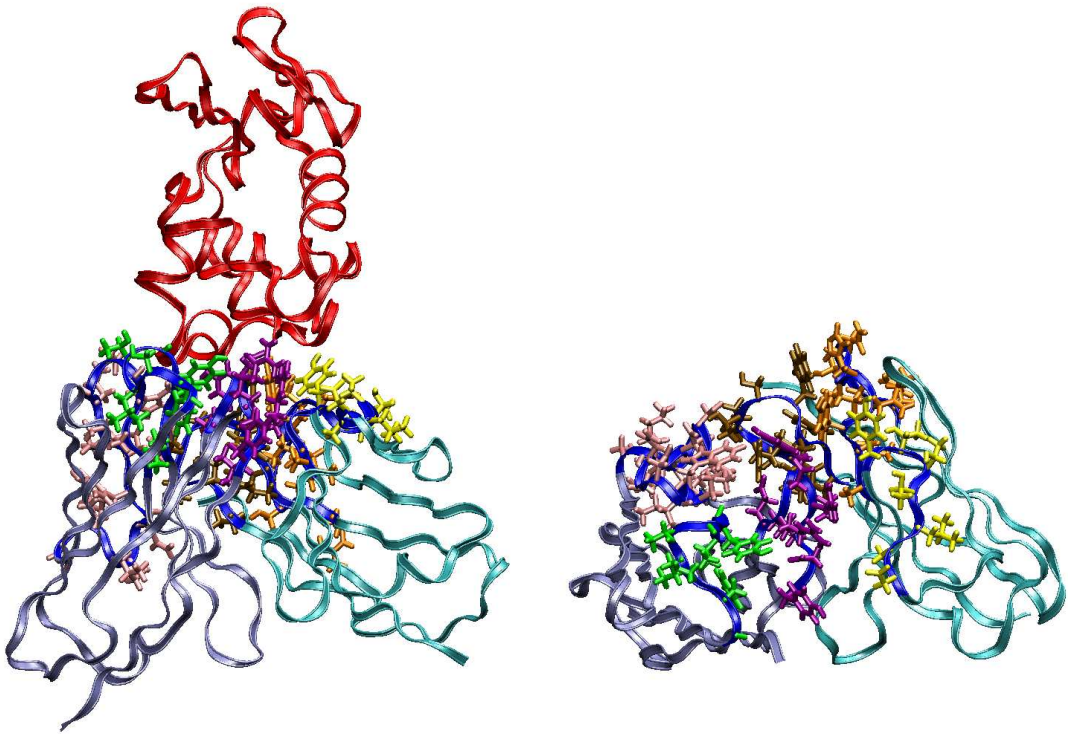
3.2.2 Search space

The Kabat definition for antibody complementarity determining region (CDR) positions was used to select antibody sites for potential computational mutation. This presented 61 positions spanning the six contiguous stretches in primary sequence, as shown in Figure 3.1. The proline at position L95 was left as wild type in all designs. For the design of single mutants, each of these residues was individually mutated to 17 other amino acids (proline and cysteine excluded). For the design of double mutations, the only pairs of positions considered were those with at least one pair of side-chain atoms within 4.75 Å (non-hydrogen atoms beyond Cβ, or the Hα1 atom for glycine). For each independent design, the one or two mutated positions were given rotameric degrees of freedom, as were side chains at nearby positions (same definition of nearby as used above). The two neutral tautomers and the protonated form of histidine were allowed.

The rotamer library was based on the backbone-independent May 2002 library from Dunbrack [91, 92], expanded by $\pm$ 10° in both χ_1 and χ_2. Prior to expansion, three histidine rotamers were added for an unsampled ring flip (64.8, -13.8; -172.8, 9.7; -68.1, -11.3), and two asparagine rotamers were added to increase sampling of the final dihedral angle rotation (-169.3, -155.7; -170.8, 159.8). Hydroxyls of serine, threonine, and tyrosine were sampled every 30°. The library contained 4,025 side-chain rotamers. A novel water library allowed for conformational freedom of crystallographic water molecules. The oxygen atom location was fixed and the hydrogen atoms were placed to create 60 symmetric water molecule rotations. A 61$^{\text{st}}$ water rotamer allowed for the water to no longer exist in the structure. In addition, each crystallographic wild-type rotamer was added in a position-specific manner to the library, using the complete cartesian representation of the side chain, rather than just the dihedral angles.

3.2.3 Energy function and model

The energy function for initial search (named "low-resolution") was the CHARMM PARAM22 all-atom parameter set [90] with no cut-offs for non-bonded interactions



CDR-H1	CDR-H2	CDR-H3	CDR-L1	CDR-L2	CDR-L3
H26-34	H50-65	H98-105	L24-34	L50-56	L89-97
green	pink	purple	orange	yellow	ochre

Figure 3.1: D1.3 CDR positions. The protein backbones are depicted by ribbons (red: lysozyme; iceblue: antibody heavy chain; cyan: antibody light chain; blue: CDRs). The D1.3 wild type side chains at the CDR positions are color-coded. (A) Antibody/lysozyme complex. (B) Antibody-only, looking down on interface in (A).

and a 4r distance-dependent dielectric constant. All energy terms were used (bond, angle, Urey-Bradley, dihedral, improper, Lennard-Jones, and electrostatic). The objective function was the difference between the bound state energy and a sum of isolated, model compounds. Each model compound conformation was the lowest energy of all side-chain rotamers with the local single amino acid backbone with an acetylated N-terminus and an N-methylamide C-terminus. The model compounds are important for canceling intrinsic side-chain energies such as ring strain (especially in PARAM22) when comparing the state energy of different protein sequences; however, the model compounds do not affect actual binding free energy predictions. In

the procedure, which enumerates all possible amino acid sequences during design, the model compounds become irrelevant; when using approximate methods to order sequences in a design with non-enumerable sequence space, the model compounds are important for sequence-to-sequence comparison.

This low-resolution energy and objective function satisfies pairwise-additivity (Equation 3.1) and is readily used in dead-end elimination algorithms.

$$E_{\text{total}} = E_{\text{const}} + \sum_{i=1}^{p} E_{\text{self}_i} + \sum_{i=1}^{p} \sum_{j>i}^{p} E_{\text{pair}_{i,j}} \tag{3.1}$$

Each summation is over the conformational search positions, p. E_{self_i} is the interaction between the residue and the fixed positions, subtracting off the energy of the isolated, model compound for that amino-acid type. The self interaction term includes intra-rotamer molecular mechanics covalent bonding terms (bonds, angles, etc.) and bonding terms to the backbone atoms of -1 or $+1$ positions if those positions are fixed. $E_{\text{pair}_{i,j}}$ is the nonbonded interaction between two positions, plus bonding terms only if the two positions are neighbors in primary sequence. Even though molecular mechanics covalent bonding terms span more than two atoms, e.g. four-atom dihedral angles, each energy term is always a function of the conformation of no more than two positions (made possible by a three-atom long backbone repeat unit), thereby maintaining pairwise additivity.

3.2.4 Initial conformational search

Each protein sequence was explicitly considered, for example, all 1,080 single mutations from 60 CDR positions with 18 amino acids each. Side chain rotamers that clashed, either with backbone atoms or with the side chains of positions not given conformational flexibility, were eliminated from the search space. The dead-end elimination and A* algorithms were used to find the global minimum energy conformation (GMEC) of each sequence and up to 10 kcal/mol of lowest-energy structures [93–99]. To select the single conformation of each sequence to be used for binding energy calculation, low-energy structures were reevaluated and reranked according to more

accurate energy functions. However, because of the extra 10° torsional sampling in the rotamer library, the strictly next-lowest-energy structures from the GMEC are usually not qualitatively different. We used the low-resolution energy function to rank-order structures that only differ by 10° dihedral angle rotations; we used a high-resolution energy function to reevaluate different low-energy side chain conformations that may, for example, exhibit different burial and trade-off between electrostatic interaction and desolvation. The expansion of χ_1 and χ_2 by $\pm$ 10° generally created families of nine rotamers (a "fleximer" [100]); we reevaluated the 30 lowest-energy structures within 10 kcal/mol that contain at least one rotamer from a new rotamer family (contain at least one new fleximer).

An in-house implementation of dead-end elimination and A* (DEE/A*) was at the core of the conformation search [101]. A novel procedure was developed to balance the speed, memory, and disk issues. A fundamental difficulty was that the input to DEE/A* is an energy cutoff from the GMEC specifying the energetic range of structures to be output, rather than upfront specifying the desired number of structures. Furthermore, even if one could specify the number of lowest-energy structures, we actually had to look through an a priori unknown number structures to find the 30 structures of unique fleximers. Our solution was to progressively increase the energy cutoff from near-zero to 10 kcal/mol. For each energy cutoff, the lowest-energy structures were output from DEE/A*. To avoid memory limitations, iterative deepening A* was used, but as a consequence, the structures were not output in order of energy. As each structure was output, it was determined if the structure contained at least one new fleximer as compared to already kept structures. If so, the structure was kept. If not, the structure only replaced the already-kept same-fleximer structure if the new structure was of lower energy. Once all structures within the energy cutoff were output, if 30 unique structures had not been obtained, then the energy cutoff was heuristically increased according to how many more unique structures were needed. The progression was stopped once 30 unique structures were obtained or the energy cutoff reached 10 kcal/mol. By evaluating structures as they were output from DEE/A*, it was not required to have a large amount of disk space for storage of all

lowest-energy structures.

Each execution of DEE/A* requires parameters specifying the schedule of different elimination routines. One can think of the different schedules as falling somewhere on a single axis of increasing/decreasing complexity. Increased scheduling complexity is necessary to solve more difficult conformational search problems, yet yields much less efficient solving of less difficult problems as compared to the time it would take for a less complex schedule. Our solution was to progressively increase the scheduling complexity. After a first attempt with a less-complex schedule, if DEE/A* did not complete within a specified amount of time, the process was aborted and a higher level of scheduling complexity was started with a longer allowed execution time.

3.2.5 Reevaluation of electrostatics

Structures were reevaluated using Poisson–Boltzmann continuum electrostatics as a substitute for the distance-dependent electrostatic energy from the molecular mechanics force field. The PARSE parameters were used for partial atomic charges and radii [62]. A locally modified version of the DELPHI computer program was used to solve the linearized Poisson–Boltzmann equation [58–61]. A dielectric constant of 4 was used for protein and explicit water, and 80 for implicit solvent regions. Ionic strength was set to match experiments at 0.145 M, modeled with a 2.0-Å Stern layer and a molecular surface generated with a 1.4-Å probe sphere. Calculations were performed with two-step focusing from 23% to 92% molecular fill of the grid. All initial calculations used one translation of a $129 \times 129 \times 129$ grid; interesting sequences were reevaluated with 10 translations of a $161 \times 161 \times 161$ grid. The nonpolar component of the solvation free energy was added as a solvent-accessible surface area (SASA) term of 5 cal/mol/Å^2 [62], calculated using the analytical surface area routine in the program CHARMM; the constant intercept of this linear model cancels for two-component binding and referencing energies to wild type. In total, the typical "high-resolution" energy function for rigid binding was comprised of van der Waals, electrostatic, and nonpolar terms. Folding energies include additional molecular mechanics covalent bonding terms for conformational change. The nonpolar term for

binding relative to wild type is typically less than 0.2 kcal/mol in magnitude and rarely as large as 0.5 kcal/mol.

3.2.6 Prediction of binding and folding

A standard calculation yielded prediction of binding affinity relative to wild type, using the predicted bound-state structure and assuming rigid binding, as well as prediction of folding stability relative to wild type, using the same structure and subtracting off energies for isolated, model compounds. The sensitivity of a prediction to the conformation of neighboring side chains was mitigated by matching neighboring side-chain conformations between mutant and wild type where possible. Inclusion of rotamers from the crystal structure for each position was critical; the wild type was given the benefit of the doubt for predicting relative binding free energies. In general, mutations predicted to destabilize the antibody by more than 3 kcal/mol were disregarded. In addition, we modeled conformational change upon binding by approximating the lower-energy unbound state with a second side-chain conformational search. The same positions from the bound-state design were allowed conformational relaxation, again with a fixed backbone but without the binding partner present. Binding free energy was the difference between the bound and unbound states. Finally, the structures for select sequences were subjected to energy-minimization of the designed positions and reevaluation of binding. All calculations for single mutations across 60 positions finished within 24 hours on a 100-CPU cluster.

3.2.7 Prediction of cooperative mutations

Double mutations predicted to be favorably cooperative were distinguished by ensuring that predicted favorability was greater than the predicted energetics for any mutation subset. Double mutation energetics were required to exceed each single mutation as well as the computed sum of each single mutation. Within the design of $18^2 = 324$ sequences, the $18-1 = 17$ single mutations were used for computing cooperativity of the $324-18 = 306$ double mutations. Cooperativity was required to exceed

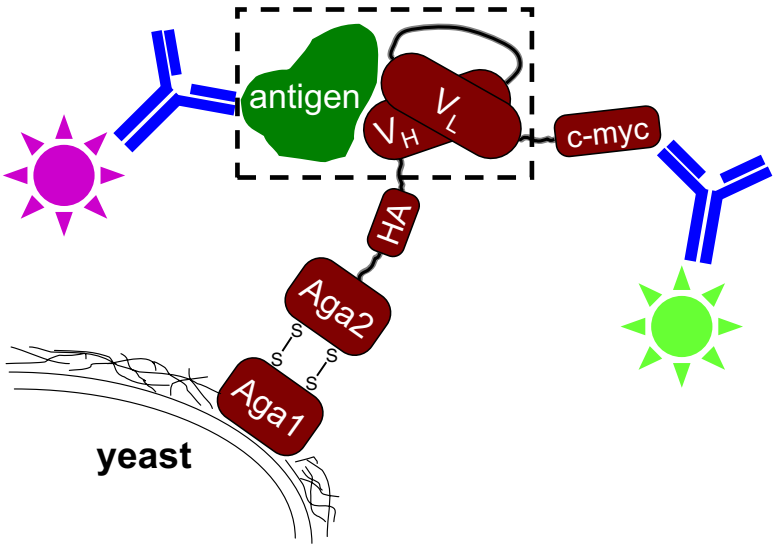


Figure 3.2: Yeast surface display schematic. The dashed box highlights the interaction of interest, the single-chain variable fragment antibody (scFv) with its antigen.

0.2 kcal/mol to avoid marginal positives, e.g. a double mutation at -0.7 kcal/mol was not considered cooperative if the two single mutations were -0.3 kcal/mol each, but would be titled cooperative if the double mutation exceeded -0.8 kcal/mol.

3.2.8 Binding affinity measurements

The single-chain format of D1.3 [4] was displayed on the surface of yeast [102, 103]. A schematic is shown in Figure 3.2. Antibody variants were constructed using site-directed mutagenesis (Stratagene, La Jolla, CA) with oligonucleotide primers from MWG Biotech (High Point, NC). Sequences were confirmed with forward and reverse sequencing (MIT CCR Biopolymers Laboratory).

The yeast strain EBY-100 was transformed with each surface-display plasmid and binding affinities were measured as previously described [104]. Briefly, each titration used 10–16 tubes with equal amounts of antibody-displaying cells and varying lysozyme concentration. Experiments were carried out at 25 °C in phosphate-buffered saline (PBS), pH 7.3–7.5, 0.145 M salt (0.167 M ionic strength). Three or more hours elapsed as binding approached equilibrium. Cells were then pelleted and washed ice-cold, and incubated with secondary reagents for detection of antibody–lysozyme

complexes.

Binding to lysozyme (Sigma Aldrich, St. Louis, MO) was detected through secondary labeling with biotinylated rabbit polyclonal anti-lysozyme antibodies (Research Diagnostics, now Fitzgerald Industries, Concord, MA) and tertiary labeling with streptavidin-phycoerythrin (Invitrogen, Eugene, OR). Polyclonal antibodies that bound yeast nonspecifically were removed prior to use. Though binding was detected of D1.3 to biotinylated lysozyme, neat lysozyme was used for all titrations to avoid any effects of biotinylated lysines near the binding interface.

The equilibrium dissociation constant (K_{d}) was fit using the mean fluorescence of only the fraction of cells that display antibody (MFU_{disp}) as a function of antigen concentration ($[Ag]$). The data were fit to Equation 3.2 as a function of the three free parameters by minimizing the sum of the squared residuals using the Solver tool in the program Excel.

$$MFU_{\mathrm{disp}} = MFU_{\mathrm{min}} + \frac{MFU_{\mathrm{range}}[Ag]}{[Ag] + K_{\mathrm{d}}} \tag{3.2}$$

MFU_{disp} was determined by subtracting the autofluorescence of the nondisplaying cells from the total fluorescence. Data from high concentrations of lysozyme where the displaying fraction was distinctly separated form the nondisplaying fraction were used to measure the nondisplaying fraction (f_{non}) and its mean fluorescence (MFU_{non}), where both values were assumed to be constant within tubes of a titration because the displaying cells were the same. MFU_{disp} was then determined using Equation 3.3.

$$MFU_{\mathrm{disp}} = \frac{MFU_{\mathrm{tot}} - f_{\mathrm{non}} MFU_{\mathrm{non}}}{1 - f_{\mathrm{non}}} \tag{3.3}$$

Data fitting using MFU_{tot} directly instead of MFU_{disp} yields an identical K_{d} value, but MFU_{min} and MFU_{range} values change. Reported K_{d} values use the average and standard deviation from independent experiments, performed with different inoculations of yeast cells and typically on different days.

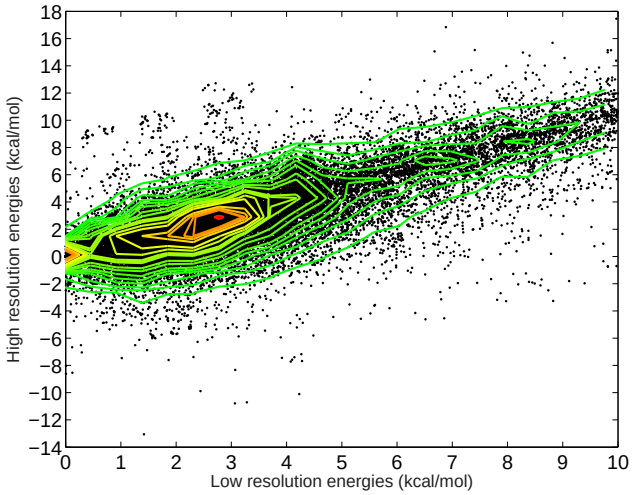


Figure 3.3: Correlation of low- and high-resolution bound-state[a] free energies within a protein sequence. Each equi-density contour encompasses an additional 5% of the data, with the outer contour containing 95% of the data. All sequences and structures designed in D1.3 are shown. Low- and high-resolution free energies are normalized to the corresponding energies of the low-resolution global minimum energy conformation (GMEC). The strong correlation demonstrates that the high-resolution minimum-energy structure for each sequence will very likely be found within the first 10 kcal/mol of low-resolution lowest-energy structures.
[a] Within a sequence, subtracting off model compound energies does not affect rank-ordering.

3.3 Results

First we attempted to redesign the model antibody D1.3 for improved binding to its antigen, hen egg-white lysozyme. The effect of single mutation to side chains excluding proline and cysteine at each of 60 complementarity determining region (CDR) positions was calculated using our two-stage hierarchical procedure.

A first result of the calculations is the correlation between the low- and high-resolution energy functions. Correlation, or rank-ordering, within a sequence is necessary for the lowest-energy structure of each sequence as ranked in the high-resolution energy function to be found by structure reevaluation. As shown in Figure 3.3, correlation within a sequence is strong (correlation coefficient of 0.77). Within the lowest 10 kcal/mol of low-resolution structures, the lowest-energy structure in high-resolution is very likely to be found.

On the other hand, if there was strong low- to high-resolution correlation of bind-

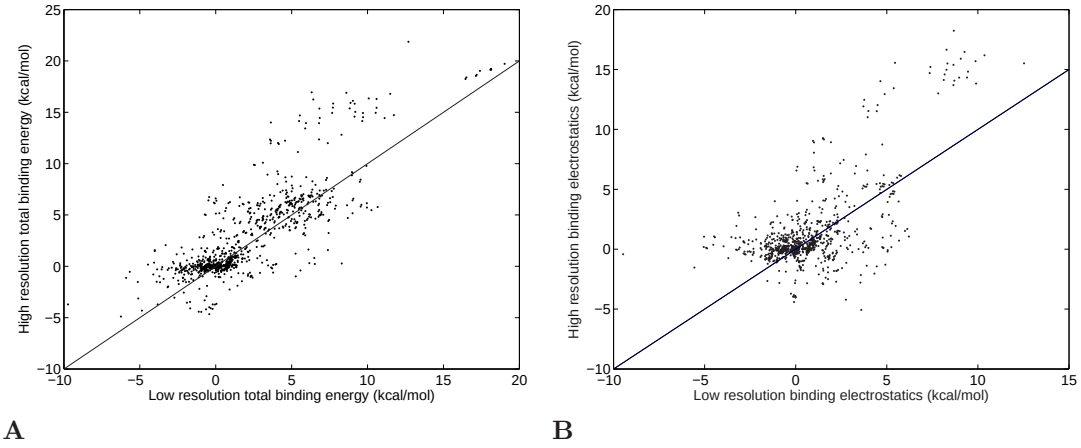


Figure 3.4: Reevaluation of relative binding energies with the high-resolution free energy function. All sequences designed in D1.3 are shown. For each sequence, the structure of lowest low- or high-resolution energy, respectively, is used to calculate the rigid binding energy relative to wild type. Though roughly correlated, the low-resolution energy function lacks quantitative accuracy for prediction within 2 kcal/mol.

ing energies from sequence-to-sequence, then the high-resolution reevaluation stage would be unnecessary. As shown in Figure 3.4, although the data are correlated (0.66 and 0.84 for total and electrostatics-only energies, respectively), there is significant deviation from the $y = x$ line. Each average absolute deviation is 1.2 kcal/mol. More importantly, for sequences with relative binding free energy less than -1 kcal/mol in low-resolution, the average absolute deviation is 1.9 and 2.1 kcal/mol, respectively, and the percentage of sequences that change from less than -1 kcal/mol to greater than 0 kcal/mol is 30% and 49%, respectively. These latter statistics indicate that predictions based on calculations with only the low-resolution energy function would not be reflective of the high-resolution and presumably more accurate energy function.

Figure 3.5 summarizes the predicted effect on binding energy relative to wild type of single mutations in D1.3. Mutations that are predicted to disrupt antibody stability by more than 3 kcal/mol are disregarded and displayed as unfavorable. Figure 3.6 displays the predictions as categorized by the solvent-accessibility of the position. Each position is classified as either interface, periphery, surface, or core. In Figure 3.6C, though the wild-type side chain does not contact lysozyme upon binding, the side

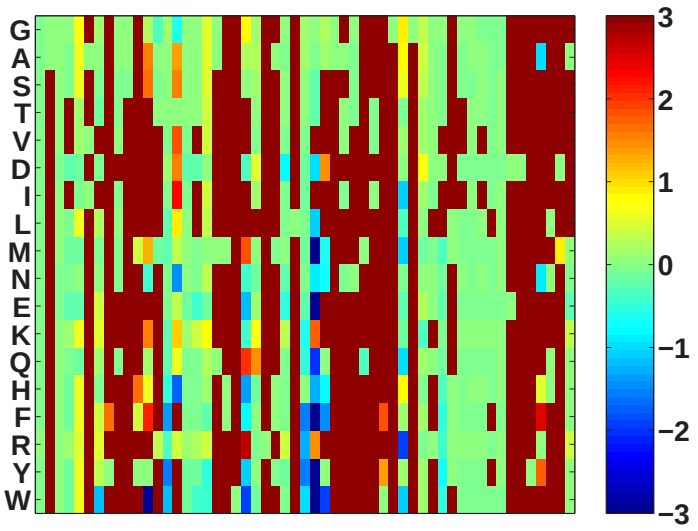


Figure 3.5: Design of single mutations in D1.3. Each column the result of design at one position in D1.3; each row is the effect of mutation to a different amino acid. Binding free energy is relative to wild type in kcal/mol. Mutations predicted to disrupt antibody stability are shown as red, unfavorable.

chains predicted to improve binding often are larger and do contact lysozyme, similar to the positions in Figure 3.6A. Therefore, the majority of mutations predicted to improved binding are at the periphery of the interface.

A striking feature of the single mutation designs is that the predictions for improved binding are dominated by mutations to large amino acids. Figure 3.7 shows the mutations that are predicted to be favorable, arranged by size of amino acid. Lysine and arginine are likely underrepresented because their net positive charge incurs in general less favorable electrostatic interaction with the positively-charged lysozyme.

Many of the predictions of favorable binding from mutation to a large amino acid exhibit improved van der Waals packing interactions outweighing disfavored net electrostatic solvation and interaction. Figure 3.8 displays the structures and energetics of three mutation examples. In all three cases, the antibody position is at the periphery of the binding interface, where fewer packing constraints allow mutation to bigger or smaller side chains, but close proximity to lysozyme facilitates improved interactions. Also, all three cases have significant predicted improvement dominated by improved van der Waals interactions. For the first two cases, the electrostatics is unfavorable.

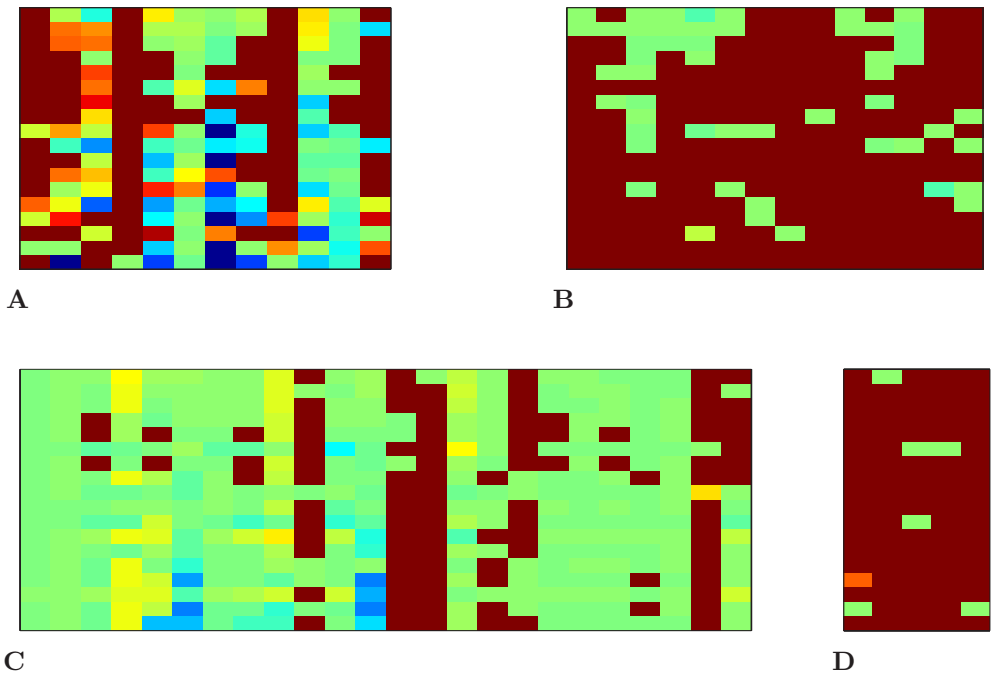


Figure 3.6: Designs arranged by type of position, from Figure 3.5. The solvent-accessibility of the wild-type side chain, before and after binding, was used to classify each position. (A) Periphery of binding interface. (B) Interior antibody core, not solvent-exposed. (C) Surface of antibody, not in contact with lysozyme. (D) Interface, full burial upon binding.

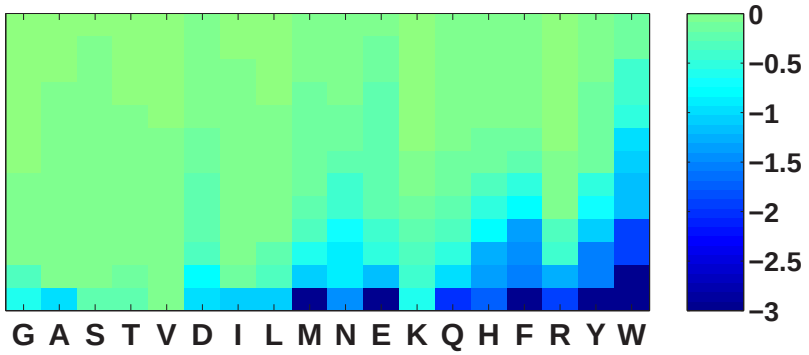


Figure 3.7: Proliferation of large amino acids. Histogram of single amino acid mutations predicted to be favorable from Figure 3.5. Relative binding free energies are in kcal/mol.

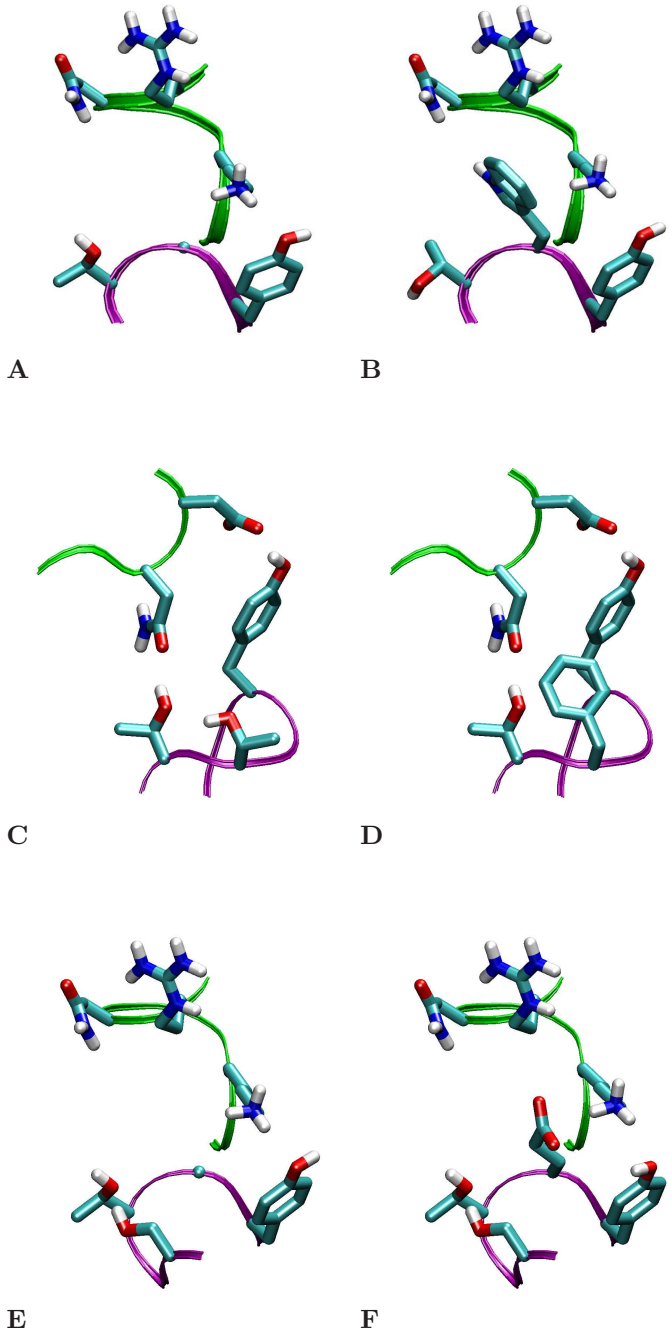


Figure	Mutation	$\Delta\Delta G_i$ (kcal/mol) vdW	nonpolar	elec.	Total
A-B	Gly(H31)Trp	−7.04	−0.56	+2.71	−4.89
C-D	Thr(L52)Phe	−1.49	−0.06	+0.20	−1.34
E-F	Gly(H31)Glu	−3.00	−0.29	−0.42	−3.71

Figure 3.8: Examples of D1.3 single mutation designs. Each pair of figures displays the wild type and a design. Binding energetics are given beneath.

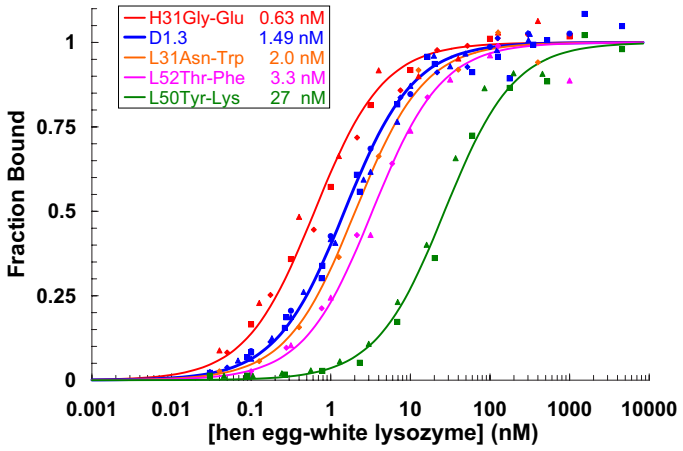


Figure 3.9: Representative experimental binding affinity curves for D1.3 mutations. Each antibody variant is displayed on the surface of yeast and lysozyme is provided in solution. After approaching equilibrium, bound complexes are detected with secondary antibodies and analyzed with flow cytometry.

Our confidence in these designs was low because many had unsatisfied hydrogen bonding, or a predicted packing improvement of a magnitude uncharacteristically large for a single amino-acid substitution. Nevertheless, 17 single mutations, most with calculated improved total binding free energy, were selected for experimental binding affinity measurement. Figure 3.9 shows representative experimental binding affinity data and Table 3.1 displays the predicted and measured energetics for the mutations. The standard deviation for most measurements is approximately 10% of the average value. In the first preliminary design round, the mutations at H32 and L53 were selected based on alignment of results from a 7-site design (data not shown). The second round consisted of positions H28, L32, L46, and L50. The third round consisted of positions H31, H99, L31, and L52. Only three mutations improved binding affinity, with 2.4-fold improvement (-0.51 kcal/mol) for the best mutation, Gly(H31)Glu.

Double mutations were computationally designed in D1.3. Instead of selecting all 1770 pairs of positions (60 choose 2), where most pairs would be out of contact with each other, we selected only the 86 pairs of nearby positions. A comparison of the number and general difficulty of designs between single and double mutations is shown in Figure 3.10. For each pair of positions, $18^2 = 324$ sequences were considered.

Table 3.1: Predicted and experimental D1.3 single mutation binding affinities.

			$\Delta\Delta G^a_{\text{binding}}$			$\Delta\Delta G^a_{\text{folding}}$	
Position	Mut'n	K_d (nM)	Exp.	$\text{Calc}_{\text{total}}$	$\text{Calc}_{\text{elec}}$	$\text{Calc}_{\text{total}}$	$\text{Calc}_{\text{elec}}$
wild type		1.49 ± 0.09					
H32 Tyr	Glu[c]	37	+1.9	−1.52	−1.54	+10.73	+3.20
L53 Thr	Arg	24	+1.6	+0.43	+3.99	−1.32	−0.90
H28 Ser	Asn[b]	1.9 ± 0.2	+0.15	−0.18	−0.19	+0.30	+0.17
H28 Ser	Gln[b]	1.70 ± 0.06	+0.08	+0.03	+0.01	−1.26	+0.15
H28 Ser	Asp	1.49 ± 0.11	+0.00	−0.75	−0.77	−0.70	−0.40
H28 Ser	Glu	1.25 ± 0.15	−0.11	−0.56	+0.09	−1.38	+0.44
L32 Tyr	Trp[c]	$>20\mu$M	$> +5$	+18.58	+1.14	+15.03	+0.07
L46 Leu	Asp[c]	$>20\mu$M	$> +5$	−0.24	−0.28	+11.38	+2.78
L46 Leu	Glu[c]	$>20\mu$M	$> +5$	−0.38	−0.38	+11.50	+6.89
L50 Tyr	Arg[c]	6.2 ± 1.8	+0.85	−1.55	+0.76	+7.29	+0.77
L50 Tyr	Lys[c]	27 ± 7	+1.71	+1.52	+0.67	+0.68	+1.46
H31 Gly	Ala[b]	2.5 ± 0.3	+0.30	−0.07	−0.15	−1.47	+0.33
H31 Gly	Glu	0.63 ± 0.13	−0.51	−3.71	−0.42	−0.46	+0.66
H31 Gly	Trp	1.5 ± 0.8	+0.00	−4.89	+2.71	−0.04	+1.37
H58 Asp	Asn	1.2 ± 0.2	−0.13	−0.16	−0.13	−0.08	−0.44
H58 Asp	Glu	1.7 ± 0.4	+0.09	−0.25	−0.21	+0.05	−1.12
H99 Arg	Tyr[c]	12	+1.2	−2.19	−1.87	+15.79	+1.17
H99 Arg	Trp[c]	4.9	+0.7	−2.91	−0.35	+1.86	+0.87
L31 Asn	Trp	2.0	+0.2	−1.22	+1.25	−1.38	+0.25
L52 Thr	Phe	3.3 ± 0.5	+0.48	−1.34	+0.20	−1.41	+0.68

[a]All values in kcal/mol.
[b]Control mutation to test shape change away from wild type.
[c]Excluded from Figure 4.5, comparison of binding energetics, due to unfavorable calculated $\Delta\Delta G_{\text{folding}}$.

The difficulty of obtaining lowest-energy conformations of each sequence is related to the number of protein positions and crystallographic water molecules allowed conformational flexibility. Predictions of improved binding affinity were filtered to select cooperative double mutations with energy improvement exceeding either single mutation or their additive sum. The only double mutations computed to be cooperative exhibited a trade-off of improved van der Waals interactions at the cost of disfavored electrostatic interactions.

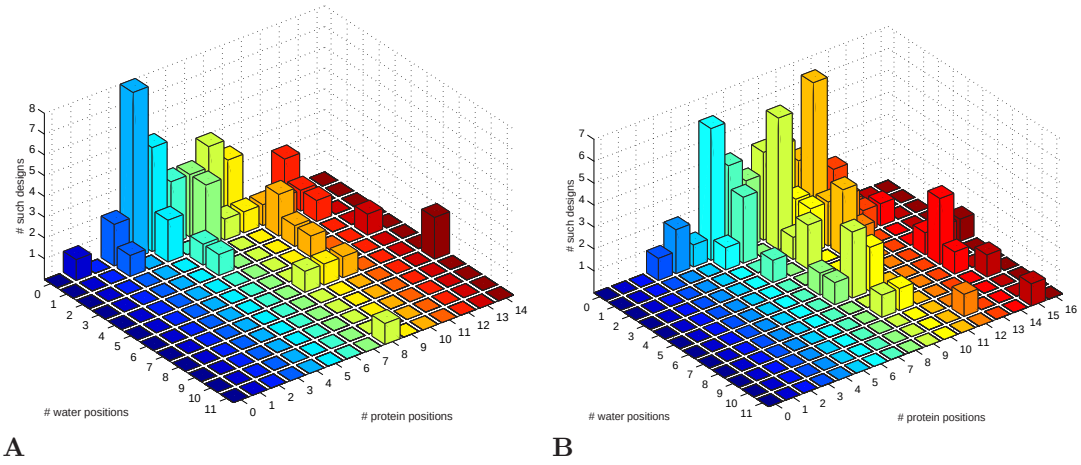


Figure 3.10: Number and difficulty of single and double mutation designs. Each design is characterized by the number of positions, water and protein, with conformational flexibility.

3.4 Discussion

Our goal for protein design is to predict amino-acid mutations that enhance protein–protein interactions. As an inverse design problem, we only need to find a subset of the likely many possible solutions, and it is not necessary to accurately predict the effect of all possible mutations. Predictions of unfavorability may be inaccurate as a result of conformational search approximations, such as a rigid protein backbone and discrete side chain rotamers, limiting the identification of low-energy states. Our systemic exploration of mutations throughout an interface is a tool to find mutation opportunities and not an attempt to recapitulate all possible effects on binding.

A key challenge in computational protein design is overcoming large conformational search complexity while evaluating atomic interactions with an accurate, physics-based energy function. We have addressed this problem with a two-stage hierarchical search procedure that takes advantage of the deterministic and powerful dead-end elimination and A* search algorithms in combination with the well-validated Poisson–Boltzmann continuum electrostatic model. Figures 3.3 and 3.4 illustrate the validity and utility of this approach. Correlation of the low- and high-resolution energy functions is necessary within a sequence and is demonstrated in Figure 3.3, but

at the same time, Figure 3.4 shows that the high-resolution reevaluation provides critical, new energetic information.

A main finding of this work is that predictions for improved binding are dominated by mutations to large amino acids (Figure 3.7), many of which exhibit improved van der Waals packing interactions outweighing disfavored net electrostatic solvation and interaction. Intuitively, a prediction for improvement is more robust if each term in the energy function is equal or favorable relative to wild type, as opposed to overall favorability as a consequence of a trade-off between individual energy terms. Moreover, examination of predicted structures made us less confident in many of the designs; qualitative hydrogen-bond analysis supported the calculated unfavorable electrostatics, but the calculated favorable van der Waals energies were less intuitive. Nevertheless, there was no fundamental reason why such predictions would be incorrect, and we therefore experimentally measured the binding affinity for a variety of designed mutations (Table 3.1).

The experimental measurements confirmed our concerns for the unreliability of predictions based on a calculated trade-off of improved van der Waals interactions at the cost of disfavored electrostatics. In addition, experiments highlighted the importance of the prediction of antibody folding stability.

Mutations at position H31 Gly provide an example of the importance of binding electrostatics. Glutamate and tryptophan are among many side chains predicted to be very favorable, mostly as a result of improved van der Waals interactions (Figure 3.8). First, experimental mutation to alanine confirmed that substitution away from the flexible glycine at the tip of the CDR H1 loop was not catastrophic, but did result in a +0.30 kcal/mol affinity loss. Next, we find that glutamate and tryptophan are not substantially improved, but instead the tryptophan is the same as wild type and the glutamate is 2.4-fold improved (-0.51 kcal/mol) relative to wild type. Only for the glutamate was the electrostatics of binding predicted to be favorable.

Mutation from aspartate to asparagine at position H58 further illustrates the potential capabilities of binding predictions based on electrostatics. Though the improvement was very modest (-0.13 kcal/mol), it was in agreement with the calculated

electrostatics (−0.13 kcal/mol). This type of mutation isolates electrostatic effects as both side chains are the same size and shape.

Two mutations were chosen to directly test the hypothesis that prediction of improved van der Waals interactions is unreliable. For Asn(L31)Trp and Thr(L52)Phe, the total binding energy is predicted to be improved (−1.22 kcal/mol and −1.34 kcal/mol, respectively), dominated by improved van der Waals interactions (−2.24 kcal/mol and −1.49 kcal/mol, respectively). The electrostatics for each is predicted to be unfavorable by differing degrees (+1.25 kcal/mol and +0.20 kcal/mol, respectively). Experimentally, neither mutant binds better than wild type (+0.2 kcal/mol and +0.48 kcal/mol, respectively).

At position H28 Ser, calculations indicated that either negatively-charge side chain would be favorable as a result of electrostatic interactions with lysozyme, particularly Lys 116. However, removal of the serine breaks an intramolecular hydrogen bond with backbone. Both asparagine and glutamine, mutations designed to test the loss of serine while remaining polar and neutral, slightly decrease binding affinity, yet asparate and glutamate, which morph isosterically from neutral to negatively-charged, recover binding affinity. The solvent-exposure of this site should negate effects of the dipole switch at the amine. This example adds evidence to support the role of antibody stability and electrostatics in binding energy prediction.

Insufficient conformational search in the bound state is unlikely to be the cause of the unreliable van der Waals predicted improvements. The typical inaccuracy is for mutation from a small amino acid (e.g. Gly, Ser, Thr) to a large amino acid (e.g. Phe, Trp). The fixed backbone and discrete rotamer approximations may result in a sub-optimal conformation for the designed larger amino acid, with additional conformation search or minimization leading to a lower energy state. However, to fix the scoring problem, energies would have to become less favorable. A more likely explanation is that improvements in interactions in the bound state by larger amino acids are realized, though these increases coincide with improvements in interactions with solvent in the unbound state and modeled insufficiently.

We found that improvements from larger amino acids were mostly not realized,

the calculated Poisson–Boltzmann continuum electrostatics term of binding seemed to be a better predictor for improvement, and that avoiding potentially destabilizing mutations was important. This led to two questions: Can electrostatics-based predictions alone be used to design binding affinity improvements in other antibodies (Chapter 4)? Is there a physics-based explanation for the seemingly inaccurate calculated packing improvements (Chapter 5)?

Chapter 4

Computational design of antibody improvement beyond *in vivo* maturation

Abstract

In Chapter 3 we presented novel methodology for the computational design of protein–protein interactions. Although the overall hierarchical approach was validated, the success rate for improved single mutations was low and affinity enhancements were small. Results indicated that selecting mutations based on the calculated Poisson–Boltzmann continuum electrostatic term for binding would yield improved results. Here we implement that idea and present designed improvement in affinity by over two orders of magnitude using iterative computational design. Four designed mutations in the anti-lysozyme model antibody D44.1 were combined to improve affinity from 4.4 nM to 43 pM, and subsequent design added a pair of mutations for net 140-fold improvement to 30 pM. We also combined three designed mutations in the clinically-approved antibody cetuximab (Erbitux) for 10-fold improvement to 52 pM, and validated predictions in two additional systems. A greater than 60% experimental success rate for single mutations was achieved using design criteria that focus on electrostatics. These results demonstrate novel computational capabilities and indicate their applicability for enhancing and accelerating development of reagents and therapeutics.

4.1 Introduction

The robust design of mutations to improve protein–protein binding affinity using computational methods remains a challenge. Success would directly impact biotechnology by accelerating the development time of reagent and therapeutic molecules, and possibly opening up the development of novel molecules inaccessible to competing technologies. The current technologies for affinity maturation are dominated by selection methods using either an *in vivo* immune response or *in vitro* directed evolution [2]. Computational techniques, while demonstrating significant progress in many applications, have not provided reliable results, that is, they have not consistently produced a list of predicted mutations that yield both a high success rate and an overall large magnitude effect [39].

The computational protein design methods developed in Chapter 3 produced marginal affinity enhancements in the anti-lysozyme model antibody D1.3. The results indicated that a revised design protocol, selecting mutations based on the calculated Poisson–Boltzmann continuum electrostatic term for binding, would improve the success rate. Here we investigate a refined computational method for designing affinity improvements in several new antibody systems. Results for computational prediction and experimental validation across different protein complexes should broaden the applicability of our conclusions for future work.

4.2 Results

In our second design attempt, we used only the electrostatic term of the computed binding free energy to predict improvements in affinity, but kept the original design procedure and full energy function for side-chain conformational search. Since there were few computed opportunities for improving D1.3 based on electrostatics, new antibodies were chosen to explore both the new method and the possibility that D1.3 is anomalous. Our second target for redesign was the antibody D44.1; like D1.3, it binds lysozyme, facilitating experiments, but its epitope is different from that of

Table 4.1: Predicted and experimental D44.1 single mutation binding affinities.

Position	Mut'n	K_d (nM)	Ratio	$\Delta\Delta G_{binding}$ (kcal/mol) Exp.	$Calc_{tot}$	$Calc_{elec}$
wild type		4.4 ± 0.5				
Selected based on $Calc_{elec}$						
L32 Asn	Gly	1.03 ± 0.12	4.2 ± 0.5	-0.86 ± 0.07	+0.99	−1.03
L92 Asn	Ala	0.53 ± 0.09	8.3 ± 1.3	-1.25 ± 0.10	−0.02	−0.86
H28 Thr	Asp	3.39 ± 0.15	1.29 ± 0.06	-0.15 ± 0.03	−0.30	−0.26
H31 Thr	Ala	9.4 ± 1.4	0.47 ± 0.07	$+0.45 \pm 0.09$	+0.31	−0.97
H31 Thr	Val	11 ± 2	0.41 ± 0.09	$+0.53 \pm 0.13$	−2.06	−0.36
H57 Ser	Ala	2.3 ± 0.6	1.9 ± 0.5	-0.37 ± 0.16	−0.59	−1.44
H57 Ser	Val	1.9 ± 0.4	2.3 ± 0.4	-0.49 ± 0.12	−2.29	−1.34
H58 Thr	Asp	1.70 ± 0.10	2.58 ± 0.16	-0.56 ± 0.04	−0.59	−0.53
H65 Lys	Asp	4.5 ± 0.5	0.97 ± 0.12	$+0.02 \pm 0.07$	−0.38	−0.42
Selected based on $Calc_{tot}$						
L32 Asn	Tyr	4.4 ± 1.0	1.0 ± 0.2	$+0.00 \pm 0.13$	−0.29	+2.59
H31 Thr	Trp	5 ± 2	0.8 ± 0.3	$+0.13 \pm 0.2$	−3.75	+1.56

D1.3. In addition, D44.1 has low nanomolar affinity, maintaining the challenge of nanomolar to picomolar affinity maturation.

Single mutations were designed at all D44.1 non-proline CDR positions and then mutations were ranked by electrostatic binding free energy term. In contrast to the results for the D1.3 design, there were many computed opportunities for electrostatic improvement. Binding affinity was measured for the nine largest-magnitude predictions, choosing no more than two mutations per position. Six out of the nine mutants were found to bind tighter than wild type, and the best mutation, Asn(L92)Ala, exhibits 8-fold improvement (Table 4.1). In addition, following the original design procedure, two mutations predicted to increase packing interactions were tested; as expected, neither mutation led to an improvement in affinity (Table 4.1; this is further investigated in Chapter 5).

The successful D44.1 single mutations were combined and tested experimentally. The six favorable mutations span five positions, where L32 and L92 are the only positions in direct contact with each other in the wild-type crystal structure. Contrary to prediction, this double mutant is not as improved as the L92 single mutant

Table 4.2: Predicted and experimental D44.1 combination mutations.

Mutations[a]	K_d (nM)	Ratio	$\Delta\Delta G_{binding}$ (kcal/mol) Exp.	$Calc_{tot}$	$Calc_{elec}$
wild type	4.4 ± 0.5				
(L32)G, (L92)A	1.37 ± 0.12	3.2 ± 0.3	−0.69 ± 0.05	+1.23	−1.68
(H57)V, (H58)D	0.83 ± 0.06	5.3 ± 0.4	−0.99 ± 0.04	−2.88	−1.87
(L32)G, (L92)A, (H57)V, (H58)D	0.23 ± 0.02	18.7 ± 1.9	−1.74 ± 0.06	−1.65	−3.55
(L92)A, (H57)V, (H58)D	0.076 ± 0.007	57 ± 5	−2.40 ± 0.05	−2.90	−2.73
(L92)A, (H28)D, (H57)V, (H58)D[b]	0.0430 ± 0.0013	102 ± 3	−2.74 ± 0.02	−3.18	−2.96

[a]Shorthand is (position) and new amino acid, e.g. (L32)G is V_L32 mutated to glycine.
[b]Referred to as "Quad".

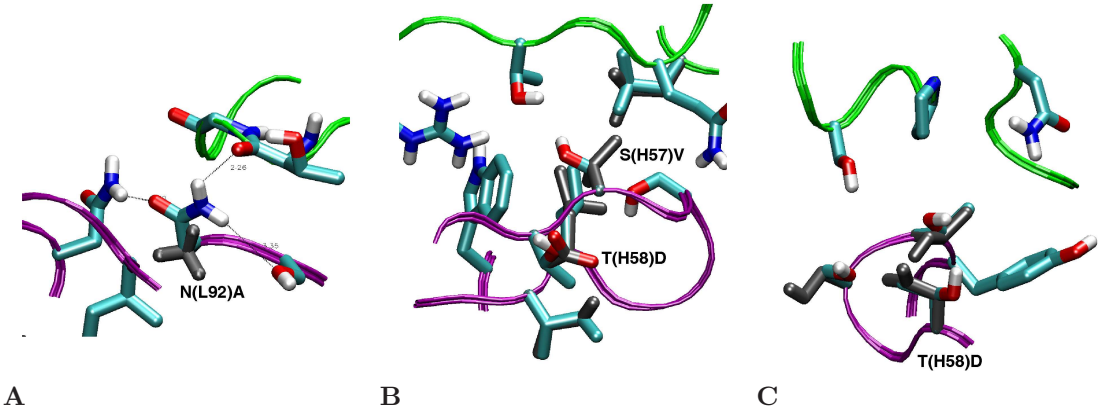


Figure 4.1: Predicted structures for D44.1 mutations. green ribbon: lysozyme backbone; magenta ribbon: antibody backbone; atom color: wild type; grey: design. (A) Asn(L92)Ala. (B) Ser(H57)Val, top; Thr(H58)Asp, bottom. (C) Thr(H28)Asp.

alone. The mutations at the other three positions are additive with L92, yielding a quadruple mutant with 43-pM affinity, 100-fold improved over wild type (Table 4.2). Figure 4.1 displays the predicted interactions for these four mutations. Measurements of dissociation and association rate constants for D44.1 and the high-affinity quadruple mutant reveal that both kinetic steps were improved, including a 23-fold slower off-rate (Table 4.3). Each ratio of k_{off} to k_{on} is in agreement with the independently measured K_d. Increasing ionic strength from 167 mM to 1.67 M screens the designed quadruple-mutant interaction by approximately 4-fold, in agreement with calculation (+0.9 kcal/mol), whereas the wild type is marginally improved at high salt, in constrast to calculation (+0.3 kcal/mol; Table 4.3).

Table 4.3: Characterization of D44.1 quadruple mutant.

	k_{off} (10^{-3} s^{-1})	k_{on} (10^6 $M^{-1}s^{-1}$)	k_{off}/k_{on} (nM)	K_d (nM)	High-salt K_d (nM)
D44.1	10.7 ± 1.0	2.52 ± 0.13	4.3 ± 0.5	4.4 ± 0.5	3.0 ± 0.8
Quad	0.466 ± 0.014	8.4 ± 1.7	0.055 ± 0.012	0.0430 ± 0.0013	0.18 ± 0.08

Two sets of calculations were made in an attempt to further improve the high-affinity D44.1 quadruple mutant. First, single mutations were designed based on the predicted structure of the quadruple mutant. These calculations reiterated many predictions seen in the original D44.1 design that were lower-ranked and not tested experimentally, as well as predictions for mutating position L32, which neighbors the Asn(L92)Ala mutation and was found experimentally to not be additive. Second, double mutations were designed at all 93 pairs of contacting positions, and triple mutations were designed at positions H32, H98, and H100 based on cooperative packing among these three positions. The double- and triple-mutant designs were filtered for favorable, cooperative predictions, requiring the double or triple mutation to exceed each single mutation and their energetic sum. Four single mutations, three double mutations, and one triple mutation were selected for experimental testing in the quadruple-mutant context (Table 4.4). Only the H35+H99 double mutant was improved relative to the quadruple mutant, at 30-nM affinity, 140-fold improved over wild type. Measurement of the individual H35 and H99 mutations revealed high cooperativity (Table 4.5), consistent with the predicted salt-bridge and hydrogen-bond rearrangements (Figure 4.2). Experimental binding affinity titrations are shown in Figure 4.3. The diminished success rate of this subsequent design round may be due to smaller magnitude predictions, antibody destabilization, or design in a modeled rather than experimentally-determined structure.

The improved double mutation in the context of the quadruple mutant, a sextuple mutant overall, is predicted to exhibit a large salt-bridge and hydrogen-bond rearrangement. The redesigned region of the antibody forms a small binding pocket for an arginine of lysozyme, using two glutamate side chains, one arginine, two tryptophans, one tyrosine, and a crystallographic water molecule. The design moves a

Table 4.4: Second round design in D44.1 quadruple mutant.

Mutation	K_d (nM)	$\Delta\Delta G_{binding}$[a] Exp.	Calc$_{tot}$	Calc$_{elec}$	$\Delta\Delta G_{folding}$[a] Calc$_{tot}$	Calc$_{elec}$
Quad	0.0430 ± 0.0013					
(L32)L	0.056 ± 0.002	+0.16	−1.37	−0.81	+0.26	−0.28
(H30)G	0.082 ± 0.006	+0.38	+0.09	−0.26	+0.14	−1.39
(H59)E	0.37 ± 0.07	+1.28	+0.49	−0.49	+3.45	+0.97
(L91)N	> 100	> +4	−2.00	−1.06	−0.67	+1.46
(L91)N, (L93)A[b]	0.29 ± 0.04	+1.13	−1.91	−1.10	−2.10	+0.39
(H98)M, (H100)Q	2.2 ± 0.2	+2.33	−0.57	−0.77	+0.30	−2.83
(H32)R, (H98)S	> 100	> +4	−2.46	−0.66	+10.86	−2.80
(H35)S, (H99)D	0.030 ± 0.003	−0.21	−4.30[c]	−4.69[c]		
(H32)W, (H98)V, (H100)Q[c]	8.1	+3.1				

[a]All values in kcal/mol, relative to the Quad mutant.
[b](L93)A mutation introduced to remove inadvertant glycosylation site from (L91)N.
[c]After minimization of design positions.
[d]Selected based on qualitative structure examination.

Table 4.5: Double mutant cycle added to D44.1 quadruple mutant.

Mutant	K_d (nM)	$\Delta\Delta G_{exp}$[a]	Ratio
Wild type	4.4 ± 0.5		
Quad	0.0430 ± 0.0013	-2.74 ± 0.02	102 ± 3
Quad+(H35)S	> 100	> +2	< 0.04
Quad+(H99)D	3.1 ± 0.4	-0.20 ± 0.07	1.41 ± 0.17
Quad+(H35)S+(H99)D ("Hex")	0.030 ± 0.003	-2.95 ± 0.07	145 ± 17

[a]All values in kcal/mol, relative to the D44.1 wild type.

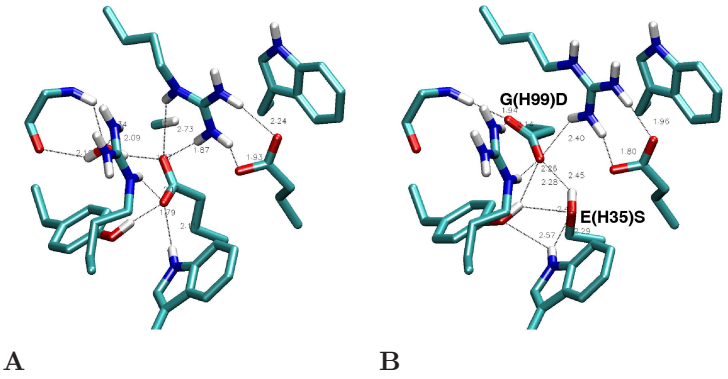


Figure 4.2: Predicted structure for D44.1 cooperative double mutation. (A) Wild type in region of double mutation. The top-center arginine is from lysozyme and all other residues are from the antibody. (B) Designed double mutation: Gly(H99)Asp in back, Glu(H35)Ser in front. The Asp is predicted to displace a crystallographic water molecule

negatively-charged side chain from the H1 loop to the H3 loop and places a serine side chain to complete intramolecular hydrogen bonding. The glutamate to aspartate mutation maintains interaction with the lysozyme arginine and an intramolecular arginine, but at a new geometry in each case. The new aspartate is also predicted to replace crystallographic water 743 and interact directly with the protein backbone amine at position H101 in its place. The design maintains an equivalent number of non-hydrogen atoms overall.

Next, we applied our electrostatics-based methods to the therapeutic antibody cetuximab (Erbitux), which binds epidermal growth factor receptor (EGFR) to block ligand binding [105–107]. Calculations revealed nine positions with opportunities for affinity-enhancing single mutations. The five mutations of largest magnitude were selected for experimental testing in the single-chain antibody format, with one mutation per position, and no two positions in close proximity. The EGFR extracellular domain (EGFR-ECD) mutant used in the assays [108] contains the point mutation Ser(418)Gly that is directly across the interface from the designed Asn(H56)Ala mutation, likely interfering with the prediction. Three of the other four mutants bind EGFR tighter than does cetuximab; there three mutations were combined to produce a triple mutant with 10-fold overall improvement, from 490 pM to 52 pM (Table 4.6,

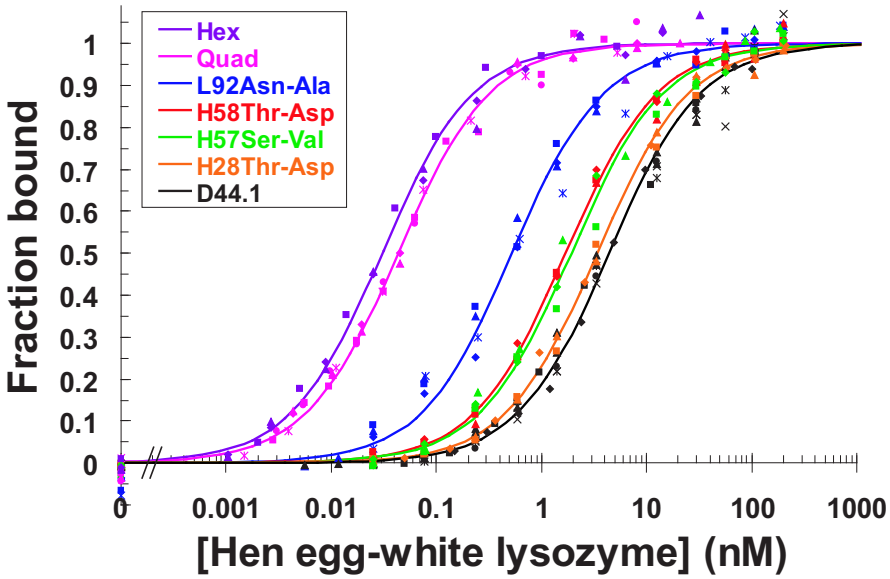


Figure 4.3: Designed high-affinity mutations in D44.1. For each variant, different symbol shapes distinguish independent measurements. From right to left: wild type; each of four single mutations used in Quad; Quad (100-fold); Hex (140-fold).

Figure 4.4).

The transferability and utility of these methods were further demonstrated by designs for which published data validates the predictions. One of our test cases was the antibody 4-4-20, which binds its small-molecule hapten, fluorescein, with 1.2 nM affinity, and was previously engineered using directed evolution to yield the antibody 4M5.3 with over 1,000-fold affinity improvement and 14 mutations [8]. Our designed single mutations in 4-4-20 revealed opportunities for improvement based on computed electrostatics at nine positions. Two of the predicted mutations, Asp(H31)His and Ser(H101)Ala, have already been shown to improve binding affinity as single mutations in 4M5.3 [7]. Interestingly, the design missed the Tyr(H102)Ser mutation in 4M5.3 due to the rigid backbone constraint. Another test case was bevacizumab (Avastin), a therapeutic antibody that binds vascular endothelial growth factor (VEGF) [109, 110], where we found five positions suitable for electrostatics-based improvement. The mutation Thr(H28)Asp and mutations at H31 and H101 are found in a published high-affinity variant of bevacizumab [111]. Our other electrostatics-based predictions in both test systems remain to be tested.

Table 4.6: Predicted and experimental cetuximab binding affinities.

Position	Mutation	K_d (nM)	Ratio	$\Delta\Delta G_{binding}$(kcal/mol) Exp.	$Calc_{tot}$	$Calc_{elec}$
wild type		0.49 ± 0.06				
L26 Ser	Asp	0.36 ± 0.17	1.4 ± 0.7	-0.2 ± 0.3	−0.40	−0.38
L31 Thr	Glu	0.20 ± 0.07	2.4 ± 0.8	-0.53 ± 0.19	−0.50	−0.48
L93 Asn	Ala	0.14 ± 0.06	3.5 ± 1.4	-0.7 ± 0.2	+1.06	−0.75
H56 Asn	Ala[a]	0.46 ± 0.16	1.1 ± 0.4	-0.04 ± 0.2	+2.25	−0.27
H61 Thr	Glu	0.44 ± 0.13	1.1 ± 0.3	-0.07 ± 0.17	−0.77	−0.77
Combinations of single mutations						
L31E, L93A		0.12 ± 0.07[b]	4 ± 2^b	-0.8 ± 0.3^b	+0.51	−1.28
L26D, L31E, L93A		0.052 ± 0.002	9.5 ± 0.3	-1.33 ± 0.02	+0.13	−1.64
L26D, L31E, L93A, H56A, H61E		0.15 ±0.13[b]	3 ± 3^b	-0.7 ± 0.5^b	+1.69	−2.61

[a]Complicated by the Ser(418)Gly mutation in the EGFR-ECD variant.
[b]Unusually large errors, by percent.

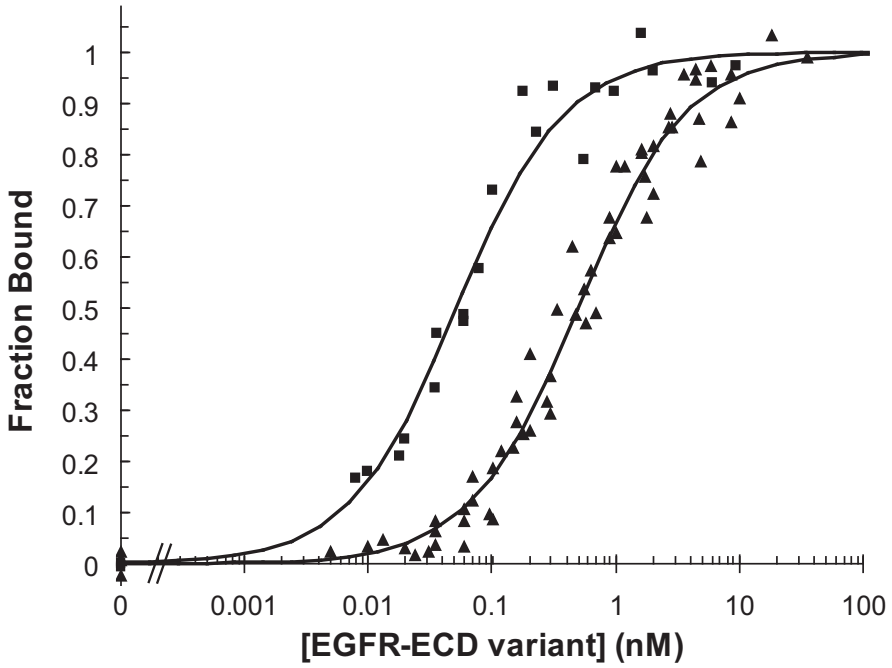


Figure 4.4: Designed high-affinity cetuximab mutant. Triangles: wild type cetuximab; squares: 10-fold improved triple mutant.

4.3 Discussion

The single mutations designed and experimentally validated in this work lead to enhanced binding affinity through one of two electrostatic mechanisms. In one mechanism, the removal of a poorly-satisfied polar group, a polar residue calculated to lose more free energy from desolvation than is recovered by interaction is mutated to a hydrophobic residue. These mutations account for the majority of improved energetics. In the second mechanism, the addition of a charged residue, net charge is changed to increase electrostatic interaction, often at the periphery of the antibody–antigen interface where desolvation is minimal. Unlike previous work using electrostatics to guide design [112, 113], these methods explicitly model the mutation, calculate a binding free energy relative to wild type, include positions that are partially or fully buried upon binding, and avoid opportunities where the mutation is predicted to destabilize the mutant protein. We find that within electrostatics-based predictions, it is equally important to consider depolarizing mutations, to reduce desolvation penalty, as it is to consider charged mutations that increase favorable interactions.

This work presents a computational alternative to directed evolution for affinity maturation. Directed evolution is adept at accumulating successive, additive mutations, but (with the exception of large-scale shuffling) is less well suited for selecting variants whose encoding DNA is further from wild type . Experimental libraries generated using error-prone PCR generally do not cover all single amino-acid mutations, let alone all pairs or greater combinations of mutations, as 13 of the 19 possible single mutations require more than one base pair change, on average. Some classes of mutation require two base pair changes — for example, mutation to either negatively-charged residue from any codon of 10 of the 18 candidate side chains. Of the 12 single mutations found to improve D1.3, D44.1, or cetuximab, 10 required two base pair changes and would therefore have been significantly more difficult to identify by experimental directed evolution. Also, the H35+H99 cooperative double mutation required concerted amino-acid mutation and three total base-pair changes. Computation has the capability to search a vastly larger space than accessible to ei-

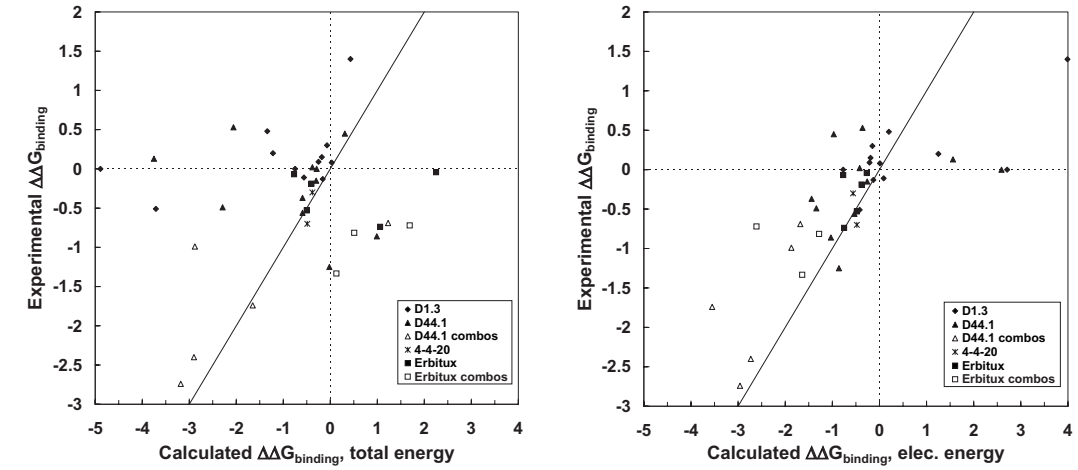


Figure 4.5: Comparison of calculated and experimental binding energetics. Energies are in kcal/mol relative to wild type, with $y = x$, $y = 0$, and $x = 0$ lines shown for clarity. Closed symbols are used for all single mutation data with a different shape for each system (D1.3 included from Chapter 3). Open symbols are used for combination mutants. (A) Calculated total free energy of binding relative to wild type. (B) Calculated electrostatic free energy term of binding relative to wild type.

ther *in vivo* maturation or experimental selection techniques, potentially discovering larger and more beneficial evolutionary steps.

Our results present several design lessons. We find that computed electrostatics alone is a better predictor for improved binding than is computed total free energy. Electrostatics-based predictions yielded fewer false positives, additional true positives, and a greater than 60% success rate for single mutations from wild type (Figure 4.5). Initial results in Chapter 3 demonstrated that predictions based on improved total free energy were dominated by mutations to larger amino acids that did not experimentally improve affinity; incorporation of improved nonpolar hydration models may improve accuracy of predicted packing changes (Chapter 5) [114–116]. Currently, predictions based on total free energy are plagued by mutations that are measured to be equal or worse than wild type, despite favorable calculation. Electrostatics does not agree well for predictions of unfavorability, but this is not our goal for designing improved affinity.

In addition, we find that designing single and double mutations allows for in-

depth conformational search and avoids having any particular design flaw spoiling all results. This work shows that it is possible to achieve large overall affinity improvements (1–2 orders of magnitude) by combining individual mutations that are predominantly additive, consistent with the conclusion in Chapter 2 that the 1800-fold affinity improvement was the result of the cumulative effect of many small changes. Mutations in close proximity to each other, such as the L32 and L92 pair, may not be additive, supporting additional work for directly predicting double and larger combination mutations, such as the cooperative and improved H35+H99 double mutation. Also, avoiding destabilizing mutations was important and crystal structure resolution did not have a significant effect, as the D44.1 and cetuximab structures are of 2.5- and 2.8-Å resolution, respectively, whereas the D1.3 structure is 1.8-Å resolution. Calculations in D1.3 showed few opportunities for electrostatics-based improvement, and accumulated evidence indicates that D1.3 is the anomaly, possibly because of the combination of the many large side chains and buried water molecules at the antibody–antigen interface.

Our results demonstrate novel capabilities for improving protein binding affinity using computational design. Maturation of the model system D44.1 by 140-fold to 30 pM, maturation of the therapeutic antibody cetuximab by 10-fold to 52 pM, and identification of known mutations in 4-4-20 and bevacizumab together indicate that our method is a significant advance for antibody design and should be effective for other antibodies and protein interactions.

4.4 Methods

4.4.1 Structure preparation

The following crystal structures were used from the Protein Data Bank: D44.1: 1MLC [117], Erbitux/cetuximab/IMC-C225: 1YY9 [106], 4-4-20: 1FLR [50], Avastin/bevacizumab/Fab-12: 1BJ1 [118]. Only the variable region (Fv) of each antibody and appropriate domain(s) of its binding partner were kept for calcula-

tions, patching exposed chain termini as either an acetylated N-terminus or an N-methylamide C-terminus. Most crystallographic water molecules and ions were removed, except for water molecules bridging the binding interface or buried away from bulk solvent. Titration states, multiple occupancies, and asparagine, glutamine, and histidine carbon/nitrogen/oxygen crystallographic uncertainties were resolved based on optimization of hydrogen-bonding in the side-chain local environments. Hydrogen-atom positions were assigned using the HBUILD facility [51] in the program CHARMM [52] with the PARAM22 all-atom parameter set [90] and the CHARMM-adapted TIP3P water model. For the fluorescein in the 4-4-20 complex, appropriate PARAM22 atom types were chosen and the partial atomic charges determined previously in Chapter 2 through fitting quantum-mechanical potentials were used.

4.4.2 Design of mutations

The methods developed in Chapter 3 were used to design antibody mutations. At the end of the procedure, mutation predictions were sorted to select those with an improved electrostatic binding free energy term. We eliminated by hand mutations that opened up significant cavities at the interface, as a continuum dielectric 80 model may be inadequate in a confined space. The removal of mutations predicted to lose significant folding or binding van der Waals energy is usually sufficient to identify these positions. One could implement a surface area method to automatically detect and eliminate cavity creation.

4.4.3 Yeast surface display constructs

The single-chain formats of D44.1 and Erbitux/cetuximab/IMC-C225 [119] were displayed on the surface of yeast [102]. The D44.1 gene was designed to match the sequence in the crystal structure [117], with a linker inserted between the heavy and light chains, and codon-optimized for yeast (DNA 2.0, Menlo Park, CA). The linker was based on the conventional $(Gly_4Ser)_3$, but modified to include 5' and 3' unique restrictions sites with conservative amino-acid changes (see Table 4.7 for details). The

gene was inserted into the pCTCON2 vector [104], a variant of pCTCON [120] with reduced genetic homology between the N-terminal and typical inter-chain glycine-rich linkers. The D44.1 journal article and structure file disagree at position H112 between a valine and a leucine, respectively. All work here used a leucine at H112, though mutation to valine was tested and did not alter binding affinity. Antibody variants were constructed using site-directed mutagenesis (Stratagene, La Jolla, CA) with oligonucleotide primers from MWG Biotech (High Point, NC). Sequences were confirmed with forward and reverse sequencing (MIT CCR Biopolymers Laboratory).

4.4.4 Measurement of binding affinity

The yeast strain EBY-100 was transformed with each surface-display plasmid and binding affinities were measured as previously described in Chapter 3. Binding to lysozyme (Sigma Aldrich, St. Louis, MO) was detected through secondary labeling with biotinylated rabbit polyclonal anti-lysozyme antibodies (Research Diagnostics, now Fitzgerald Industries, Concord, MA) and tertiary labeling with streptavidin-phycoerythrin (Invitrogen, Eugene, OR). Polyclonal antibodies that bound yeast nonspecifically were removed prior to use. Experiments were carried out at 25 °C in phosphate-buffered saline (PBS), pH 7.3–7.5, 0.145 M salt (0.167 M ionic strength), except for high-salt experiments where NaCl was added to reach 1.67 M ionic strength. For cetuximab, the variant 404SG engineered for expression in yeast of the extracellular domain of epidermal growth factor receptor was produced solubly [108]. This variant contains two mutations in domain III, Phe(380)Ser, and Ser(418)Gly at the interface with cetuximab. Binding was detected through secondary labeling with a biotinylated mouse monoclonal anti-FLAG antibody (Sigma Aldrich, St. Louis, MO) and tertiary labeling with streptavidin-phycoerythrin.

4.4.5 Measurement of binding kinetics

Off-rates were determined by labeling displaying cells with excess lysozyme, washing free lysozyme from solution, and then labeling with excess biotinylated lysozyme

Table 4.7: Sequence details for new D44.1 surface display plasmid within pCTCON2.

	DNA sequence	Amino-acid sequence
HA	TACCCATACGACGTTCCAGACTACGCT	YPYDVPDYA
linker	CTGCAGGCTAGTGGTGGAGGAGGCTCTGGTGGA GGCGGTAGCGGAGGCGGAGGGTCG	LQASGGGGSGGGGSGGGGS
NheI	GCTAGC	AS
V_H	CAAGTTCAGTTACAAGAAAGTGGTGCCGAAGTT ATGAAGCCAGGTGCATCCGTCAAGATCTCTTGT AAGGCTACTGGTTATACATTTTCAACTTACTGG ATTGAATGGGTTAAACAAAGACCCGGTCATGGT CTAGAATGGATTGGTGAAATTTTACCAGGTAGT GGAAGCACTTACTATAATGAGAAATTCAAAGGC AAAGCCACTTTTACAGCAGATACTTCTTCAAAT ACCGCTTATATGCAACTGTCTAGCCTAACCAGC GAAGATAGTGCAGTTTATTATTGCGCTAGAGGT GATGGAAATTATGGATATTGGGGTCAAGGAACA ACATTGACTGTTTCTAGT	QVQLQESGAEVMKPGASVKI SCKATGYTFSTYWIEWVKQR PGHGLEWIGEILPGSGSTYY NEKFKGKATFTADTSSNTAY MQLSSLTSEDSAVYYCARGD GNYGYWGQGTTLTVSS
KasI- linker-SpeI	GGCGCCGGAGGTTCAGGCGGCGGTGGTTCCGGT GGAGGTACTAGT	GAGGSGGGGSGGGTS
V_L	GATATTGAACTGACCCAAAGTCCAGCTACTTTA TCTGTCACCCCTGGTGATTCCGTTTCTTTGTCA TGCAGAGCATCTCAAAGCATTTCCAACAACTTA CATTGGTATCAACAAAAGTCCCATGAGTCCCCT CGTCTACTTATCAAGTACGTATCTCAGAGTTCA AGCGGAATACCTTCACGTTTCTCAGGTTCCGGT TCAGGCACAGATTTTACATTAAGCATCAACTCC GTTGAAACTGAAGACTTCGGTATGTATTTTTGT CAGCAATCTAACAGCTGGCCCAGAACATTCGGC GGTGGCACCAAGCTTGAAATTAAG	DIELTQSPATLSVTPGDSVS LSCRASQSISNNLHWYQQKS HESPRLLIKYVSQSSSGIPS RFSGSGSGTDFTLSINSVET EDFGMYFCQQSNSWPRTFGG GTKLEIK
linker	GCTGGCGCT	AGA
BamHI	GGATCC	GS
c-MYC	GAACAAAAGCTTATTTCTGAAGAGGACTTG	EQKLISEEDL
stop-stop	TAATAG	**
XhoI	CTCGAG	LE

as a function of time. Mean fluorescence data were fit to a single exponential for dissociation time. On-rates were determined by labeling displaying cells with set lysozyme concentrations and then washing away free lysozyme as a function of time and labeling with the anti-lysozyme antibodies. Mean fluorescence data were fit to a single exponential for apparent association time, and a linear plot of this time constant versus initial lysozyme concentration yielded the intrinsic on-rate.

Chapter 5

Development of an improved nonpolar solvation model

Abstract

In Chapter 3, computational protein design methods were developed for the rational improvement of binding affinity. The procedure was extended in Chapter 4 to focus on the electrostatic free energy term of binding, resulting in significant affinity improvements. However, the problem of mutations to larger side chains dominating thet predictions due to seemingly inaccurate calculated packing interactions remained. Here we investigate a physics-based improvement to the nonpolar term of solvation free energy. A continuum van der Waals interaction energy model was implemented efficiency and combined with a surface area cavitation model to form a new nonpolar free energy model. The model was parameterized to reproduce the experimental solvation free energies of alkanes. The two nonpolar models were compared for single mutation protein design calculations in the D1.3–lysozyme system and three addition anti-lysozyme antibodies. The new nonpolar model attenuates the prediction of larger side chains as expected, but does not address the few most egregious designs. Possible improvements to the model are outlined.

5.1 Introduction

In Chapter 3 we developed methods for the computational design of improved protein–protein binding affinity. Predictions in the anti-lysozyme model antibody D1.3 were dominated by unintuitive mutations to large amino acids calculated to make improved van der Waals packing interactions at the binding site periphery. Experimental characterization of 17 single mutations, as well as two mutations to larger amino acids in D44.1 in Chapter 4, led us to conclude that improvement was indeed necessary in the methods or energy function. At the same time, predictions based on calculated electrostatics improvements were taken advantage of to achieve single mutation high success rates and combination mutants with greatly improved affinities (Chapter 4). In this chapter we explore a physics-based improvement to the energy function to address the design of larger amino acids.

The two main limitations in computational protein design are conformational search thoroughness and energy function accuracy. Insufficient conformational sampling, such as from fixed backbone or discrete rotameric side chain assumptions, or from heuristic search algorithms, can cause the true low energy conformation of a sequence to be missed. On the other hand, an inaccurate energy function can result in the correct low energy conformation to be scored worse than decoy conformations, or can produce an incorrect relative binding energy for the correct conformation of a mutant sequence.

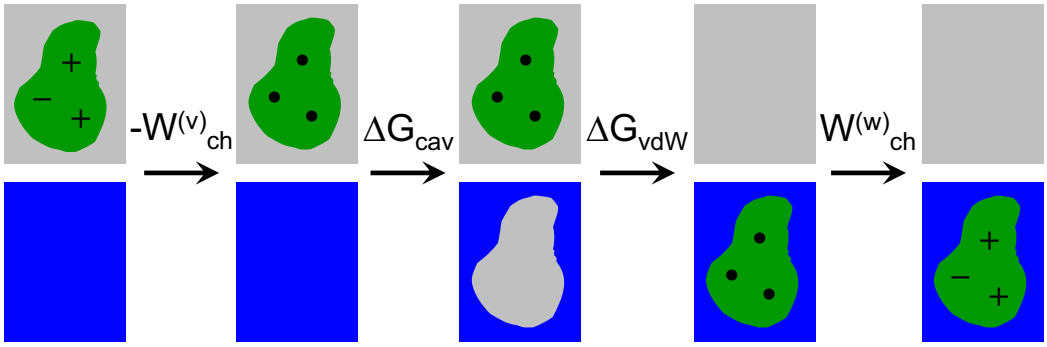
The over-prediction of large amino acids in affinity design is most likely an energy function issue. The calculated favorable van der Waals interactions from larger residues will not be attentuated by increased conformational search, rather, the calculated interactions would only improve given more sampling. Since improvement is relative to wild type, this argument makes the reasonable assumption that the small wild type side chain will not likely find a better conformation for interaction than that of the crystal structure or the repacking of the wild-type sequence. More likely, the error is from the energy function in correctly selecting the lowest-energy conformation of a sequence or in calculating the binding energy for that lowest-energy structure.

Since the characteristic feature of the large amino acids predictions is the calculation of greatly improved van der Waals interactions, the appropriate question is: Are the improved van der Waals interactions too large, or should there be a different term in the energy function that counteracts the van der Waals? Four types of interaction to be accounted for are protein–protein and protein–water, van der Waals and electrostatics. Side chain entropy loss upon binding, though ignored in the energy functions, is unlikely to change the predictions based on an order of magnitude analysis: a glycine to tryptophan mutation at most completely fixes two dihedral degrees of freedom for approximately 1.2 kcal/mol, compared to the several kcal/mol van der Waals interaction differences being calculated. Calculation of binding requires the trade-off of bound-state and unbound-state interactions. Upon mutation from a small to large amino acid, increased bound-state protein–protein interactions are in principle counteracted by increased unbound-state protein–water interactions, but these offseting energetics are often calculated asymmetrically using an atomistic Lennard-Jones potential and a non-atomistic nonpolar term of solvation, respectively.

The interactions of a protein with its aqueous environment are accounted for using a continuum solvation model, necessarily avoiding the computational complexity of explicit water. As depicted in Figure 5.1, solvation can be broken down into four steps, grouped as net electrostatics (ΔG_{elec}), cavitation (ΔG_{cav}), and van der Waals (ΔG_{vdW}). Cavitation and van der Waals are commonly further grouped into a total nonpolar solvation term ($\Delta G_{\text{nonpolar}}$). With the electrostatics modeled with Poisson–Boltzmann continuum solvation, we had adopted the common solvent-accessible surface area (SASA) nonpolar term shown in Equation 5.1.

$$G_{\text{nonpolar}} = \gamma SASA + b \tag{5.1}$$

A variety of work has indicated deficiencies in and developed alternatives to the surface-area approach to the nonpolar term of solvation [43, 114, 115, 121–126]. The main limitation of a surface-area approach is that atoms that contribute little or no solvent-exposed surface area can in fact interact favorably with solvent. Second,



$$\Delta G_{\text{solvation}} = -W_{\text{ch}}^{(\text{v})} + \Delta G_{\text{cav}} + \Delta G_{\text{vdW}} + W_{\text{ch}}^{(\text{s})}$$
$$\Delta G_{\text{solvation}} = \Delta G_{\text{elec}} + \Delta G_{\text{cav}} + \Delta G_{\text{vdW}}$$
$$\Delta G_{\text{solvation}} = \Delta G_{\text{elec}} + \Delta G_{\text{nonpolar}}$$

Figure 5.1: Stepwise solvation process. Transfer of a protein from vacuum (gray) to water (blue). Left to right: electrostatic work to neutralize charges in vacuum environment, cavitation energy to create a solvent cavity, van der Waals energy between protein and solvent upon transfer from vacuum into solvent, electrostatic work to re-charge the protein in the aqueous environment.

surface-area approaches tend to use a single energy per unit area independent of atom-type, or a parameterized value for each of hydrophilic and hydrophobic surface area, but do not account for the full chemical diversity of atom types and their different parameterizations within molecular mechanics force fields.

We implemented a continuum solvent van der Waals interaction energy model based on work by Levy and co-workers [43]. This continuum van der Waals model aims to combine the computational efficiency of continuum models with the detail of explicit solvent calculations. We combined it with a surface area cavitation model to create a full nonpolar solvation model and parameterized it to match the experimental solvation free energies of alkanes. Unlike the solvent van der Waals term, the cavitation term is expected to be modeled adequately by a simple surface area term [124, 125, 127].

5.2 Methods

5.2.1 Continuum van der Waals model

The continuum van der Waals model as described by Levy and co-workers [43] assumes that the average water atom number density outside of the protein volume is constant. The protein is modeled with a standard molecular mechanics force field and the solvent region is modeled as a continuous smear of atoms of a water model. The interaction energy between protein and solvent is the sum of the interactions of each protein atom with solvent; the interaction of a protein atom with solvent is computed by an integral from point to volume, rather than the traditional point-to-point van der Waals evaluation. Nevertheless, the same force field parameters are used.

Atom–solvent interaction is calculated using Equation 5.2,

$$U_{\mathrm{vdW}}(\mathrm{i}) = \int_{\mathrm{solvent}} \rho_{\mathrm{w}} u_{\mathrm{vdW}}^{(\mathrm{i})}(\mathbf{r} - \mathbf{r}_i) d^3\mathbf{r} \tag{5.2}$$

where ρ_{w} is the water bulk number density taken to be 0.0336 $\mathrm{\AA}^{-3}$ and $\mathrm{u}_{\mathrm{vdW}}^{(\mathrm{i})}$ is the Lennard-Jones pair potential for atom type i and water. For the CHARMM-adapted TIP3P water model, Equation 5.2 is modified to count both water oxygen and water hydrogen interactions, where the hydrogen bulk number density is twice that of water/oxygen density. The total protein–solvent interaction energy is then simply the sum over all protein atoms,

$$U_{\mathrm{vdW}} = \sum_{i=1}^{n} U_{\mathrm{vdW}}(\mathrm{i}) \tag{5.3}$$

where n is the number of atoms in the protein.

The boundary between protein and solvent is taken as the solvent-accessible surface, which is the closest approach of the center of a probe sphere rolled over the protein set of spherical atoms. However, the radius of this probe sphere, though commonly taken to be 1.4 Å to approximate a water molecule, is left as a free parameter. This is the only free parameter of the continuum van der Waals interaction model

and is used to balance the assumptions that the solvent region is of constant density and approaches the protein to a discrete, solvent-accessible surface.

Levy and co-workers used the continuum van der Waals model with the OPLS force field [128] and TIP4P water model [129], parameterizing the solvent probe radius to match protein–solvent interaction energies from explicit solvent molecular dynamics calculations. Since we were to add a cavitation term to address the full nonpolar solvation energy, which adds one or two parameters itself, we decided to fit all free parameters simultaneously to the experimental solvation free energies of alkane molecules.

5.2.2 Numerical solution

To compute the interaction given by Equation 5.2, the integral was converted to one with a bounded region as done previously [43]. The integral from an atom to the unbounded region of solvent is equivalent to the analytical integral from the atom to all space, minus the overcounted interaction from the atom to the interior protein region. In Equation 5.4,

$$U_{\mathrm{vdW}}^{(\mathrm{i})} = U_{\mathrm{vdW}}^{(\mathrm{i})}(\mathrm{isolated}) - \int_{\mathrm{protein}} \rho_{\mathrm{w}} \Theta(|\mathbf{r} - \mathbf{r}_i| - R_{\mathrm{i}}) u_{\mathrm{vdW}}^{(\mathrm{i})}(\mathbf{r} - \mathbf{r}_i) d^3\mathbf{r} \tag{5.4}$$

the first term can be computed analytically for atom i with radius R_{i}. The second term corrects for overcounted interaction in the region outside atom i yet inside the protein region, and the step function Θ handles integrating only outside of atom i.

A C++ program was written for numerically solving the continuum van der Waals integration given by the second term in Equation 5.4. The basic approach was to divide the bounded region into volume elements and sum the analytical integral to each volume element. Cubic volume elements were inadequate for the inherently spherical geometries. In one implementation, spherical volume elements based on the atom of integration were used, with fixed divisions in θ and ϕ angular space, and progressively-spaced radial divisions to take advantage of the r^{-6} dependence of the interaction. The fine three-dimensional discretization was set in advance, and setup

new for each atom. In a second implementation, a coarse initial discretization was used, and elements that were not provably all-inside or all-outside were recursively subdivided until the value of their integral fell below a threshold.

5.2.3 Cavitation model

The cavitation energy was modeled as directly proportional to the molecular (solvent-excluded) surface area [124, 125, 127] as given by Equation 5.5.

$$G_{\text{cav}} = \gamma MSA + b \tag{5.5}$$

The molecular surface area was calculated using a solvent probe radius of 1.4 Å in the analytical surface routine of the program MSMS [130]. The slope, γ, and intercept, b, of this model were free parameters.

5.2.4 Fit to alkane experimental data

To determine the free parameters of the new nonpolar solvation model, we sought to recapitulate the experimental solvation free energies of different alkanes. Alkanes are well-suited for the task because their minimal electrostatic component of solvation means that the nonpolar component is critical for matching the experimental data. In addition, the linear alkanes have been used previously to determine the solvent-accessible surface area nonpolar model parameters for the PARSE parameter set for Poisson–Boltzmann continuum electrostatics [62]. The branched and cyclic alkanes can not be modeled adequately by a SASA model for nonpolar solvation, data that support the adoption of a more physics-based energy function. Some atoms in branched and cyclic alkanes are partially buried from solvent and contribute more to solute-solvent interaction than their contribution to the surface area would otherwise indicate.

Alkane data from Cabani *et al.* were used [131]. Alkanes that contained a quaternary carbon were skipped due to lack of CHARMM parameters. Each alkane was modeled in its lowest-energy conformation, selected after minimizing all possible all-

staggered conformations for acyclic compounds and minimizing known low-energy states for cyclic compounds. The small but non-zero electrostatic contribution to the solvation free energy was calculated using Poisson–Boltzmann continuum electrostatics with an internal dielectric constant of 2 and transfer from dielectric 80 to vacuum dielectric 1.

There were three free parameters in total: cavitation slope, cavitation intercept, and continuum van der Waals solvent probe radius. These parameters were determined by best fit to the experimental solvation free energies of the linear, branched, and cyclic alkanes. Values for the solvent probe radius were enumerated in 0.01 Å increments from 0.6 Å to 1.5 Å. For each value, the optimal cavitation slope and intercept value were determined using analytical linear best-fitting, and the sum of squared residuals was recorded. The three parameters that gave the lowest overall residual were selected. The fit was repeated for both PARAM22 and PARAM19 molecular mechanics, but all work here used the all-atom PARAM22 set.

5.2.5 Protein test systems

In addition to the anti-lysozyme model antibody D1.3 used in Chapter 3, three more anti-lysozyme antibodies were used for the computational evaluation of the new nonpolar model. The antibodies HyHEL-5, HyHEL-8, and D44.1 bind at a second and third lysozyme epitope, and crystal structures are available (Protein Data Bank codes 1YQV [132], 1NDG [48], and 1MLC [117], respectively). The structures were prepared for calculation following the procedures outlined in Chapter 3.

5.3 Results

The first step for investigation of the new nonpolar solvation term was the efficient implementation of a numerical solver for the integration in Equation 5.4. The integral was calculated using a spherical grid centered on each atom. The refinement of the grid was controlled by two parameters corresponding to the radial volume element size, $\Delta\rho$, and the angular divisions, $\Delta\phi$. Division of the angular coordinate θ was

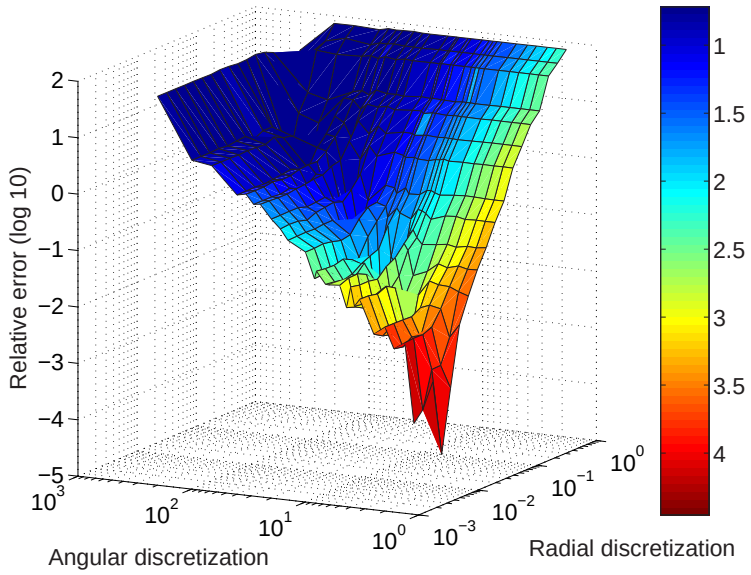


Figure 5.2: Convergence of continuum van der Waals numerical solution. Radial discretization controls initial width in Å of volume elements; angular discretization controls angular sweep of ϕ and θ. Error is relative to the finest discretization. Surface is colored according to time in solver, on a log 10 scale.

coupled to ϕ for simplicity. Figure 5.2 shows the convergence with increasing discretization for the interaction between the D1.3–lysozyme complex and water. The relative error steadily decreases as the two grid parameters are reduced together. Along any particular slice in contant $\Delta\rho$ or $\Delta\phi$, the relative error reaches a plateau due to the accuracy limitation of the current value of the other parameter. The solver performs optimally with parameters chosen along the crease in the surface where the shortest time is used for a given relative error tolerance.

The solver accuracy was set to achieve 0.1 kcal/mol accuracy in the calculation of relative binding energies. Figure 5.3 shows the comparison of faster, less accurate solver settings with a more accurate solver setting for the calculation of the continuum van der Waals of binding for D1.3 sequence and structure variants. The deviation of each data set from $y = x$ (taking into account the 1 kcal/mol vertical offset for plotting clarity) was used to evaluated the accuracy in relative binding energy prediction for particular solver parameters.

The continuum van der Waals interaction model with the surface area cavitation model were parameterized to reproduce the experimental solvation free energies of

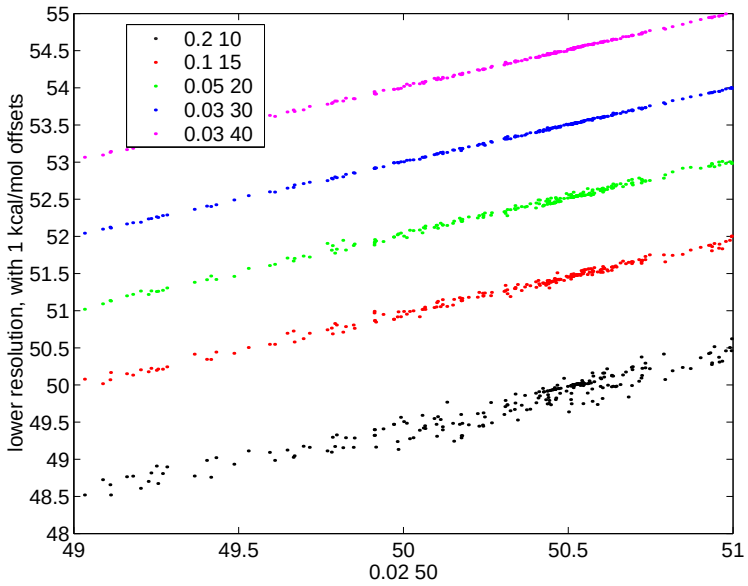


Figure 5.3: Convergence of relative binding energies for D1.3 sequence and structure variants. A more accurate and more computationally expensive discretization was used to calculate the continuum van der Waals on the x-axis. Less accurate yet faster solutions are plotted on the y-axis, with 1 kcal/mol offsets for clarity. The legend refers to the radial discretization (e.g. 0.02 Å) and the angular discretization in number per 180 degrees (e.g. 50 produces 3.6 degrees).

alkanes. There were 7 linear alkanes (C2–C8), 5 branched alkanes, and 7 cyclic alkanes. The previous, simple solvent-accessible surface area (SASA) model for nonpolar solvation adequately fit the linear alkane data (Figure 5.4A) as has been demonstrated previously. However, the SASA model could not fit the full set of alkanes because of the cyclic molecules (Figure 5.4B). On the other hand, the new nonpolar model with continuum van der Waals interaction and cavitation was capable of fitting all of the alkane data (Figure 5.4C). Figure 5.4D shows the details of the fit, illustrating that the favorable van der Waals interaction with solvent is greater for the cyclic alkanes than otherwise predicted based on surface area. The experimental values (open squares) are not described well by solvent-accessible surface area, nor by molecular (solvent-excluded) surface area (data not shown). The fit parameters are: continuum van der Waals probe radius = 0.75 Å, cavitation slope = 52.5 cal/mol/Å^2, cavitation intercept = −1.6 kcal/mol (PARAM19 fit values were 0.73 Å, 103.5 cal/mol/Å^2, and −1.4 kcal/mol, respectively).

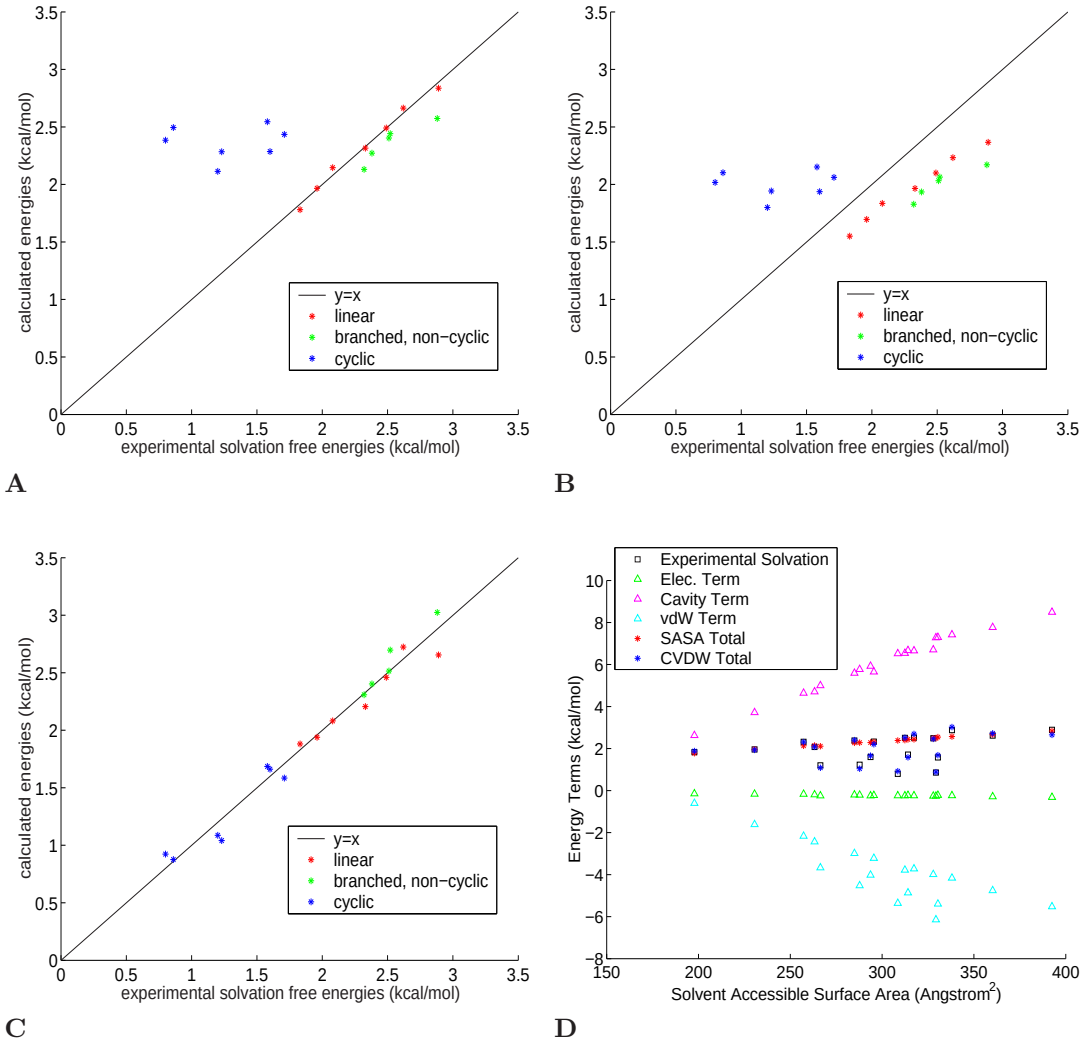


Figure 5.4: Parameterization of nonpolar solvation models to alkane data. In (A–C), calculated solvation energy is plotted versus the experimental measurement. Red: linear; green: branched; blue: cyclic. (A) Solvent-accessible surface area (SASA) model fit to only the linear alkane data. (B) SASA model fit to all data. (C) New nonpolar model fit to all data. (D) Details of the new fit. Energies are plotted as a function of solvent-accessible surface area, with the experimental values as hollow square targets.

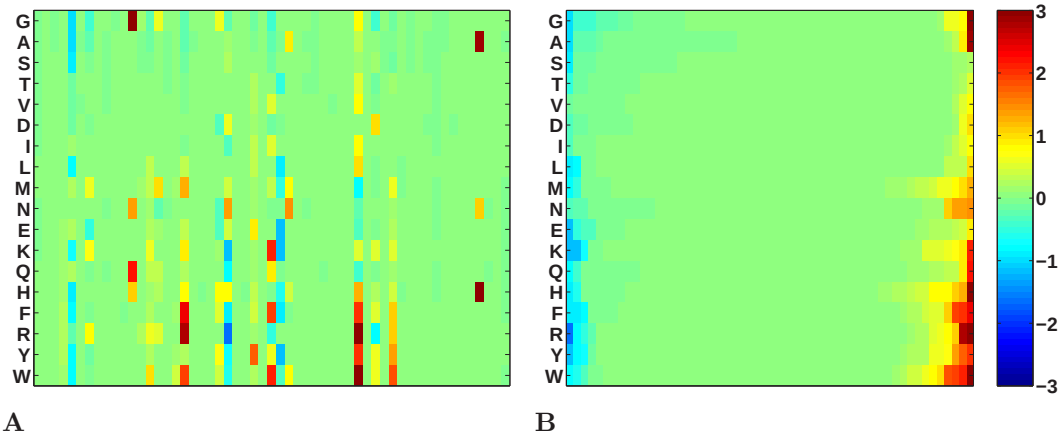


Figure 5.5: Effect of continuum van der Waals on D1.3 single mutation designs. (A) Each column is a position in the antibody and the rows are the predicted effects on binding energy relative to wild type in kcal/mol for each single mutation. Each energy is the difference between the solvent-accessible surface area (SASA) nonpolar model and the continuum van der Waals plus cavitation nonpolar model, with yellow/red indicating that the mutation is predicted to be less favorable for binding in the new model. (B) Results are sorted by row to highlight the overall attenuation of large amino acids with the new nonpolar model.

The new nonpolar model, combining continuum van der Waals interaction with cavitation, was applied to the reevaluation of designed D1.3 single mutations for improving binding affinity to its antigen, hen egg-white lysozyme. Previously, designs were dominated by mutation to larger amino acids. Replacing the solvent-accessible surface area (SASA) nonpolar term with the new nonpolar term resulted in the overall attenuation of large amino acids as shown in Figure 5.5. Figure 5.5A reports the effect of the new nonpolar model and Figure 5.5B arranges the same data by row to highlight the effect by amino acid. The accumulation of yellow/red less-favorable energies for the design of large amino acids is consistent with the goal that large amino acids would not be as highly ranked with a new, more physics-based energy function.

A direct comparison of the old and new nonpolar models for predicting the effect of mutating a small residue to a larger side chain is shown in Figure 5.6. Small is defined as a wild-type side chain with four or less non-hydrogen atoms (Ala, Cys, Asp, Gly, Ile, Leu, Met, Asn, Pro, Ser, Thr, Val), and large is defined as five or more non-hydrogen atoms (His, Phe, Tyr, Trp) excluding arginine to avoid complications

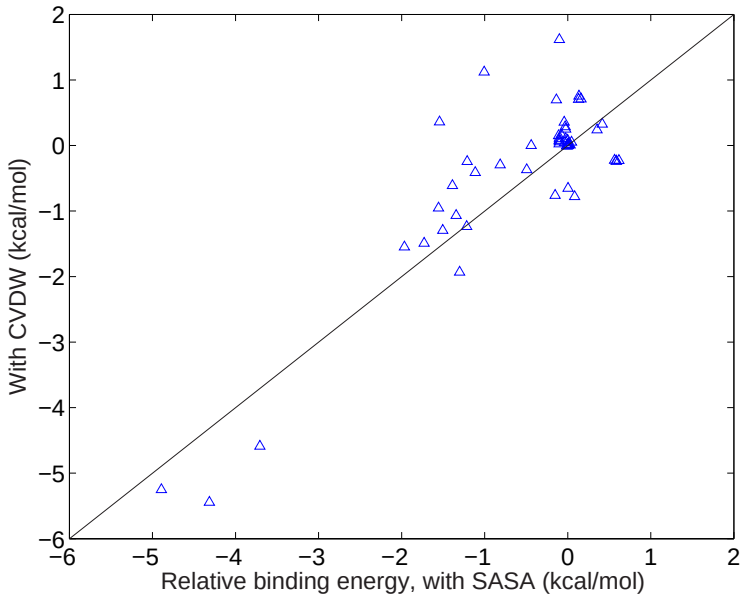


Figure 5.6: Effect of CVDW for small-to-big mutations. The total binding energy relative to wild type in kcal/mol is plotted with the nonpolar term evaluated with either the old or new models.

in analysis due to the positive charge. There are two main points about this data. First, the three mutations that are predicted to be distinctly favorable for binding relative to wild type (H31 Gly mutated to Trp, Phe, Tyr) are not attenuated in the new model as had been thought; those predictions are in fact larger in magnitude than before. Second, the overall effect is nevertheless to attenuate the predictions as had been thought. Predictions that were at least -0.5 kcal/mol favorable in the original energy function exhibit an average attenuation of $+0.36$ kcal/mol.

The new nonpolar model was next applied to the reevaluation of designs in three additional anti-lysozyme antibody systems. The effect of the new model on the relative binding free energies as sorted by type of amino acid is shown in Figure 5.7. The three new systems do not show the strong preference for attenuating large side chains as for D1.3. There are fewer changes in HyHEL-5 (Figure 5.7B) with a slight preference to attenuate the larger side chains, and the changes in HyHEL-8 (Figure 5.7C) are also more spread out, but with some preference to attenuation of the larger side chains. The new nonpolar model changes the D44.1 designs much less (Figure 5.7D) and surprisingly attenuates the two smallest side chains the most. These results will

be discussed below.

5.4 Discussion

We explored an improved model for nonpolar solvation after computational design for improved binding affinity (Chapter 3) revealed predictions that were dominated by mutation to larger amino acids. Analysis of the physical interactions involved in binding, the calculation techniques, and recent progress in the field of biophysical modeling suggested that this direction was necessary.

This work consisted of three steps: (1) development of software for the efficient and accurate numerical solution of the improved model, (2) determination of the free parameters in the model, and (3) evaluation of the parameterized model for protein–protein relative binding free energy predictions.

A C++ solver for the irregular-geometry integration in Equation 5.4 was developed that satisfies accuracy and efficiency goals. A single protein–protein relative binding free energy can be calculated to within 0.1 kcal/mol on the order of 1 minute, which is less time than required for the Poisson–Boltzmann continuum electrostatics calculation that is executed in parallel.

The new nonpolar model, a combination of the continuum van der Waals interaction model of Levy and co-workers [43] and a surface-area cavitation model, contains three free parameters. Parameter values were fit to reproduce the experimental solvation free energies of linear, branched, and cyclic alkanes, as the previous model was incapable of capturing this data. This fit was necessary for the success of the new model, but possibly not sufficient.

The new nonpolar model was evaluated for its effect on the predicted binding free energies relative to wild type for single mutations compared to the previous solvent-accessible surface area (SASA) model. The results were mixed. On one hand, the few, most notable predictions of large favorability from mutation from a small to large amino acid were not greatly altered by the new nonpolar model, but we had expected these predictions to be decreased in magnitude. Nevertheless, as a whole,

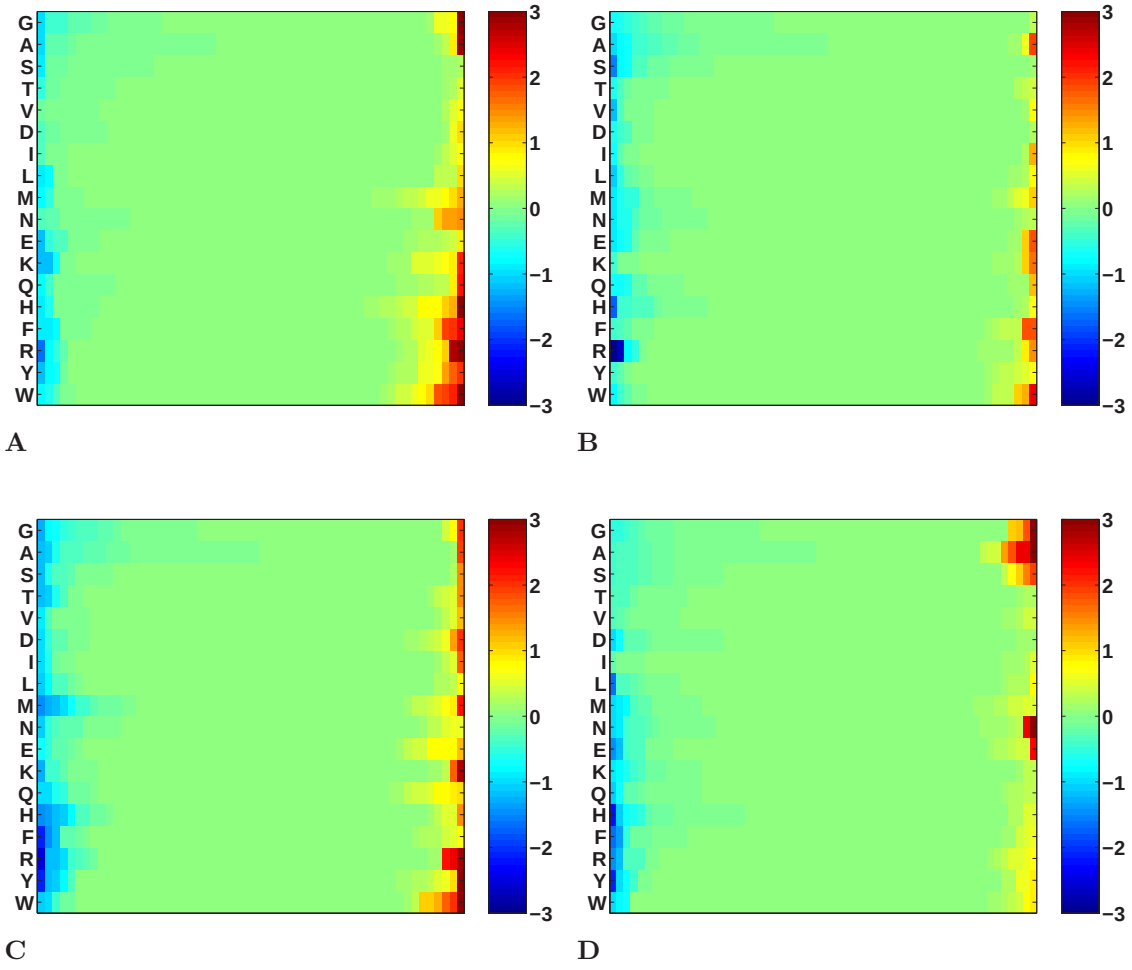


Figure 5.7: Continuum van der Waals effect in four systems. Single mutations were designed at each antibody position and the difference between the old and new non-polar models for binding free energy relative to wild-type is shown in kcal/mol. The results were are sorted by amino acid. (A) D1.3, as shown in Figure 5.5B. (B) HyHEL-5. (C) HyHEL-8. (D) D44.1.

the new model did attenuate the predictions of large amino acids. The fourth test system, D44.1–lysozyme, is the only exception, as several mutations to glycine and alanine are made less favorable with the new nonpolar model.

Further work should be continued to explore the new nonpolar energy model. The current parameterization exhibits the desired qualitative behavior, but lacks quantitative success for the few largest-magnitude predictions of large amino acids. Simultaneous fitting of the one continuum van der Waals and two cavitation model parameters to the alkane data may be sufficiently consistent with a family of parameter sets and not contrained enough for a particular most-effective parameter set. The fact that the PARAM19 and PARAM22 independent fits yielded different cavitation slope parameters indicates a possible problem; one might expect the cavitation term to be independed of the molecular mechanics parameter set. Instead, one could fit the probe radius in the continuum van der Waals term directly to interactions from explicit water molecular dynamics simulations, as done by Levy and co-workers [43]. In a second step, the cavitation slope and intercept parameters could be determined by fitting the alkane experimental solvation free energy data as in this work, only with the continuum van der Waals probe radius pre-determined.

In addition, one should investigate whether alternate conformations of the alkanes with Boltzmann-weighting would improve the fit [133]. Also, it may be necessary to add a second continuum van der Waals parameter, fourth overall, to allow a different probe radius and corresponding molecular surface boundary for the continuous water oxygen atoms versus the water hydrogen atoms. Previous work used a TIP4P water model with Lennard-Jones parameters only on the oxygen, whereas the CHARMM-adapted TIP3P water model includes van der Waals interactions from all three atoms.

Chapter 6

General conclusions

This thesis has presented novel methods for the analysis and design of high-affinity protein interactions. Analysis of the electrostatic component of binding in a high-affinity antibody variant elucidated original mechanisms for the role of mutations that had been accumulated through directed evolution. Development of computational protein design methods and their application to the design of affinity-enhancing mutations met with mild success in the first anti-lysozyme model antibody D1.3. The calculations and experimental measurements identified two directions that were each subsequently explored. First, new prediction methods that focused on a calculated improved electrostatic component of binding free energy led to robust and significant improvements in binding affinity in a number of systems, including a 10-fold improvement in binding of the therapeutic antibody cetuximab (Erbitux) to its ligand, epidermal growth factor receptor (EGFR). Second, an investigation of a new model for the nonpolar component of the solvation free energy showed encouraging progress but necessary future work.

One theme in this thesis was the use of the Poisson–Boltzmann continuum electrostatic model. This model was at the core of tools used in Chapter 2 to analyze anti-fluorescein antibodies and reveal novel insights into the molecular mechanisms of four affinity-enhancing mutations. This successful analysis lent credibility to the use of the model within protein design methods. A central conclusion of Chapter 3 was that the developed protein design methods might be improved using a particular focus on calculated electrostatic free energies. In Chapter 4 we implemented that idea and created large and robust binding affinity improvements in several systems. The mutations were either polar to hydrophobic side chains, e. g. asparagine to alanine, or polar to charged side chains, e. g. threonine to aspartate. We found it as important

to consider depolarizing mutations, to reduce the side chain desolvation penalty, as to consider charged mutations that increase favorable interactions. The electrostatic term of the binding free energy was a better predictor for experimental improvement than the calculated total binding free energy, and prediction based on the naive total computed energy would have missed some of the most favorable mutations that were predicted to lose van der Waals interactions. As an inverse design problem, we only needed to use a subset of all possible ways to improve binding affinity. There were likely to be multiple affinity-enhancing mutations, with design success only needing to identify a subset that achieved overall improvement. At the same time, however, our scientific interest for the long-term development of protein design led us subsequently to an investigation of the underlying physical models.

A tenet of this thesis was the use of thorough and guaranteed optimization tools to address the protein design methods and energy function. In the protein design development described in Chapter 3, we chose to consider many small search spaces, single and double mutations across a protein–protein interface, instead of a few large regions of the interface or a single full interface design. With multiple independent and small problems, we were able to use computational optimization to rank-order the structures with the lowest calculated energy in each space. In Chapter 3 this approach led to unintuitive mutations to large side chains dominating the predictions for improved binding free energy. We could isolate the problem as an energy function issue apart from conformational sampling because of our exhaustive search approach. Larger side chains were calculated to make radically improved van der Waals packing interactions across the periphery of the binding interface, and designs of multiple simultaneous positions were saturated by these unintuitive predictions. We concluded that an investigation was necessary, and this conclusion was greatly facilitated by the use of optimization.

In Chapter 5 we investigated the over-prediction of large side chains by addressing a physical deficiency in the model. A new model for the nonpolar component of solvation free energy was developed using a continuum van der Waals interaction model from Levy and co-workers combined with a surface area cavitation model.

The new nonpolar model attenuated the prediction of larger side chains, as expected, but did not address the few most egregious designs. This work was an improvement overall, but future work is needed. More rigorous parameterizations are expected to give improved results, and alternate related work may address the problems as well. We expect that physics-based developments in the underlying model will improve not only our own results, but work in other systems and problems as well.

A goal of this thesis was the development of robust computational design methods. We sought a method that would produce a high success rate for designed improvements, as well as a method that would be transferable for successful use in other protein systems. Initial work in the anti-lysozyme model antibody D1.3 gave marginal improvements. On the other hand, improvements of 1–2 orders of magnitude were achieved in anti-lysozyme D44.1 and anti-EGFR cetuximab with their binding partners. The method was further validated with anti-fluorescein 4-4-20 and anti-vascular endothelial growth factor (VEGF) bevacizumab (Avastin). Moreover, calculations on HyHEL-5, HyHEL-8, and Herceptin with their respective ligands indicate opportunities for single mutations to improve binding affinity, yet untested. Therefore, the electrostatics-based design method appears generally capable of improving the binding affinity of an antibody/antigen system starting from low-nanomolar binding affinity, with D1.3 as a notable exception. We explored a possible structural basis for D1.3 as an anomaly and its implications on the design procedure.

We postulate that the antibody D1.3 might not be a "model" system, representative of low-nanomolar antibody–protein interactions in general. The D1.3–lysozyme interface contains a much greater-than-average number of interfacial buried water molecules, many of which interact with each other in small pools. In addition, D1.3 presents primarily large side chains at the interface (tyrosine, tryptophan, arginine). The binding-site interactions may be at a type of local minimum for the generic packing problem, where a large number of concerted mutations would be necessary to repack the binding interface into a lower-energy state. Our single- and double-mutation designs are possibly less adept at mutating away from the largest residues; they are possibly more adept at mutating away from the small- or medium-sized side

chains, where options for mutation to bigger, smaller, hydrophobic, and charged side chains all exist without dramatic shape change. We expect that some small fraction of cases may be like D1.3, at least in the general aspect of poor designability with the methods developed here.

Results in this thesis have implications for the protein and nucleotide libraries used in directed evolution experiments and the future of computational protein design. A protein variant library generated using error-prone PCR at the DNA level generally does not translate to covering all single amino acid mutations, let alone all possible double mutations. The nature of the genetic code results in peculiarities in mutation frequencies. For example, 10 of the 18 neutral or positively-charged protein side chains can not be mutated to either negatively-charged side chain via any single DNA base pair change. Computation is not fundamentally biased to the genetic code. We found overall that 10 of the 12 single mutations that improve binding affinity in systems that had been previously optimized by *in vivo* maturation required two base pair changes, where the likelihood by chance for 10 or more out of 12 is approximately 20%. Also, the designed affinity-enhancing double mutation requires three base-pair changes. Computation is an engineering approach complementary to directed evolution that has the capability of addressing a vastly larger protein sequence space, capable of discovering unique and valuable protein variants.

This thesis has presented novel methods for the rational improvement of binding affinity. Order-of-magnitude improvements were accessible in most systems by the combination of additive single mutations. Methods were validated on the therapeutically-relevant FDA-approved anti-EGFR antibody cetuximab, where 10-fold improvement in binding was achieved, and on the model anti-lysozyme antibody D44.1 with 140-fold improvement in binding free energy. Improvements were made in both on- and off-rates, which may be important for applications that require specifically faster association or slower dissociation rates. Overall, these results demonstrate novel computational capabilities and indicate their applicability for enhancing and accelerating development of reagents and therapeutics.

Appendix A

Evolution of an interloop disulfide bond in high-affinity fibronectin type III antibody mimics: Molecular convergence with single-domain camelid and shark antibodies[1]

Abstract

Antibody mimics based on the ^{10}Fn3 immunoglobulin-like scaffold were previously selected by yeast surface display to bind with high affinity to the model antigen hen egg-white lysozyme. A striking feature of the highest-affinity, 350 pM variant was a pair of selected cysteine residues at positions 28 and 77 in adjacent binding loops. Here we use molecular modeling to estimate the likelihood for each inter-loop pair of positions forming a disulfide bond. The modeling results and previous mutagenesis and evolution experiments all support the hypothesis that the two cysteines selected in the high-affinity ^{10}Fn3 variant form a disulfide bond required for the strong protein–protein interaction. The selection of this cysteine pair is structurally analogous to the natural evolution of disulfide bonds found in new antigen receptors of cartilaginous fish and in camelid heavy-chain variable domains.

[1]Portions of this chapter have been accepted as:

Lipovsek D., Lippow, S. M., Hackel, B. J., Gregson, M. W., Cheng, P., Kapila, A. & Wittrup, K. D. Evolution of an interloop disulfide bond in high-affinity antibody mimics based on fibronectin type III domain and selected by yeast surface display: Molecular convergence with single-domain camelid and shark antibodies. *J. Mol. Biol.* In Press (2007).

A.1 Introduction

The tenth human fibronectin type III domain (^{10}Fn3) has been previously engineered to bind with high affinity and specificity to arbitrary protein targets [134–139]. This antibody mimic has structural homology to immunoglobulin folds, resembling a V_H-only antibody with its loops BC, DE, and FG that connect beta-sheet elements analogous to antibody CDR-1, CDR-2, and CDR-3 loops [140, 141]. Wild-type ^{10}Fn3 possesses attractive qualitives for development as a therapeutic scaffold: small size (10 kD), high expression in *E. coli*, and high stability (82°C melting temperature [138]).

Lipovsek and Wittrup investigated the use of yeast surface display and directed evolution for the selection and affinity maturation of ^{10}Fn3 variants. With randomization of only 14 residues and libraries of size 10^7–10^9, they were able to isolate and affinity mature molecules with binding affinities as high as 350 pM for the model antigen hen egg-white lysozyme. A striking feature of the highest-affinity antibody mimic was a pair of cysteine residues on adjacent loops, positions 28 and 77. This cysteine pair was found in experiments to be critical for the high-affinity binding. We began a collaboration to use molecular modeling to address whether the cysteine pair is likely to be forming a disulfide bond in solution. In addition, we estimated the potential for other inter-loop pairs of positions to form disulfide bonds.

A.2 Results

Disulfide bonds were modeled between all 49 pairs of BC loop and FG loop positions using an NMR average structure of ^{10}Fn3 as a template. For each pair of positions, two free cysteines were modeled for reference in addition to the disulfide-bonded form. Due to the lack of experimental data on side-chain conformation in the selected loops, the remaining twelve BC-loop and FG-loop positions were modeled as either all glycine, all alanine, or all serine. Each structure was subjected to a restrained minimization protocol. The final total energy of the disulfide bond relative to the free

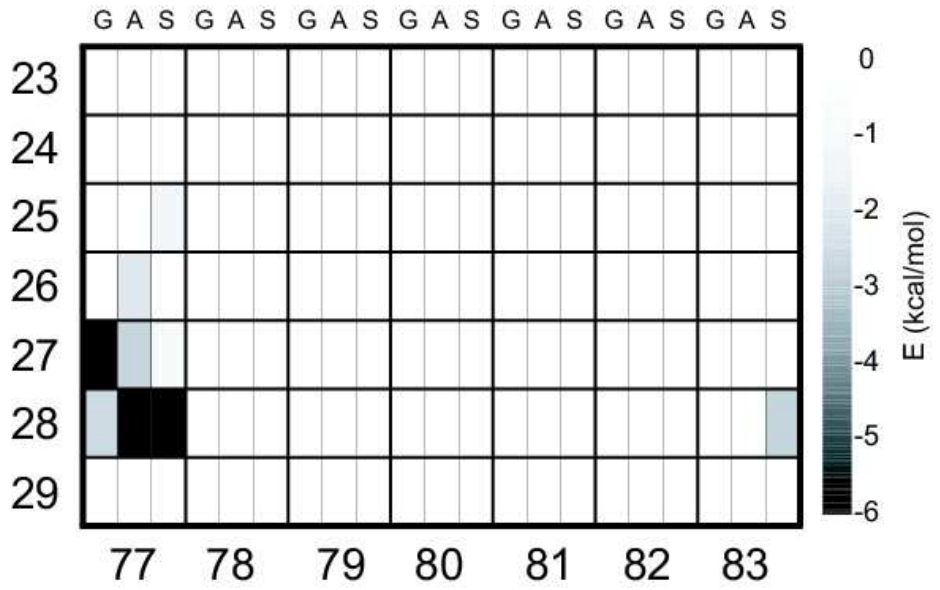


Figure A.1: Modeling of disulfide bonds between cysteines in BC and FG loops. The numbers on the vertical axis refer to the position of the cysteine in the BC loop, and the numbers on the horizontal axis refer to the position of the cysteine in the FG loop. Energies of disulfide formation for each pair of BC and FG cysteines were calculated following restrained minimization using CHARMM, with all the remaining residues in the BC and FG loops set to glycine (first column for a given pair of residues), alanine (second column), or serine (third column). Calculated energies are color-coded, with a darker shade of gray corresponding to a lower (more negative) energy, and thus to a higher likelihood of formation of the disulfide bond. White: 0 kcal/mol or higher; black: −6 kcal/mol or lower.

cysteines is shown in Figure A.1, and structural statistics for models with favorable disulfides are presented in Table A.1.

A disulfide bond between the BC loop and the FG loop was calculated to be favorable for only five pairs of positions. Only for the pair 28-77 was a disulfide bond found to be favorable across all three side chain backgrounds, and this disulfide was top-ranked for both the alanine and serine backgrounds. The glycine background most favored a disulfide between positions 27 and 77.

The model of the 28-77 disulfide bond with alanine at the other twelve BC loop and FG loop positions is shown in Figure A.2. Before minimization, this modeled structure was deformed due to a long sulfur–sulfur bond length, but otherwise it was free of steric clashes. A standard disulfide conformation was achieved through the restrained minimization, primarily by rotation of the protein backbone at position 77.

Table A.1: Disulfide bonds between BC and FG loops of ^{10}Fn3 that are predicted to be energetically favorable.

Cys[1]	Back-ground[2]	C_α–C_α distance[3] (Å)			RMSD[4] (Å)			Energy[5]
		^{10}Fn3[6]	DiS[7]	Ref [8]	^{10}Fn3/diS	^{10}Fn3/Ref	DiS/Ref	(kcal/mol)
27-77	Gly	9.9	5.8	9.1	1.18	1.10	0.62	−9.4
28-77	Ser	8.1	5.6	7.8	1.18	1.11	0.52	−6.9
28-77	Ala	8.1	5.6	7.8	1.20	1.12	0.53	−5.6
28-83	Ser	12.1	6.4	11.0	1.26	1.11	0.73	−1.4
27-77	Ala	9.9	6.0	9.6	1.20	1.13	0.67	−1.3
28-77	Gly	8.1	5.7	6.8	1.17	1.12	0.54	−1.2
26-77	Ala	13.2	6.1	13.3	1.27	1.12	0.95	−0.8
25-77	Ser	12.6	6.3	12.4	1.29	1.12	0.98	−0.3
27-77	Ser	9.9	6.2	10.3	1.19	1.11	0.72	−0.2

[1]Positions of the two modeled cysteines in loops BC and FG.
[2]Amino-acid residues assigned to the remaining 12 positions in BC and FG loops.
[3]Distance between the C_α atoms at the positions identified in column Cys.
[4]Root mean square distance between two structures, including all equivalent atoms.
[5]Calculated energy of a disulfide bond relative to the free cysteines.
[6]NMR solution structure of wild-type human ^{10}Fn3.
[7]Modeled structure with a disulfide between the cysteines in the two positions listed under Cys.
[8]Modeled structure with two free cysteines in the two positions listed under Cys.

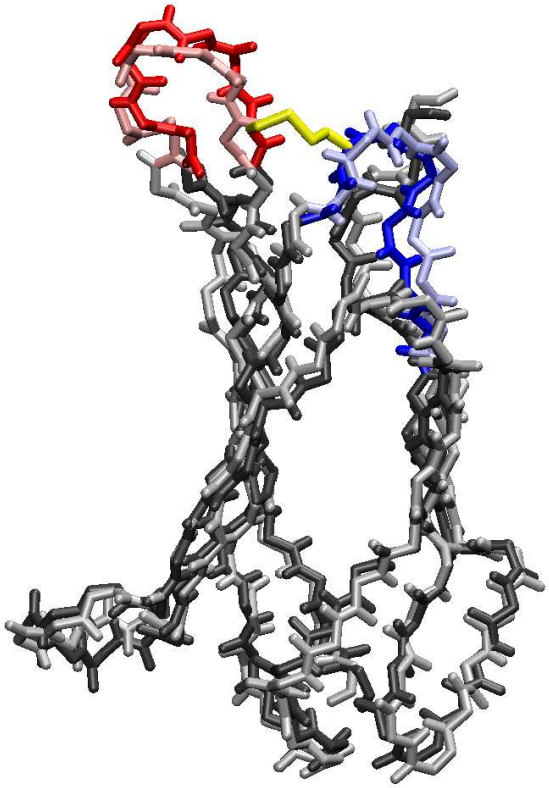


Figure A.2: Model of a ^{10}Fn3 variant with a disulfide bond between Cys 28 and Cys 77, superimposed on the NMR structure of wild-type ^{10}Fn3. The disulfide model was generated using harmonically restrained minimization of the NMR structure with a disulfide bond built between cysteines at positions 28 and 77, and with alanine at the other twelve positions in BC and FG loops. Dark red: FG loop in C28-C77 model; dark blue: BC loop in C28-C77 model; dark gray: the rest of main chain of the C28-C77 model; yellow: disulfide bond between Cys 28 and Cys 77 in the C28-C77 model, excluding hydrogen atoms; light red: FG loop in wild-type ^{10}Fn3; light blue: BC loop in wild-type ^{10}Fn3; light gray: the rest of main chain of wild-type ^{10}Fn3. Side chains are not shown for clarity.

The goal of the disulfide-bond modeling was to suggest pairs of positions that are capable of forming a disulfide bond, not to rule out pairs of positions that cannot form a disulfide bond. Several simplifications used in the calculations may result in missing a low-energy conformation that would favor a particular disulfide bond. The glycine, alanine, and serine backgrounds used in modeling are only three of 4×10^{15} possible sequences for each disulfide, and the minimization protocol undersamples side chain and backbone conformational space. Nevertheless, the modeling demonstrates that disulfide bonds can be introduced between several positions on the BC and the FG loop without disrupting the ^{10}Fn3 fold. The cysteine pair found in the high-affinity clone, Cys 28 and Cys 77, corresponds to the lowest disulfide-bond energy in the alanine and serine backgrounds, and to one of the two favorable disulfide-bond energies in the glycine background. Interestingly, Cys 27 and Cys 77, which correspond to the lowest calculated disulfide-bond energy in the glycine background, were found in another selected clone.

A.3 Discussion

The most striking feature in the sequences in lysozyme-binding antibody mimics was the prevalence of cysteine, which was present in selected clones at more than twice its frequency in naive libraries. We hypothesized that two cysteines on adjacent CDR-like loops of ^{10}Fn3 form a disulfide bond, which contributes to both thermodynamic stability of the ^{10}Fn3 fold and to its binding to hen egg-white lysozyme. The hypothesis that two cysteines on adjacent CDR-like loops of ^{10}Fn3 form a disulfide bond was tested using structural modeling, separate affinity-maturation experiments, and separate site-directed mutagenesis; all three approaches yielded results consistent with such a disulfide bond.

The most likely mechanism in which the 28-77 disulfide contributes to lysozyme binding is entropic, by reducing the flexibility of loops BC and FG in unbound state. Nevertheless, it is also possible that the disulfide itself is a part of the lysozyme-binding surface on the antibody-mimic molecule. Indirect support for this possibility

comes from the presence of two adjacent cysteines in the BC loop of a different isolated variant (K_{d} = 10 nM), which may form a vicinal disulfide bond [142]. An affinity maturation strategy that fixed the FG loop (except for position 77, which was randomized between serine and cysteine) and re-randomized the BC loop did not yield any mutants with a Cys 28 – Cys 77 pair, suggesting that most of the binding energy of the variant to lysozyme was provided by its selected BC loop, or by the combination of its BC loop and the 28-77 disulfide.

Whereas this is the first report of an interloop disulfide bond selected in ^{10}Fn3-based antibody mimics, analogous disulfides have been discovered in naturally evolved variable antibody domains from camelids (llamas and camels) [143, 144] and in new antigen receptors (IgNAR) from cartilaginous fish (sharks, skates, rays, and chimaeras) [145, 146]. In contrast to variable domains produced by immune systems of most other animals, which bind antigens in $V_{\mathrm{H}}/V_{\mathrm{L}}$ pairs , camelid and IgNAR variable domains bind antigens singly. A possible explanation for the prevalence of interloop disulfides in these domains is that the added structural constraint of an interloop disulfide contributes to domain stability [146–149], rigidity of the antigen-binding site [144, 147, 148], or structural diversity [145, 147, 149], all of which may contribute to antigen binding. The evolutionary distance between camelids and cartilaginous fish and the rarity of interloop disulfides in non-camelid mammals suggest that the interloop disulfides present in both types of single-variable-domain antibodies arose by convergent evolution [146, 148].

The *in-vitro* selection of ^{10}Fn3-based antibody mimics is the third example of convergent evolution to an interloop disulfide under pressure of antigen binding by a single-domain, immunoglobulin-like structure. Furthermore, this is the first system in which the role of such a disulfide bond has been tested experimentally; we have demonstrated the loss of antigen binding when the disulfide is disrupted by site-directed mutagenesis, and the tendency of the system to re-evolve the disulfide after the cysteine positions are randomized during affinity maturation. We expect that the analogous interloop disulfides in natural camelid and shark antibodies will prove similarly critical for their antigen binding.

A.4 Methods

The NMR, restrained energy-minimized average structure of the tenth type III module of human fibronectin (PDB code 1TTG [140]) was prepared for computational modeling by removing all water molecules and building on hydrogen atoms using the HBUILD facility [51] and the PARAM22 all-atom parameter set [90] in CHARMM [52]. Disulfide bonds were modeled between all 49 pairs of BC loop and FG loop positions. For each pair of positions, two models were created: one with a disulfide bond, and one with two free cysteines. The remaining twelve positions in BC and FG loop were modeled as either all glycine, all alanine, or all serine. The cysteine and serine side chains were built in their default conformation.

The structure with two free cysteines and the structure with a disulfide bond were each energy-minimized using no non-bonded cut-offs and a 4r distance-dependent dielectric constant. The minimization protocol included a harmonic restraint of 0.1 kcal/mol/Å^2 on each atom in the protein to preserve the native structure where possible and to reduce the effects of the imperfect energy function. Each structure was minimized to convergence, the harmonic restraints were then reset for the new atomic coordinates, and the process was repeated for a total of three complete minimizations. Similar results were obtained with two or four rounds of minimization.

The energy reported for a disulfide bond between a particular pair of positions is the difference between total energies, after minimization, of the disulfide-bonded structure and the corresponding structure with two free cysteines. Similar results were obtained when directly using the final energy after minimization for each disulfide. RMSD calculations were performed between all atoms common to both structures, using the McLachlan algorithm [150] as implemented in the program ProFit V2.2 (Martin, A.C.R., http://www.bioinf.org.uk/software/profit/). Figure A.2 was drawn using Visual Molecular Dynamics [151].

Bibliography

[1] J. Foote and H. N. Eisen. Kinetic and affinity limits on antibodies produced during immune responses. *Proc. Natl. Acad. Sci. U.S.A.*, 92(5):1254–1256, 1995.

[2] A. M. Levin and G. A. Weiss. Optimizing the affinity and specificity of proteins with molecular display. *Mol. BioSyst.*, 2(1):49–57, 2006.

[3] R. E. Hawkins, S. J. Russell, M. Baier, and G. Winter. The contribution of contact and non-contact residues of antibody in the affinity of binding to antigen. the interaction of mutant D1.3 antibodies with lysozyme. *J. Mol. Biol.*, 234(4):958–64, 1993.

[4] J. J. VanAntwerp and K. D. Wittrup. Fine affinity discrimination by yeast surface display and flow cytometry. *Biotechnol. Progr.*, 16(1):31–37, 2000.

[5] D. Lipovsek, S. M. Lippow, B. J. Hackel, M. W. Gregson, P. Cheng, A. Kapila, and K. D. Wittrup. Evolution of an interloop disulfide bond in high-affinity antibody mimics based on fibronectin type III domain and selected by yeast surface display: Molecular convergence with single-domain camelid and shark antibodies. *J. Mol. Biol.*, In Press, 2007.

[6] A. Razai, C. Garcia-Rodriguez, J. Lou, I. N. Geren, C. M. Forsyth, Y. Robles, R. Tsai, T. J. Smith, L. A. Smith, R. W. Siegel, M. Feldhaus, and J. D. Marks. Molecular evolution of antibody affinity for sensitive detection of botulinum neurotoxin type A. *J. Mol. Biol.*, 351(1):158–169, 2005.

[7] K. S. Midelfort and K. D. Wittrup. Context-dependent mutations predominate in an engineered high-affinity single chain antibody fragment. *Protein Sci.*, 15(2):324–34, 2006.

[8] E. T. Boder, K. S. Midelfort, and K. D. Wittrup. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl. Acad. Sci. U.S.A.*, 97(20):10701–5, 2000.

[9] C. P. Graff, K. Chester, R. Begent, and K. D. Wittrup. Directed evolution of an anti-carcinoembryonic antigen scFv with a 4-day monovalent dissociation half-time at 37 degrees C. *Protein Eng. Des. Sel.*, 17(4):293–304, 2004.

[10] M. Schmidt and K. D. Wittrup. Unpublished results, 2007.

[11] M. J. Feldhaus, R. W. Siegel, L. K. Opresko, J. R. Coleman, J. M. Feldhaus, Y. A. Yeung, J. R. Cochran, P. Heinzelman, D. Colby, J. Swers, C. Graff, H. S. Wiley, and K. D. Wittrup. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat. Biotechnol.*, 21(2):163–70, 2003.

[12] M. C. Kieke, E. Sundberg, E. V. Shusta, R. A. Mariuzza, K. D. Wittrup, and D. M. Kranz. High affinity T cell receptors from yeast display libraries block T cell activation by superantigens. *J. Mol. Biol.*, 307(5):1305–15, 2001.

[13] R. A. Buonpane, B. Moza, E. J. Sundberg, and D. M. Kranz. Characterization of T cell receptors engineered for high affinity against toxic shock syndrome toxin-1. *J. Mol. Biol.*, 353(2):308–321, 2005.

[14] S. M. Malakauskas and S. L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.*, 5(6):470–475, 1998.

[15] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

[16] J. R. Desjarlais and T. M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci.*, 4(10):2006–2018, 1995.

[17] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[18] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282(5393):1462–7, 1998.

[19] J. Reina, E. Lacroix, S. D. Hobson, G. Fernandez-Ballester, V. Rybin, M. S. Schwab, L. Serrano, and C. Gonzalez. Computer-aided design of a PDZ domain to recognize new target sequences. *Nat. Struct. Biol.*, 9(8):621–627, 2002.

[20] L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–190, 2003.

[21] W. Yang, A. L. Wilkins, Y. M. Ye, Z. R. Liu, S. Y. Li, J. L. Urbauer, H. W. Hellinga, A. Kearney, P. A. van der Merwe, and J. J. Yang. Design of a calcium-binding protein with desired structure in a cell adhesion molecule. *J. Am. Chem. Soc.*, 127(7):2085–2093, 2005.

[22] F. V. Cochran, S. P. Wu, W. Wang, V. Nanda, J. G. Saven, M. J. Therien, and W. F. DeGrado. Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. *J. Am. Chem. Soc.*, 127(5):1346–1347, 2005.

[23] J. M. Shifman and S. L. Mayo. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl. Acad. Sci. U.S.A.*, 100(23):13274–13279, 2003.

[24] L. A. Joachimiak, T. Kortemme, B. L. Stoddard, and D. Baker. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein–protein interface. *J. Mol. Biol.*, 361(1):195–208, 2006.

[25] J. J. Havranek and P. B. Harbury. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.*, 10(1):45–52, 2003.

[26] D. N. Bolon, R. A. Grant, T. A. Baker, and R. T. Sauer. Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 102(36):12724–9, 2005.

[27] J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, B. L. Stoddard, and D. Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659, 2006.

[28] X. I. Ambroggio and B. Kuhlman. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J. Am. Chem. Soc.*, 128(4):1154–1161, 2006.

[29] D. F. Green, A. T. Dennis, P. S. Fam, B. Tidor, and A. Jasanoff. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry*, 45(41):12547–59, 2006.

[30] J. K. Lassila, J. R. Keeffe, P. Oelschlaeger, and S. L. Mayo. Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. *Protein Eng. Des. Sel.*, 18(4):161–163, 2005.

[31] P. Oelschlaeger and S. L. Mayo. Hydroxyl groups in the beta beta sandwich of metallo-beta-lactamases favor enzyme activity: A computational protein design study. *J. Mol. Biol.*, 350(3):395–401, 2005.

[32] A. Korkegian, M. E. Black, D. Baker, and B. L. Stoddard. Computational thermostabilization of an enzyme. *Science*, 308(5723):857–860, 2005.

[33] A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Rothlisberger, and D. Baker. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.*, 15(12):2785–2794, 2006.

[34] J. K. Lassila, H. K. Privett, B. D. Allen, and S. L. Mayo. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.*, 103(45):16710–16715, 2006.

[35] M. A. Dwyer, L. L. Looger, and H. W. Hellinga. Computational design of a biologically active enzyme. *Science*, 304(5679):1967–1971, 2004.

[36] G. Song, G. A. Lazar, T. Kortemme, M. Shimaoka, J. R. Desjarlais, D. Baker, and T. A. Springer. Rational design of intercellular adhesion molecule-1 (ICAM-1) variants for antagonizing integrin lymphocyte function-associated antigen-1-dependent adhesion. *J. Biol. Chem.*, 281(8):5042–5049, 2006.

[37] V. D. Sood and D. Baker. Recapitulation and design of protein binding peptide structures and sequences. *J. Mol. Biol.*, 357(3):917–927, 2006.

[38] R. W. Dixon, R. J. Radmer, B. Kuhn, P. A. Kollman, J. Yang, C. Raposo, C. S. Wilcox, L. A. Klumb, P. S. Stayton, C. Behnke, I. Le Trong, and R. Stenkamp. Theoretical and experimental studies of biotin analogues that bind almost as tightly to streptavidin as biotin. *J. Org. Chem.*, 67(6):1827–37, 2002.

[39] L. A. Clark, P. A. Boriack-Sjodin, J. Eldredge, C. Fitch, B. Friedman, K. J. Hanf, M. Jarpe, S. F. Liparoto, Y. Li, A. Lugovskoy, S. Miller, M. Rushe, W. Sherman, K. Simon, and H. Van Vlijmen. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci.*, 15(5):949–60, 2006.

[40] C. Pabo. Molecular technology. designing proteins and peptides. *Nature*, 301(5897):200, 1983.

[41] K. E. Drexler. Molecular engineering — an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.*, 78(9):5275–5278, 1981.

[42] K. S. Midelfort, H. H. Hernandez, S. M. Lippow, B. Tidor, C. L. Drennan, and K. D. Wittrup. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J. Mol. Biol.*, 343(3):685–701, 2004.

[43] R. M. Levy, L. Y. Zhang, E. Gallicchio, and A. K. Felts. On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. *J. Am. Chem. Soc.*, 125(31):9523–30, 2003.

[44] G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz, and R. C. Stevens. Structural insights into the evolution of an antibody combining site. *Science*, 276(5319):1665–1669, 1997.

[45] H. R. Wu, G. Beuerlein, Y. Nie, H. Smith, B. A. Lee, M. Hensler, W. D. Huse, and J. D. Watkins. Stepwise in vitro affinity maturation of Vitaxin, an alpha(v)beta(3)-specific humanized mAb. *Proc. Natl. Acad. Sci. U.S.A.*, 95(11):6037–6042, 1998.

[46] J. Yin, E. C. Mundorff, P. L. Yang, K. U. Wendt, D. Hanway, R. C. Stevens, and P. G. Schultz. A comparative analysis of the immunological evolution of antibody 28B4. *Biochemistry*, 40(36):10764–10773, 2001.

[47] J. Valjakka, A. Hemminki, S. Niemi, H. Soderlund, K. Takkinen, and J. Rouvinen. Crystal structure of an in vitro affinity- and specificity-matured anti-testosterone Fab in complex with testosterone — improved affinity results from small structural changes within the variable domains. *J. Biol. Chem.*, 277(46):44021–44027, 2002.

[48] Y. L. Li, H. M. Li, F. Yang, S. J. Smith-Gill, and R. A. Mariuzza. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat. Struct. Biol.*, 10(6):482–488, 2003.

[49] T. Sagawa, M. Oda, M. Ishimura, K. Furukawa, and T. Azuma. Thermodynamic and kinetic aspects of antibody evolution during the immune response to hapten. *Mol. Immunol.*, 39(13):801–808, 2003.

[50] M. Whitlow, A. J. Howard, J. F. Wood, Jr. Voss, E. W., and K. D. Hardman. 1.85 A structure of anti-fluorescein 4-4-20 Fab. *Protein Eng.*, 8(8):749–61, 1995.

[51] A. T. Brunger and M. Karplus. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins*, 4(2):148–56, 1988.

[52] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM — a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[53] F. A. Momany and R. Rone. Validation of the general-purpose QUANTA(R)3.2/CHARMM(R) force-field. *J. Comput. Chem.*, 13(7):888–900, 1992.

[54] D. F. Green and B. Tidor. Evaluation of ab initio charge determination methods for use in continuum solvation calculations. *J. Phys. Chem. B*, 107(37):10261–10273, 2003.

[55] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, Ö. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komáromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian 98* (Gaussian, Inc., Pittsburgh, PA, 1998).

[56] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges — the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[57] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.*, 8(7):1381–92, 1999.

[58] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, 94(19):7684–7692, 1990.

[59] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules — theory and applications. *Annu. Rev. Biophys. Bio.*, 19:301–332, 1990.

[60] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution - method and error assessment. *J. Comput. Chem.*, 9(4):327–335, 1988.

[61] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system — solvation energies, binding-energies, and conformational-analysis. *Proteins*, 4(1):7–18, 1988.

[62] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.

[63] G. A. Lazar, J. R. Desjarlais, and T. M. Handel. De novo design of the hydrophobic core of ubiquitin. *Protein Sci.*, 6(6):1167–1178, 1997.

[64] M. Shimaoka, J. M. Shifman, H. Jing, J. Takagi, S. L. Mayo, and T. A. Springer. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.*, 7(8):674–8, 2000.

[65] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci.*, 5(5):895–903, 1996.

[66] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 94(19):10172–10177, 1997.

[67] B. I. Dahiyat, D. B. Gordon, and S. L. Mayo. Automated design of the surface positions of protein helices. *Protein Sci.*, 6(6):1333–1337, 1997.

[68] K. Raha, A. M. Wollacott, M. J. Italia, and J. R. Desjarlais. Prediction of amino acid sequence from structure. *Protein Sci.*, 9(6):1106–1119, 2000.

[69] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U.S.A.*, 97(19):10383–10388, 2000.

[70] S. A. Marshall and S. L. Mayo. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, 305(3):619–631, 2001.

[71] J. M. Shifman and S. L. Mayo. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.*, 323(3):417–423, 2002.

[72] C. M. Kraemer-Pecore, J. T. J. Lecomte, and J. R. Desjarlais. A de novo redesign of the WW domain. *Protein Sci.*, 12(10):2194–2205, 2003.

[73] W. Yang, L. M. Jones, L. Isley, Y. M. Ye, H. W. Lee, A. Wilkins, Z. R. Liu, H. W. Hellinga, R. Malchow, M. Ghazi, and J. J. Yang. Rational design of a calcium-binding protein. *J. Am. Chem. Soc.*, 125(20):6165–6171, 2003.

[74] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U.S.A.*, 101(18):6946–6951, 2004.

[75] A. V. Morozov, J. J. Havranek, D. Baker, and E. D. Siggia. Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, 33(18):5781–5798, 2005.

[76] L. Jiang, B. Kuhlman, T. A. Kortemme, and D. Baker. A solvated rotamer approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins*, 58(4):893–904, 2005.

[77] A. V. Morozov and T. Kortemme. Potential functions for hydrogen bonds in protein structure prediction and design. In *Peptide Solvation and H-Bonds*, volume 72 of *Advances in Protein Chemistry*, pages 1–+. 2006.

[78] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.*, 332(2):449–460, 2003.

[79] M. A. Dwyer, L. L. Looger, and H. W. Hellinga. Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 100(20):11255–11260, 2003.

[80] J. J. Havranek, C. M. Duarte, and D. Baker. A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.*, 344(1):59–70, 2004.

[81] N. Dobson, G. Dantas, D. Baker, and G. Varani. High-resolution structural validation of the computational redesign of human U1A protein. *Structure*, 14(5):847–856, 2006.

[82] X. Z. Hu and B. Kuhlman. Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins*, 62(3):739–748, 2006.

[83] J. Mendes, R. Guerois, and L. Serrano. Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, 12(4):441–446, 2002.

[84] T. Kortemme and D. Baker. Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.*, 8(1):91–97, 2004.

[85] C. L. Vizcarra and S. L. Mayo. Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.*, 9(6):622–626, 2005.

[86] S. A. Marshall, C. L. Vizcarra, and S. L. Mayo. One- and two-body decomposable Poisson–Boltzmann methods for protein design calculations. *Protein Sci.*, 14(5):1293–1304, 2005.

[87] D. B. Gordon, S. A. Marshall, and S. L. Mayo. Energy functions for protein design. *Curr. Opin. Struct. Biol.*, 9(4):509–513, 1999.

[88] K. J. M. Hanf. *Protein design with hierarchical treatment of solvation and electrostatics*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, U.S.A., 2002.

[89] T. N. Bhat, G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dallacqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak. Bound water molecules and conformational stabilization help mediate an antigen–antibody association. *Proc. Natl. Acad. Sci. U.S.A.*, 91(3):1089–1093, 1994.

[90] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[91] Jr. Dunbrack, R. L. and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–81, 1997.

[92] R. L. Dunbrack. Rotamer libraries in the 21(st) century. *Curr. Opin. Struct. Biol.*, 12(4):431–440, 2002.

[93] I. Lasters, M. Demaeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side-chains. *Protein Eng.*, 8(8):815–822, 1995.

[94] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.

[95] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.

[96] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.*, 66(5):1335–1340, 1994.

[97] D. B. Gordon and S. L. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, 19(13):1505–1514, 1998.

[98] D. B. Gordon and S. L. Mayo. Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure*, 7(9):1089–1098, 1999.

[99] D. B. Gordon, G. K. Hom, S. L. Mayo, and N. A. Pierce. Exact rotamer optimization for protein design. *J. Comput. Chem.*, 24(2):232–243, 2003.

[100] J. Mendes, A. M. Baptista, M. A. Carrondo, and C. M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins*, 37(4):530–543, 1999.

[101] M. D. Altman. *Computational ligand design and analysis in protein complexes using inverse methods, combinatorial search, and accurate solvation modeling.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, U.S.A., 2006.

[102] E. T. Boder and K. D. Wittrup. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.*, 15(6):553–7, 1997.

[103] E. T. Boder and K. D. Wittrup. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Method. Enzymol.*, 328:430–44, 2000.

[104] G. Chao, W. L. Lau, B. J. Hackel, S. L. Sazinsky, S. M. Lippow, and K. D. Wittrup. Isolating and engineering human antibodies using yeast surface display. *Nature Protocols*, 1(2):755–768, 2006.

[105] J. D. Sato, T. Kawamoto, A. D. Le, J. Mendelsohn, J. Polikoff, and G. H. Sato. Biological effects in vitro of monoclonal antibodies to human epidermal growth factor receptors. *Mol. Biol. Med.*, 1(5):511–29, 1983.

[106] S. Li, K. R. Schmitz, P. D. Jeffrey, J. J. 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.

[107] M. L. Jaramillo, Z. Leon, S. Grothe, B. Paul-Roc, A. Abulrob, and M. O'Connor McCourt. Effect of the anti-receptor ligand-blocking 225 monoclonal antibody on EGF receptor endocytosis and sorting. *Exp. Cell Res.*, 312(15):2778–90, 2006.

[108] Y. S. Kim, R. Bhandari, J. R. Cochran, J. Kuriyan, and K. D. Wittrup. Directed evolution of the epidermal growth factor receptor extracellular domain for expression in yeast. *Proteins*, 62(4):1026–1035, 2006.

[109] L. G. Presta, H. Chen, S. J. O'Connor, V. Chisholm, Y. G. Meng, L. Krummen, M. Winkler, and N. Ferrara. Humanization of an anti-vascular endothelial growth factor monoclonal antibody for the therapy of solid tumors and other disorders. *Cancer Res.*, 57(20):4593–9, 1997.

[110] N. Ferrara, K. J. Hillan, H. P. Gerber, and W. Novotny. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. *Nat. Rev. Drug Discov.*, 3(5):391–400, 2004.

[111] Y. Chen, C. Wiesmann, G. Fuh, B. Li, H. W. Christinger, P. McKay, A. M. de Vos, and H. B. Lowman. 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.

[112] T. Selzer, S. Albeck, and G. Schreiber. Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, 7(7):537–541, 2000.

[113] B. A. Joughin, D. F. Green, and B. Tidor. Action-at-a-distance interactions enhance protein binding affinity. *Protein Sci.*, 14(5):1363–1369, 2005.

[114] J. A. Wagoner and N. A. Baker. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. U.S.A.*, 103(22):8331–6, 2006.

[115] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.*, 23(5):517–29, 2002.

[116] E. Gallicchio and R. M. Levy. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.*, 25(4):479–499, 2004.

[117] B. C. Braden, H. Souchon, J. L. Eisele, G. A. Bentley, T. N. Bhat, J. Navaza, and R. J. Poljak. 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.

[118] Y. A. Muller, Y. Chen, H. W. Christinger, B. Li, B. C. Cunningham, H. B. Lowman, and A. M. de Vos. 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.

[119] W. Wels, R. Beerli, P. Hellmann, M. Schmidt, B. M. Marte, E. S. Kornilova, A. Hekele, J. Mendelsohn, B. Groner, and N. E. Hynes. EGF receptor and p185erbB-2-specific single-chain antibody toxins differ in their cell-killing activity on tumor cells expressing both receptor proteins. *Int. J. Cancer*, 60(1):137–44, 1995.

[120] D. W. Colby, B. A. Kellogg, C. P. Graff, Y. A. Yeung, J. S. Swers, and K. D. Wittrup. Engineering antibody affinity by yeast surface display. *Method. Enzymol.*, 388:348–58, 2004.

[121] J. W. Pitera and W. F. van Gunsteren. The importance of solute–solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.*, 123(13):3163–3164, 2001.

[122] Y. Su and E. Gallicchio. The non-polar solvent potential of mean force for the dimerization of alanine dipeptide: the role of solute–solvent van der waals interactions. *Biophys. Chem.*, 109(2):251–260, 2004.

[123] J. Wagoner and N. A. Baker. Solvation forces on biomolecular structures: A comparison of explicit solvent and Poisson–Boltzmann models. *J. Comput. Chem.*, 25(13):1623–1629, 2004.

[124] H. S. Ashbaugh, E. W. Kaler, and M. E. Paulaitis. A universal surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.*, 121(39):9243–9244, 1999.

[125] E. Gallicchio, M. M. Kubo, and R. M. Levy. Enthalpy–entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B*, 104(26):6271–6285, 2000.

[126] A. K. Felts, Y. Harano, E. Gallicchio, and R. M. Levy. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins*, 56(2):310–21, 2004.

[127] J. Pitarch, V. Moliner, J.-L. Pascual-Ahuir, E. Silla, and I. Tunon. Can hydrophobic interactions be correctly reproduced by the continuum models? *J. Phys. Chem.*, 100(23):9955–9959, 1996.

[128] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[129] W. L. Jorgensen and J. D. Madura. Temperature and size dependence for Monte-Carlo simulations of TIP4P water. *Mol. Phys.*, 56(6):1381–1392, 1985.

[130] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[131] S. Cabani, P. Gianni, V. Mollica, and L. Lepori. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.*, 10(8):563–595, 1981.

[132] G. H. Cohen, E. W. Silverton, E. A. Padlan, F. Dyda, J. A. Wibbenmeyer, R. C. Willson, and D. R. Davies. Water molecules in the antibody–antigen interface of the structure of the Fab HyHEL-5–lysozyme complex at 1.7 Angstrom resolution: comparison with results from isothermal titration calorimetry. *Acta Crystallogr. D*, 61:628–633, 2005.

[133] Robert B. Hermann. Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 74(10):4144–4145, 1977.

[134] A. Koide, C. W. Bailey, X. L. Huang, and S. Koide. The fibronectin type III domain as a scaffold for novel binding proteins. *J. Mol. Biol.*, 284(4):1141–1151, 1998.

[135] J. Richards, M. Miller, J. Abend, A. Koide, S. Koide, and S. Dewhurst. Engineered fibronectin type III domain with a RGDWXE sequence binds with enhanced affinity and specificity to human $\alpha v\beta 3$ integrin. *J. Mol. Biol.*, 326(5):1475–1488, 2003.

[136] E. Karatan, M. Merguerian, Z. H. Han, M. D. Scholle, S. Koide, and B. K. Kay. Molecular recognition properties of Fn3 monobodies that bind the Src SH3 domain. *Chem. Biol.*, 11(6):835–844, 2004.

[137] A. Koide, S. Abbatiello, L. Rothgery, and S. Koide. Probing protein conformational changes in living cells by using designer binding proteins: Application to the estrogen receptor. *Proc. Natl. Acad. Sci. U.S.A.*, 99(3):1253–1258, 2002.

[138] M. H. Parker, Y. Chen, F. Danehy, K. Dufu, J. Ekstrom, E. Getmanova, J. Gokemeijer, L. Xu, and D. Lipovsek. Antibody mimics based on human fibronectin type three domain engineered for thermostability and high-affinity binding to vascular endothelial growth factor receptor two. *Protein Eng. Des. Sel.*, 18(9):435–444, 2005.

[139] L. H. Xu, P. Aha, K. Gu, R. G. Kuimelis, M. Kurz, T. Lam, A. C. Lim, H. X. Liu, P. A. Lohse, L. Sun, S. Weng, R. W. Wagner, and D. Lipovsek. Directed evolution of high-affinity antibody mimics using mRNA display. *Chem. Biol.*, 9(8):933–942, 2002.

[140] A. L. Main, T. S. Harvey, M. Baron, J. Boyd, and I. D. Campbell. The 3-dimensional structure of the 10th type-III module of fibronectin — an insight into RGD-mediated interactions. *Cell*, 71(4):671–678, 1992.

[141] C. D. Dickinson, B. Veerapandian, X. P. Dai, R. C. Hamlin, N. H. Xuong, E. Ruoslahti, and K. R. Ely. Crystal-structure of the 10th type-III cell-adhesion module of human fibronectin. *J. Mol. Biol.*, 236(4):1079–1092, 1994.

[142] O. Carugo, M. Cemazar, S. Zahariev, I. Hudaky, Z. Gaspari, A. Perczel, and S. Pongor. Vicinal disulfide turns. *Protein Eng.*, 16(9):637–639, 2003.

[143] S. Muyldermans, T. Atarhouch, J. Saldanha, Jarg Barbosa, and R. Hamers. Sequence and structure of V-H domain from naturally-occurring camel heavy-chain immunoglobulins lacking light-chains. *Protein Eng.*, 7(9):1129–1135, 1994.

[144] A. Desmyter, T. R. Transue, M. A. Ghahroudi, M. H. D. Thi, F. Poortmans, R. Hamers, S. Muyldermans, and L. Wyns. Crystal structure of a camel single-domain V-H antibody fragment in complex with lysozyme. *Nat. Struct. Biol.*, 3(9):803–811, 1996.

[145] R. L. Stanfield, H. Dooley, M. F. Flajnik, and I. A. Wilson. Crystal structure of a shark single-domain antibody V region in complex with lysozyme. *Science*, 305(5691):1770–1773, 2004.

[146] K. H. Roux, A. S. Greenberg, L. Greene, L. Strelets, D. Avila, E. C. McKinney, and M. F. Flajnik. Structural analysis of the nurse shark (new) antigen receptor (NAR): Molecular convergence of NAR and unusual mammalian immnnoglobulins. *Proc. Natl. Acad. Sci. U.S.A.*, 95(20):11804–11809, 1998.

[147] K. B. Vu, M. A. Ghahroudi, L. Wyns, and S. Muyldermans. Comparison of llama V-H sequences from conventional and heavy chain antibodies. *Mol. Immunol.*, 34(16-17):1121–1131, 1997.

[148] V. A. Streltsov, J. N. Varghese, J. A. Carmichael, R. A. Irving, P. J. Hudson, and S. D. Nuttall. Structural evidence for evolution of shark Ig new antigen receptor variable domain antibodies from a cell-surface receptor. *Proc. Natl. Acad. Sci. U.S.A.*, 101(34):12444–12449, 2004.

[149] M. Diaz, R. L. Stanfield, A. S. Greenberg, and M. F. Flajnik. Structural analysis, selection, and ontogeny of the shark new antigen receptor (IgNAR): identification of a new locus preferentially expressed in early development. *Immunogenetics*, 54(7):501–512, 2002.

[150] A. D. McLachlan. Rapid comparison of protein structures. *Acta Crystallogr. A*, 38(NOV):871–873, 1982.

[151] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.