© 2021

Zachary Fritz

ALL RIGHTS RESERVED

DEVELOPMENT AND APPLICATIONS OF A COARSE-GRAINED PROTEIN ENERGETICS MODEL

By

ZACHARY R. FRITZ

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Biomedical Engineering

Written under the direction of

Martin Yarmush

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2021

ABSTRACT OF THE DISSERTATION

DEVELOPMENT AND APPLICATIONS OF A COARSE-GRAINED PROTEIN ENERGETICS MODEL

By: ZACHARY FRITZ

Dissertation Director:

Martin L. Yarmush

Proteins carry out a staggering number of functions within the human body and the biological world at large, and are often the primary targets of drugs in treatment of disease. For this reason, an understanding of protein behavior—how a protein's sequence and structure determines its stability, function, and interactions with other molecules—is critical for pharmaceutical development and other biotech-based industries. Recent advancements in some areas of protein modelling, especially the application of neural-network based machine learning to protein structure prediction, have been very promising, but there is still much work to do to fully understand how proteins fold. Other protein modelling techniques, like molecular dynamics simulations, are powerful but hampered by very short timescales that don't capture the full spectrum of protein behavior. Additionally, many of these techniques are computationally intensive, require a high degree of user expertise, and do not breakdown protein energy contributions at the amino acid level.

We have developed a coarse-grained protein energetics model, the Hidden Symmetry Model (HSyM), that is able to extract per-residue interaction energy data from sequence and structure data. The model is scale invariant and only requires a single structure of a protein's native

conformation; consequently, its calculations can take a matter of seconds and be done on a basic PC. We have demonstrated potential applications of HSyM by successfully using it to predict mutation-induced thermostability shifts in T4 lysozyme, and to predict the binding affinities of engineered peptides for an antibody with limited structural information for the molecules involved. Preliminary work on a fully integrated microfluidic device that would use these model-engineered peptides to carry out diagnostic blood assays is presented to show a potential clinical use for the model. We also present work to further optimize HSyM by using a simplified statistical mechanical "toy model" that takes into account solvent-residue interactions. We hope that with further refinement the Hidden Symmetry Model will have a far-reaching impact on the fields of computational drug design, protein engineering, and biomedical and biotechnology research in general.

Acknowledgement

I would first and foremost like to thank both of my advisers, Drs. Martin Yarmush and Lawrence Williams. They were instrumental in keeping me focused, encouraging me to fully embrace my project, and not letting minor setbacks (botched experiments, negative results) discourage me. They were also chiefly responsible for molding me into a much more confident scientist and scientific communicator.

I would also like thank the other two members of my dissertation committee, Drs. Rene Schloss and Jeffrey Zahn. I regarded Dr. Schloss as a sort of “lab mom”: she was always there to bounce ideas off of and provide much needed moral support. I definitely relied on both her wisdom and optimism to survive grad school. Dr. Zahn provided key insights for the microfluidic aspect of my project and general tips for my writing and presentations. I also greatly enjoyed taking some of his courses as a student.

Dr. Anil Shrirao was an indispensable part of my graduate school tenure; I am not sure what I would have done without his expertise in microfluidics, device design, and scientific communication. He was yet another who kept me afloat during rough times.

Also within the Biomedical Engineering Department I would like to thank Dr. Mark Pierce for his valuable assistance and guidance with the optical detection parts of the device work; Lawrence Stromberg and Mary Ellen Presa for all their help with academic and administrative duties; and my Graduate Director and Department Chair Drs. Joseph Freeman and David Shreiber.

I was grateful to have the opportunity to mentor outstanding undergraduates like Parinita Jain and Gustavo Rios-Delgado in the lab. I also have to thank the undergrad members of the

Williams lab: Yonatan Attali, Mohammad Omar Ani, and Nicholas Syracuse. Much of my later work with the protein energetics model would not have been possible without them.

Within the larger Rutgers community, I have to thank Drs. Nilgun Tumer, Xiao-ping Li, and Michael Pierce in the School of Environmental and Biological Sciences for instructing me in the technique of surface plasmon resonance and allowing me access to their powerful Biacore T200 (funding support from NIH grant 1S10OD026750-01). I also have to thank Dr. Zoltan Szekely in the College of Pharmacy for teaching me solid-support peptide synthesis and LC-MS.

I made many lasting friendships with my fellow students in the Yarmush and Berthiaume labs: Ileana Marrero-Berrios, Paulina Krzyszczyk, Mollie Davis, Xiomara Perez, Noah Chen, Hwan June Kang, Josh Leipheimer, Robert Rosen, and Talia Greenstein. Besides all the good times we had, all of these people have helped me in some way, whether it was teaching me an experimental technique or being an audience for me to practice my presentations on. I also owe much to my friends in other labs: Alison Acevedo, Cosmas Mwikirize, Maria Qadri, Elliot Dolan, Stephanie Fung, Joseph Molde, Fernando Rebolledo, Mariagemiliana Dessì, Ryan Guasp, Yoliem Alarcon, Meghan Arnold, Carolina Bobadilla, and Ilija Melentijevic.

I would like to thank Ann Stock and the Rutgers Biotechnology Training Program for providing academic, career prep, and financial support in the form of a fellowship and grant (NIH T32 GM135141), as well as Janet Alder and the Rutgers iJOBS program for additional career preparation support.

Last but certainly not least, I must thank my parents, George Fritz and Monica Camuñez, and my sister, Theresa Camuñez, for their love and unconditional support. They have always encouraged my passion for science and I would not have gotten this far without them.

Table of Contents

List of Tables

List of Illustrations

Chapter 1: Introduction

Why Understanding Protein Behavior Matters

Proteins have been called “nature’s robots” [1]. Besides providing key structural materials in cells and tissues, these molecular machines are responsible for carrying out the majority of biological processes, including signal transduction [2], intercellular communication [3], immune system recognition [4], regulating genetic transcription and translation [5], catalyzing biochemical reactions [6], and much more. Proteins are able to perform this vast diversity of functions due to (1) their ability to fold and assume a variety of conformations and (2) their ability to bind and interact with other molecules, including other proteins, nucleic acids, and small molecules [7-9]. Understanding and modeling the thermodynamics driving these two aspects of protein behavior is critical for unlocking the full potential of protein-based applications. For example, most drugs are targeted towards proteins involved in disease mechanisms and usually function by binding to proteins to inhibit their interactions with other molecules, mark them for recognition by other proteins or immune cells, and/or change or destabilize their conformation [10]. Knowledge of the target protein’s structure and thermodynamic organization (conformational stability and interaction energy distribution) is crucial for designing drugs that will bind to the correct site on the protein, bind with the desired affinity, and carry out their intended functions [11]. Additionally, the burgeoning field of rational protein engineering requires accurate modeling and prediction of how novel designed proteins will behave; these engineered molecules have applications in medicine/pharmaceuticals, agriculture [12], energy [13], food production and processing [14], and other fields.

The Current State of Protein Modeling

The term “protein modeling” is broad and refers to a variety of techniques with an array of different specializations and functions. These techniques can often be classified as “template based”, which rely on known structures and sequences of proteins, or “template-free/de novo” methods. It should be noted that both template and template-free methods have owed much to the expansion of structural and genomic databases like the Protein Data Bank, as even in de novo methods elements of known protein structures, sequence alignments, and data on correlated mutations can be used to help predict secondary structure features [15] and residue contact maps [16], or be used in training machine learning methods [17].

A protein’s structure and stability are inextricably tied to its energy landscape, as proteins have evolved to assume a native conformation, or “ground state”, that corresponds to a free energy minimum. Consequently, most protein models consist of two elements: a mathematical energy function based on the conformation of the atoms in the system (sometimes referred to as a microstate), and an algorithm for sampling this function and selecting preferred conformations [18]. Energy functions may make use of physics-based force fields that calculate the system’s energy based on the sum of all atomic interactions (e.g. covalent bonds, electrostatic effects, Van der Waals forces) in the microstate [19]. Knowledge-based or statistical force fields instead make use statistical analysis of data from protein structural databases to make predictions of atomic packing, contacts, and conformational features and make energy scores for a given conformation [20]. Additionally, energy functions can be coarse-grained, which typically merge side chain atoms into a single entity or only analyze proteins at the residue level, and result in a smoothed energy landscape [21]. These models are less computationally demanding than atomistic energy functions but are less accurate and may lead to false energy minima. For this reason many protein modeling packages and protocols, like ROSETTA, use a combination of

coarse-grained modeling for initial rough structure prediction and all-atom reconstructions for refinement [22].

Sampling algorithms range from simple gradient-based methods that always proceed with steps that lower free energy, to Monte Carlo methods that employ randomized substitutions and/or conformational changes and grade the favorability of these based on the magnitude of the calculated energy change (the Metropolis criterion) [18]. As with coarse-grained models, gradient-based sampling can lead to identifying false energy minima that do not actually correspond to a protein's ground state. Monte Carlo methods tend to be more accurate but require substantially more computing power to perform tasks like secondary structure fragment assembly and sidechain rotamer sampling. Molecular dynamics (MD) simulations constitute a special class of energy function sampling and are often used to further refine the results of other models. These programs use Newtonian forcefields (derived from the gradient of the potential energy function) applied to every atom or coarse-grained pseudoatom to calculate atomic trajectories iteratively over discrete, very small (on the order of a femtosecond) time steps [23]. While this technique may better recapitulate how proteins behave than other methods, it is very much limited by the small timescales it can accurately render, which are often much shorter than biologically relevant phenomena like protein folding, allosteric shifts, and other global protein motions. Longer duration MD runs, especially with large proteins, complexes, and/or explicit solvent molecules, can require the use of supercomputer clusters in order to handle the multitudes of calculations over the millions to trillions of timesteps necessary [21].

Two central applications of these models have involved elucidating the sequence-structure relationship of proteins: prediction of a protein's native structure from its sequence alone, and the converse of designing a protein sequence based on a known or desired structure. With regards to the former application, recent major advancements in machine learning methods like

artificial neural networks have yielded very promising results, culminating in DeepMind's AlphaFold scoring a record best median global distance test score of 92.4 out of 100 in the 2020 Critical Assessment of Protein Structure Prediction competition [24]. While some challenges remain, including accurate structure prediction of protein complexes, the path forward in structure prediction seems very encouraging. It remains to be seen whether these same machine learning tools will also revolutionize protein design and engineering. While both current template-based [25, 26] and de novo methods [27, 28] have yielded some success, many obstacles still hinder the field, including accurately modelling flexible proteins and domains, balancing polar and nonpolar interactions, modelling explicit water molecules, and creating functional enzymes [18]. Many of these issues are also present in the related field of structure-based computer-aided drug design (CADD), which uses techniques like molecular docking and MD to design small molecule drugs that bind to protein targets [29]. **Table 1.1** summarizes some of the currently used protein model components discussed in this section.

Model Component	Example(s)	Advantages	Disadvantages
Gradient-based sampling	Sidechain torsion angle sampling in a RosettaDock algorithm [30]	• Quickly finds local energy minima	• Not suitable for identifying global energy minima
Monte Carlo-based sampling	Sidechain rotamer sampling in ROSETTA3 [31]; protein chain movements in CABS [32]	• Able to avoid false energy minima • Generally faster than MD • Customizable for a particular application or system • Capable of large-scale or unphysical moves for broader sampling	• Not suitable for predicting binding kinetics or other time-dependent behavior • Not well suited for use with explicit solvent • High computational demand
Template-based Modeling	SWISS-MODEL [33]; MODELLER [34]	• Applicable to 60-70% of proteins • Useful for target protein/ligand binding design	• Limited to proteins with structures similar to known structures in the PDB
Molecular Dynamics (MD)	Desmond [35]; GROMACS [36]	• Highly accurate sampling of conformational space • Able to simulate explicit solvent molecules	• Not well suited for simulating longer timescale phenomena (folding, sometimes binding) • High computational demand
Physics-Based Forcefields	CHARMM19; GROMOS	• Well suited for all-atom models	• May have less accurate energy scoring than knowledge-based potentials
Knowledge-Based Statistical Forcefields	CABS forcefield [32]; DOPE [37]	• May be better suited for coarse-grained models than physics based	• Not transferable between different types of systems (e.g., single protein to intermolecular complex)
Docking	DOCK; AutoDock	• Able to simulate binding behavior too slow for MD	• Increased flexibility or size of ligand decreases accuracy
Deep Learning/Neural Networks	AlphaFold [24]	• Currently has the best structure prediction ability	• "Black box" effect might hinder true understanding of how proteins fold • May be biased against novel/under-represented protein folds • Needs development for predicting intermolecular interactions

Table 1.1: Summary of Components Used in Current Protein Modelling Techniques

Per Residue Interaction Energy & The Hidden Symmetry Model

While most protein modelling techniques utilize an energy function to calculate the total energy of the system (protein(s) and solvent), only a specialized few attempt to decompose a protein's interaction free energy on a per-residue basis. Those that do may rely on calculations obtained from computationally intensive methods like MD which may not fully represent a protein's behavior at a biologically relevant timescale [38], and/or they often attempt to attribute interaction energies to only a few, relatively short-range interactions (H-bonds, Van der Waals forces, etc.) [39], which may not take into account "hidden", non-local effects [40, 41]. An accurate, rapid, and easy to interpret residue energy breakdown could provide invaluable insights into which residues or types of interactions are critical for stabilizing a protein, binding to a ligand or protein interface, or responsible for inducing conformational changes. We have developed the Hidden Symmetry Model (HSyM) as a fast, user-friendly, and computationally inexpensive method to extract protein residue interaction energy contributions from sequence and structure data [42]. Unlike all-atom and coarse-grained Monte Carlo and MD based methods, this coarse-grained model avoids the computationally taxing task of molecular ensemble generation by only analyzing a protein's ground state, or lowest energy conformation; this approximation was based on the observation that protein folding—specifically, amino acid side chain burial and resulting solvent accessible surface area—exhibits scale-invariant behavior typical of phase transitions [43]. As a result, the model's algorithm is capable of generating an interaction energy map of a protein in a matter of seconds, and computation using our Python-based program (POLARIS) can be carried out on a standard PC.

A summary of the model's workflow is shown in **Figure 1.1**. In brief, the program uses sequence and structural (atomic coordinates) data from a user-inputted PDB file to assign every amino acid a unitless value, μ, directly related to that residue's interaction energy. This value is

calculated based on the identity of the amino acid, its secondary structure-related symmetry with neighboring residues, and the identities of these local neighbors. An in-depth discussion of the model's equations, parameters, and algorithm is provided in Chapter 2 of this dissertation.

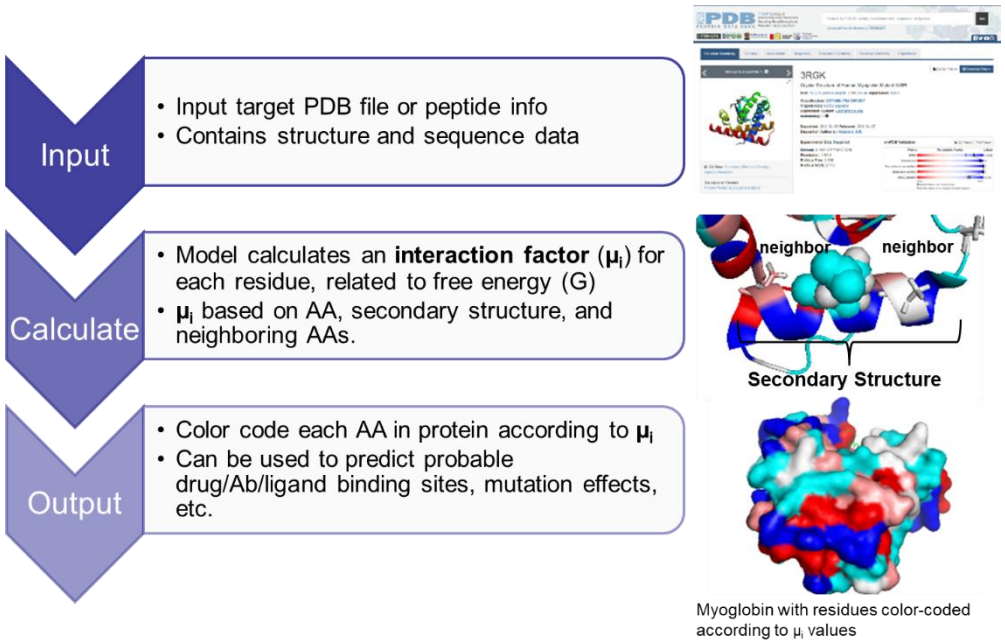


Figure 1.1: The Hidden Symmetry Model Workflow. HSyM is able to convert sequence and structure data from a PDB file to energy data by calculating an interaction factor (μ, unitless but directly proportional to a residue's interaction free energy contribution) for every residue in the protein. Myoglobin structure was obtained from PDB ID: 1MBN [44].

Potential Model Applications

HSyM's ability to rapidly provide a residue interaction energy map of a protein lends itself to a variety of potential applications. **Figure 1.2** categorizes and identifies several of these. For example, the identification of patches of higher-than-average μ residues on a protein's surface could mark these regions as "hot spots" that are likely candidates for drug or ligand binding sites and/or as active sites in enzymes. Identifying these target binding sites is a critical step in the design of small molecule and biologic drugs. Additionally, the model shows promise for

predicting thermostability changes associated with mutations, as demonstrated in Chapter 2. The engineering of more stable variants of proteins via targeted mutations has profound implications for designing biologics, vaccines, and enzyme catalysts that provide higher production yields and which maintain their native conformation and function at a broader range of temperatures and storage conditions [45]. Identifying key sites for mutation may also aid in the design of so-called “protein switches”, proteins that are able to assume a new stable conformation at specific conditions [46, 47].

A related protein engineering application is using HSyM to tune the binding affinity in protein-protein or protein-peptide interfaces, with the latter application being demonstrated in Chapter 3. By identifying and mutating selected flanking residues adjoining a binding site on one molecule, the model posits that the interaction energies of the conserved (non-mutated) residues involved in the binding can be changed, thereby altering the free energy of binding and thus the affinity of the intermolecular interaction. Besides being useful for the design of biologics like monoclonal antibodies, cytokines, and vaccines, this method of protein engineering could potentially be utilized in designing more sensitive diagnostic immunoassays or tissue engineering scaffolds that target certain cellular receptors/markers.

All in all, HSyM may be used as a potent in silico prescreening tool in protein engineering and drug design, allowing the user to rapidly and easily rank and eliminate candidates based on quantities like thermostability or binding affinity before proceeding to either more computationally intensive modeling (atomistic Monte Carlo methods or MD simulations) or experimental validation/screening (phage display, microarrays, functional assays, etc.).

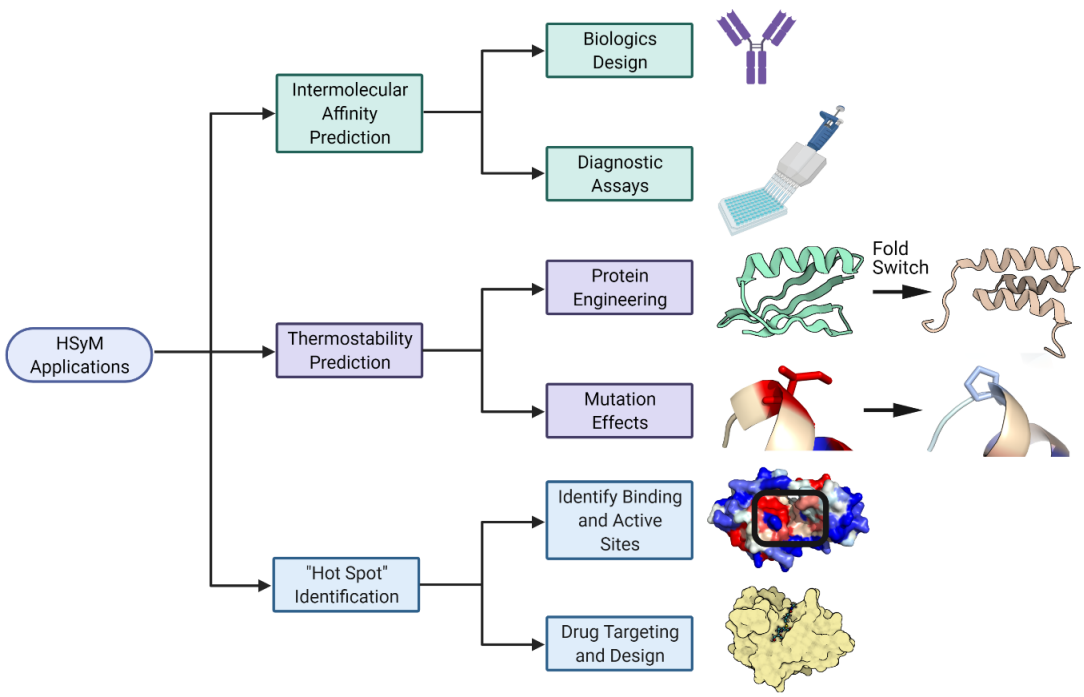


Figure 1.2: Potential Applications of the Hidden Symmetry Model. The illustrations for the "Drug Targeting and Design", "Identifying Binding and Active Sites"/"Mutation Effects", and "Protein Engineering" applications were created using PDB entries 4EYL, 3FA0, 2KDM, and 2KDL, respectively [46, 48, 49]. This figure was created using BioRender.

Thesis Objective and Summary

The dearth of user-friendly, computationally simple protein models that provide a decomposition of residue interaction energy presents a hindrance to a more complete understanding of protein behavior and to protein engineering efforts. We have developed HSyM to address this problem, with the central objectives of (1) developing a model that accurately gives a per-residue interaction energy decomposition of a protein based only on ground state sequence and structure data and (2) demonstrating possible applications of this model for medicine and biotechnology.

Chapter 2 of this thesis describes HSyM in detail and benchmarks it by successfully predicting mutation-induced thermostability changes with a study of the protein T4 lysozyme. Chapter 3 describes a peptide engineering study in which HSyM was used to design a suite of peptides and was successfully able to predict their relative binding affinities for an antibody, even without any structural data for the peptides or antibody target. This chapter discusses a possible use for this in designing diagnostic immunoassays for cancer autoantibodies with improved sensitivities. Chapter 4 is focused on further optimizing the model and addressing some of its flaws, primarily through introducing parameters that accurately account for residues' interactions with water. A statistical mechanical analysis of a simplified two dimensional "toy model" that simulates protein folding/unfolding and residue-solvent interactions was used to arrive at these parameter values. Chapter 5 presents preliminary work on a fully-integrated microfluidic device that we envision could carry out automated blood/sera immunoassays using peptides designed with the method in Chapter 3. Finally, Chapter 6 summarizes our work to date, discusses the remaining limitations of the model, and explores proposed future work.

References

1. Tanford, C.; Reynolds, J., *Nature's robots: a history of proteins*. OUP Oxford: 2003.
2. Lee, M. J.; Yaffe, M. B., Protein Regulation in Signal Transduction. *Cold Spring Harbor Perspectives in Biology* **2016,** *8* (6).
3. Ahmed, K. A.; Xiang, J., Mechanisms of cellular communication through intercellular protein transfer. *J Cell Mol Med* **2011,** *15* (7), 1458-1473.
4. Kurtz, J., Memory in the innate and adaptive immune systems. *Microbes and Infection* **2004,** *6* (15), 1410-1417.
5. Latchman, D., *Gene regulation*. Taylor & Francis: 2007.
6. Robinson, P. K., Enzymes: principles and biotechnological applications. *Essays Biochem* **2015,** *59*, 1-41.
7. James, L. C.; Tawfik, D. S., Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences* **2003,** *28* (7), 361-368.
8. Letovsky, S.; Kasif, S., Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **2003,** *19* (suppl_1), i197-i204.
9. Hegyi, H.; Gerstein, M., The relationship between protein structure and function: a comprehensive survey with application to the yeast genome 11Edited by G. von Heijne. *Journal of Molecular Biology* **1999,** *288* (1), 147-164.

10. Scott, D. E.; Bayly, A. R.; Abell, C.; Skidmore, J., Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery* **2016,** *15* (8), 533-550.
11. Śledź, P.; Caflisch, A., Protein structure-based drug design: from docking to molecular dynamics. *Current opinion in structural biology* **2018,** *48*, 93-102.
12. Engqvist, M. K. M.; Rabe, K. S., Applications of Protein Engineering and Directed Evolution in Plant Research. *Plant Physiol* **2019,** *179* (3), 907-917.
13. Wen, F.; Nair, N. U.; Zhao, H., Protein engineering in designing tailored enzymes and microorganisms for biofuels production. *Curr Opin Biotechnol* **2009,** *20* (4), 412-419.
14. Kapoor, S.; Rafiq, A.; Sharma, S., Protein engineering and its applications in food industry. *Critical Reviews in Food Science and Nutrition* **2017,** *57* (11), 2321-2329.
15. Mackenzie, C. O.; Grigoryan, G., Protein structural motifs in prediction and design. *Current opinion in structural biology* **2017,** *44*, 161-167.
16. Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C., Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* **2011,** *6* (12), e28766.
17. Jones, D. T.; Kandathil, S. M., High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **2018,** *34* (19), 3308-3315.
18. Kuhlman, B.; Bradley, P., Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **2019,** *20* (11), 681-697.
19. Vanommeslaeghe, K.; MacKerell Jr, A., CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta (BBA)-General Subjects* **2015,** *1850* (5), 861-871.
20. Cossio, P.; Granata, D.; Laio, A.; Seno, F.; Trovato, A., A simple and efficient statistical potential for scoring ensembles of protein structures. *Scientific Reports* **2012,** *2* (1), 351.
21. Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A., Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **2016,** *116* (14), 7898-7936.
22. Das, R.; Baker, D., Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **2008,** *77*, 363-382.
23. Vlachakis, D.; Bencurova, E.; Papangelopoulos, N.; Kossida, S., Current state-of-the-art molecular dynamics methods and applications. *Advances in protein chemistry and structural biology* **2014,** *94*, 269-313.
24. Service, R. F., 'The game has changed.' AI triumphs at protein folding. *Science* **2020,** *370* (6521), 1144.
25. Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; Baker, D., Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **2013,** *501* (7466), 212-216.
26. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453* (7192), 190-195.
27. Murphy, G. S.; Sathyamoorthy, B.; Der, B. S.; Machius, M. C.; Pulavarti, S. V.; Szyperski, T.; Kuhlman, B., Computational de novo design of a four-helix bundle protein--DND_4HB. *Protein science : a publication of the Protein Society* **2015,** *24* (4), 434-45.
28. Lin, Y.-R.; Koga, N.; Tatsumi-Koga, R.; Liu, G.; Clouser, A. F.; Montelione, G. T.; Baker, D., Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences* **2015,** *112* (40), E5478.

29. Śledź, P.; Caflisch, A., Protein structure-based drug design: from docking to molecular dynamics. *Curr Opin Struct Biol* **2018,** *48*, 93-102.
30. Wang, C.; Schueler-Furman, O.; Baker, D., Improved side-chain modeling for protein–protein docking. *Protein Science* **2005,** *14* (5), 1328-1339.
31. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **2011,** *487*, 545-574.
32. Kolinski, A., Protein modeling and structure prediction with a reduced representation. *Acta biochimica Polonica* **2004,** *51* (2), 349-71.
33. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **2018,** *46* (W1), W296-w303.
34. Webb, B.; Sali, A., Comparative Protein Structure Modeling Using MODELLER. *Current protocols in bioinformatics* **2016,** *54*, 5.6.1-5.6.37.
35. Bowers, K. J.; Chow, D. E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. In *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*, SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, 11-17 Nov. 2006; 2006; pp 43-43.
36. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005,** *26* (16), 1701-1718.
37. Shen, M. y.; Sali, A., Statistical potential for assessment and prediction of protein structures. *Protein science* **2006,** *15* (11), 2507-2524.
38. Serçinoğlu, O.; Ozbek, P., gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. *Nucleic Acids Research* **2018,** *46* (W1), W554-W562.
39. Melagraki, G.; Ntougkos, E.; Rinotas, V.; Papaneophytou, C.; Leonis, G.; Mavromoustakos, T.; Kontopidis, G.; Douni, E.; Afantitis, A.; Kollias, G., Cheminformatics-aided discovery of small-molecule Protein-Protein Interaction (PPI) dual inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF-κB Ligand (RANKL). *PLOS Computational Biology* **2017,** *13* (4), e1005372.
40. Mark, A. E.; van Gunsteren, W. F., Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J Mol Biol* **1994,** *240* (2), 167-76.
41. Dill, K. A., Additivity principles in biochemistry. *The Journal of biological chemistry* **1997,** *272* (2), 701-4.
42. Williams, L. J.; Schendt, B. J.; Fritz, Z. R.; Attali, Y.; Lavroff, R. H.; Yarmush, M. L., A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme. *Technology (Singap World Sci)* **2019,** *7* (1-2), 12-39.
43. Moret, M. A.; Zebende, G. F., Amino acid hydrophobicity and accessible surface area. *Physical Review E* **2007,** *75* (1), 011920.
44. Watson, H., The stereochemistry of the protein myoglobin. *Prog. Stereochem* **1969,** *4* (299), 5.
45. Hsieh, C.-L.; Goldsmith, J. A.; Schaub, J. M.; DiVenere, A. M.; Kuo, H.-C.; Javanmardi, K.; Le, K. C.; Wrapp, D.; Lee, A. G.; Liu, Y.; Chou, C.-W.; Byrne, P. O.; Hjorth, C. K.; Johnson, N. V.;

Ludes-Meyers, J.; Nguyen, A. W.; Park, J.; Wang, N.; Amengor, D.; Lavinder, J. J.; Ippolito, G. C.; Maynard, J. A.; Finkelstein, I. J.; McLellan, J. S., Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **2020,** *369* (6510), 1501.
46. Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N., A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (50), 21149-54.
47. Wei, K. Y.; Moschidi, D.; Bick, M. J.; Nerli, S.; McShan, A. C.; Carter, L. P.; Huang, P.-S.; Fletcher, D. A.; Sgourakis, N. G.; Boyken, S. E.; Baker, D., Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proceedings of the National Academy of Sciences* **2020,** *117* (13), 7208.
48. King, D. T.; Worrall, L. J.; Gruninger, R.; Strynadka, N. C., New Delhi metallo-β-lactamase: structural insights into β-lactam recognition and inhibition. *Journal of the American Chemical Society* **2012,** *134* (28), 11362-5.
49. Mooers, B. H.; Tronrud, D. E.; Matthews, B. W., Evaluation at atomic resolution of the role of strain in destabilizing the temperature-sensitive T4 lysozyme mutant Arg 96 --> His. *Protein science : a publication of the Protein Society* **2009,** *18* (5), 863-70.

Chapter 2: A Protein Interaction Free Energy Model Based on Amino Acid Residue Contributions: Assessment of Point-Mutation Stability of T4 Lysozyme

Note: This chapter is reproduced from the following publication:

A Protein Interaction Free Energy Model Based on Amino Acid Residue Contributions: Assessment of Point-Mutation Stability of T4 Lysozyme, *Technology (Singap World Sci)* 2019, 7 (1-2), 12-39. doi:10.1142/s233954781950002x

Lawrence J. Williams[1], Brian J. Schendt[1], Zachary R. Fritz[2], Yonatan Attali[1], Robert H. Lavroff[1], Martin L. Yarmush[2]

[1]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey

[2]Department of Biomedical Engineering, Rutgers, The State University of New Jersey

Correspondence should be addressed to L.J.W. (lwilliams@chem.rutgers.edu) or M.L.Y. (yarmush@soe.rutgers.edu).

Contribution

Working collaboratively with the other authors, I (Zachary Fritz) was involved with every aspect of this work, including writing and editing code, conducting calculations, analyzing data, creating all of the figures, and writing the manuscript.

Abstract

Here we present a model to estimate the contribution of each amino acid residue to the interaction free energy of a given protein. Protein interaction energy is described in terms of per-residue interaction factors, μ. Multibody interactions are implicitly captured in μ through the combination of amino acid terms (γ) guided by local conformation indices (σ). The model enables construction of an interaction factor heat map for a protein in a given fold, prima facie assessment of the degree of residue-residue interaction, and facilitates a qualitative and quantitative evaluation of protein association properties. The model was used to compute thermal stability of T4 bacteriophage lysozyme mutants across 7 sites. Qualitative assessment of mutational effects provides a straightforward rationale as to whether a particular site primarily perturbs native or non-native states, or both. The presented model was found to be in good agreement with experimental mutational data (R^2 = 0.73) and suggests a means by which to convert structure space into energy space.

Significance

The assessment of protein thermal stability as a function of amino acid sequence enables the testing of molecular models. Whereas most quantitative studies rely on all-

atom descriptions and dynamic simulations across many time scales to assess free energy changes, we introduce a model that uses only coarse structural data: protein sequence and a list of side chain-side chain contacts. The model uses amino acid-specific fractional exponents to describe the protein sequence and assigns each residue to one of four states. Despite ignoring all molecular fine structure – including the details of backbone and side chain conformation and all specific interactions – we find that changes in folding free energy can be estimated with good accuracy and that protein energetics can be interpreted and visualized in terms of per-residue contributions. The key implication of the study is that structure space may be readily convertible into energy space.

Introduction

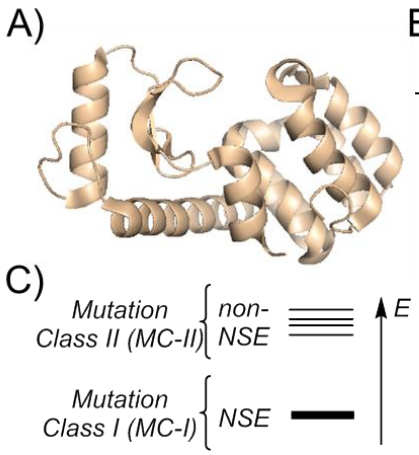
Biological processes are fundamentally realized at the atomic and molecular level and are mostly carried out by proteins. Molecular life science is teeming with protein studies and derived biotechnologies. Despite its tremendous success, biotechnology and its foundational molecular sciences lack a simple means to relate interaction energy to protein primary sequence, structure, and function.[1, 2] Advancement of such an understanding would significantly improve the pace and capabilities of the biotechnology industry.[3] Without it the community has grown to rely on intelligent guesswork made possible by genomics, proteomics, informatics, and other experimental tools for protein manipulation and structure determination.[4] The seminal insights of the founders of protein science[5-10] largely have given way to

methods of protein characterization that depend on all-atom computations.[11-16] These powerful tools have not led to much-desired qualitative guidelines, intuitive understanding, or prima facie approximations of per-residue contributions to interaction free energy.[17-20] It is an open question whether the difficulty in identifying governing principles of protein behavior is intrinsic to proteins or whether the difficulty is due to obfuscation by the surfeit of details generated by all-atom computations.

We began to explore the possibility of a simple model to understand protein energetics derived solely from protein sequence and approximate structure.[21, 22] A scale-invariant, non-atomistic approach could provide a straightforward understanding of proteins and enable prima facie analysis of sequence-dependent behavior.[23-27] Moreover, it could substantively impact the biotechnology industry by simplifying the description of intra- and inter-molecular protein interactions, facilitating protein design and engineering, and describing a direct connection between genomic sequence, protein structure, and biological function. At present, proteins are engineered to have enhanced binding affinity, thermostability, catalytic activity, etc., through synthetic and recombinant technologies as well as computational modeling.[28, 29] However, this process can be labor intensive and largely empirical without the guidance provided by a predictive understanding of interaction free energy.

To overcome these limitations, we introduce a simple model to predict the per-residue protein interaction energy and demonstrate its applicability to understand thermal stability of the well-studied T4 bacteriophage lysozyme protein (**T4L**, **Fig 2.1A**).

We illustrate the ideas by focusing on single point mutations of the wild-type protein, the use of a fractional exponent scale (**Fig. 2.1B**), and classification of mutations that are expected to primarily impact the energetics of the native state ensemble (NSE, **Fig. 2.1C**). Our findings suggest that this model has the potential to: (a) understand the effects of mutations on protein thermal stability, (b) approximate well the interaction free energy changes that bear on the native state ensemble, and (c) convert protein structure space into energy space.



B)

AA	~γ x10^3 (sdev)	AA	~γ x10^3 (sdev)
C	246 (4)	H	153 (2)
V	238 (4)	T	135 (1)
I	222 (2)	P	121 (2)
Y	222 (2)	N	113 (1)
M	221 (5)	Q	105 (2)
F	218 (3)	S	101 (1)
L	197 (3)	E	94 (1)
W	174 (3)	D	87 (2)
A	157 (6)	R	78 (3)
G	156 (5)	K	69 (1)

Figure 2.1. Towards a reduced complexity description of proteins. **(A)** Lysozyme of T4 bacteriophage (T4L, PDB ID: 3fa0). **(B)** Scale invariant fractional values for the 20 canonical amino acids and with the standard deviation (sdev) determined from the power law: $\alpha_r \propto N^{-\gamma}$ (see: Ref. 31, text, and Appendix). **(C)** Mutations are classified based on whether the impact of mutation is expected to be primarily on the native state ensemble (NSE) – Mutation class I (MC-I) or on the non-native state ensemble (non-NSE) or both the NSE and the non-NSE – Mutation class II (MC-II).

Summary of the Model

Our model has three key features: (1) a high precision (fractal exponent) parameter for each amino acid residue (γ), (2) an index of four possible states that each residue can adopt (σ), and (3) normalization of a unique subset of residues for each amino acid in the sequence to determine their interaction factor (μ). These are briefly described below (see also **S2.1**).

γ-Parameter. There are over 400 scales describing attempts to parameterize various properties of amino acids residues in proteins, such as structural propensities, hydrophobicity, etc.[30] However, there is only one scale-invariant set (**Fig. 2.1B**) and it is based on accessible surface area measurements. [24, 25, 31] Using high resolution crystallographic data, peptide segments as small as only three amino acid residues in length were computationally separated from the parent protein as static objects and the accessible area of the central residue was measured. The procedure was repeated for larger and larger segments, each time measuring the accessible surface area of the central residue. Examination of thousands of segments revealed that each amino acid displayed scale invariance and obeyed a power law of the form $\alpha_r = N^{-\gamma}$, where α_r is the relative accessible surface area of the amino acid side chain and N is the length of the peptide segment analyzed (see also, **S2.1**). Although asymptotic reduction of amino acid sidechain accessible surface area was noted early on[32, 33] and anticipated by steric crowding, it was not until the report of the γ scale[31] that this behavior was shown to be distinctly different for each of the 20 amino acids. The exponents are fractional and range from approximately 1/16 (for lysine) to 1/4 (for cysteine) (**Fig 2.1B**). The possible implications of this previously hidden dilation symmetry led us to formulate a model, which we term the Hidden Symmetry Model (HSyM), based on the information likely to be contained in these exponents and based on notions of how the effects captured by these terms might be propagated along the protein sequence (for the details of these conjectures, see **Appendix**).

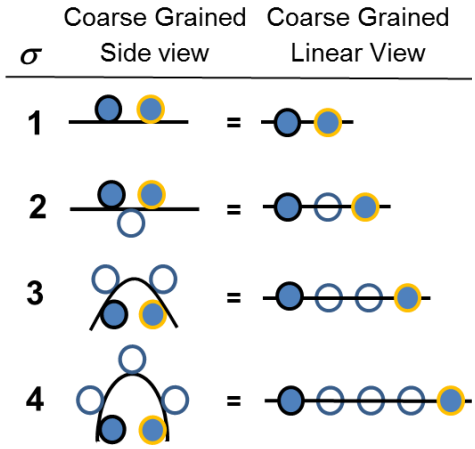


Figure 2.2. σ-index classification of local protein conformation. Amino acid sidechains are coarsened as blobs (circles). The degree of separation between a residue (blue-filled black circle) and the closest-linked nearest neighbor (CLNN, blue-filled orange circle) defines σ. Simplified schematic of each closest-linked nearest neighbor index is shown for σ = **1**, **2**, **3**, and **4** (two views).

σ-Parameter. Although protein conformation is conventionally defined by the backbone and side chain torsion angles,[34-36] we anticipated advantages of an alternative conformational classification, which we term σ-index. The σ-index is a structural classification based on the closest-linked nearest neighbors (CLNN) of each amino acid residue. Briefly, CLNN are those residues with sidechains near each other and separated by less than 4 residues in the sequence (see **S2.1** for a detailed description of CLNNs and sigma assignment). The CLNN designation can be σ = 1, 2, 3, or 4 (**Fig. 2**). This index describes canonical secondary motifs in recognizable terms, but also describes loops, turns, and coil structures. For example, a residue (*i*) that is part of a canonical beta strand would be assigned σ = 2, since *i*+2 and/or *i*-2 would be CLNN. Had that region of the protein been folded as an extended 3_{10} helix the residue (*i*) would be assigned σ = 3, since *i*+3 and/or *i*-3 would be the CLNN. The CLNN approach thereby delimits

local conformation according to relative amino acid side chain arrangement. This coarse treatment of conformation considers many variations of the local backbone and side chain dihedral angles as equivalent and accommodates small dynamical fluctuations.[37, 38]

μ-Parameter. Effective per-residue contributions to interaction energy (μ) are determined as a function of *γ* and *σ* terms (**Eq. 1**). Normalization of intrinsic contribution (*γ*) of the *determining set* (θ_d^i) gives μ_i. The set, θ_d^i, consists of residue *i*, the CLNN (*i* + *σ*), along with a subset of residues nearby in the sequence that would constitute an extended regular *σ* motif – a surface composed of contiguous amino acid residue sidechains centered on residue *i*. A physical interpretation of this is as a hydrated surface centered on the residue of interest, for example, the sort of extended motif that would be revealed upon partial unfolding.[19, 39, 40] Each residue of this surface is considered critical and is therefore included in determining the interaction factor. Specifically, $i \pm n\sigma \in \theta_d^i$ where n = 1, 2, 3, …, and $n\sigma \leq 10$. The details of interaction factor determination and the related free energy calculations are given in **Appendix**. Briefly, the free energy of interacting residues *i* and *j* is related to the product of the interaction factors, μ, **Eq. 2**, where *G* is free energy, ε relates medium effects, $\tau_{i,j}$ is an interaction efficiency term, and λ is a scaling factor with units of energy. Single state analysis requires an estimate of entropic contributions (see **Appendix**). For the purposes of this study we set $\varepsilon = \tau_{i,j} = 1$ and we estimate λ in this study through fitting the **T4L** thermal data (see below). The implicit multi-body interactions subsumed in the interaction factors (μ) aim to reflect solvation and other non-local effects that are hidden or obscured by multi-scale correlations. The scaling problem confounds protein free energy calculations, especially per-residue free energy contributions,[17, 41, 42] but as shown below, this approach appears to describe interaction free energy successfully.

$$\mu_i^\sigma = (2\xi_\sigma + 1)^{-1} \sum_{n=-\xi_\sigma}^{\xi_\sigma} \gamma_{i+\sigma n} \ , \ \xi_\sigma = \left\lfloor \frac{\xi}{\sigma} \right\rfloor \quad \textbf{Eq. 1}$$

$$\Delta G \sim -\varepsilon\lambda \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\,\mu_i\mu_j \qquad \textbf{Eq. 2}$$

Results and Discussion

Native vs Non-native Ensemble Perturbations. The energy gap between the native and non-native ensembles determines the thermal stability of a protein fold.[43, 44] A decrease in thermal stability of a mutant protein reflects destabilization of the native state, stabilization of non-native states, or both; the converse holds as well. Experimental measures of thermal stability ($\Delta\Delta G_{exp}$) do not reveal which state or states are perturbed by mutation. Theoretical appraisal of thermal stability does not require knowledge of which states are perturbed provided the energetics of a suitably representative set of the very large number of conformational microstates are computed accurately across appropriate timescales.[45-47] This is difficult to achieve with certainty. The determination of mutational effects based on analysis of the native state alone is particularly challenging, since knowledge of whether the mutation primarily influences the native state is required.

Our model enables speculation of whether a mutation primarily influences native state stability and by how much ($\Delta\Delta G_{calc}$) based on evaluation of the native state alone. First, we categorize the impact of mutation on thermal stability as belonging to one of two classes (**Fig 2.1C**). Mutation Class I (MC-I) are those mutants that are expected to primarily change the interaction factors of buried residues. Such changes are expected to correlate with protein stability, since these changes bear directly on stabilizing the native state. This presumes that the non-native state involves different residue-residue contacts and/or different backbone conformations.[48, 49] MC-I mutations are expected to minimally perturb the non-native state energetics. Mutation Class II (MC- II) are those mutants that change the interaction factors of surface residues or both surface and buried residues and therefore are expected to perturb

either the non-native states or both native and non-native states. This analysis suggests that the determination of whether a mutation is Class I can be assessed qualitatively and quantitatively by examining the ground state structure alone.

The set of residues whose μ factor would change upon mutation of site *i* determines the effect of the mutation on thermal stability. We term this group of residues the *member set* of site *i* (θ_m^i). It is important to note that θ_m^i and θ_d^i are not necessarily identical. The member set is composed of those residues whose μ values are impacted by mutation at site *i*, whereas the determining set is composed of those residues that are normalized to give the μ value of site *i*. Analysis of θ_m^i can be used to assess whether mutation of residue *i* will be of Class I, as illustrated below.

μ-Profile of T4 Lysozyme. T4L is one of the most extensively studied proteins.[50-53] The availability of high-resolution structures for the wild-type protein and many of its mutant forms, combined with the corresponding thermal data, make this protein an ideal candidate for comparative analysis.[54] We focused on these structures since of the single point mutations of this protein do not significantly alter the native structure. Each possible μ_i factor for the wild-type sequence was calculated

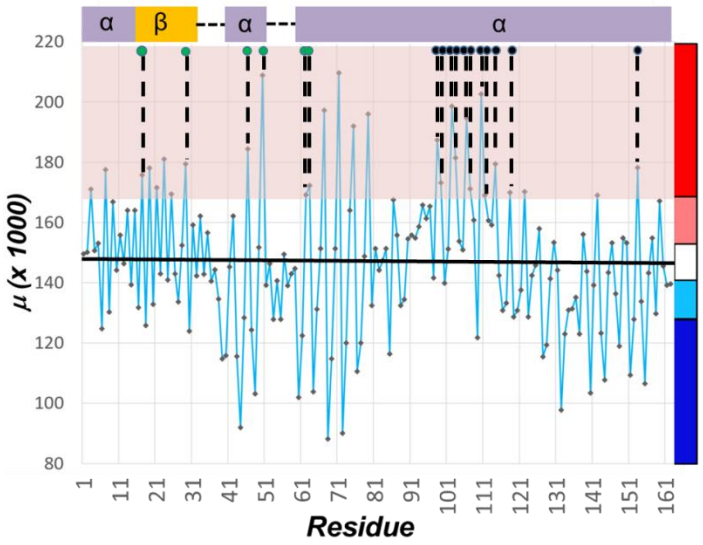


Figure 2.3. Computed μ-values of T4L. Per-residue values calculated according to **Eq. 1** (blue lines in plot provided for clarity only). Color gradient (right) used to generate heat maps in **Figure 2.4**, where red corresponds to the highest values (hot) and blue to the lowest (cold). Most of the hottest thirty-one residues make contiguous contacts and twenty-one aggregate in three clusters. Two of the largest are indicated (dotted lines). The first cluster (green) includes six of the hottest residues in the *N*-terminal domain. The second (black) includes eleven residues and is located in the *C*-terminal domain. The values and specific residue color used to determine the per-residue contributions and for visualization are given in **Table S2.4**.

(**Table S2.1**). We used the high resolution structure of this protein (PDB ID: 3FA0[55]) to assign a σ-index to each residue and the residue-residue contacts (see **S2.1** for details). Additionally, all mutants were assumed to retain these assignments. **Figure 2.3** is a plot of the computed per-residue μ-factors for the native fold, along with the conventional α-helical and β-sheet motif assignments (top) and a color gradient (far right) to categorize and visualize relative μ values of the protein. These range from red (high/hot), to salmon, white, cyan, and then blue (low/cold). The protein structure heat-map, shown in **Figure 2.4A** and 2.**4B**, provides a simple way to

simultaneously visualize the approximate relative magnitude of interaction factors and the protein structure. For example, the protein exterior is dominated by cold residues. The interior, however, is dominated by hot residues, whose interactions significantly contribute to stability ($G \sim \mu_i\mu_j$). **Figure S2.1** shows the heat maps of several other classic globular proteins, which also follow this trend. This model suggests that proteins may be organized such that the interaction factors are optimized. The canonical hydrophobic amino acid residues,[56] i.e. residues with non-polar side chains, tend to be hot, whereas those with polar side chains, especially those with ionic functionality, tend to be cold. Approximately 25% of the residues in each of these classes, however, do not hold to these trends. Thus, the nature of the side-chain without regard for backbone conformation (σ-index) and sequence context is not a reliable predictor of μ factors. For example, in T4 lysozyme, some of the archetypal nonpolar residues are cold, such as L121, and some of the charged residues are hot, such as the catalytically relevant E11 (**Figs. 2.4E**, **F**).

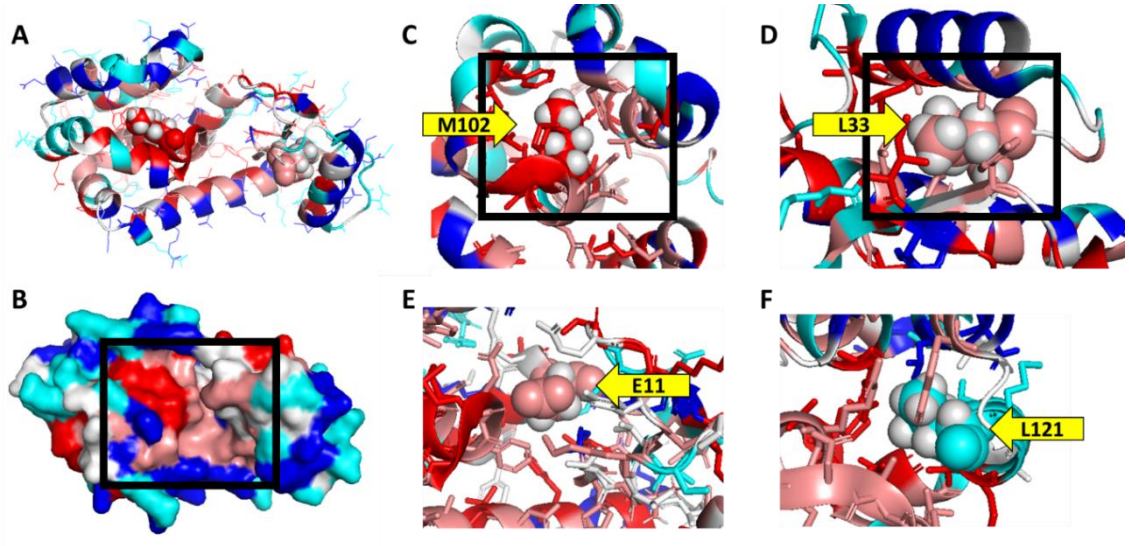


Figure 2.4. Heat maps of T4 lysozyme. (**A**) The protein interior is predominantly hot (red and salmon interstitial lines) and the exterior predominantly cold (blue and cyan surface lines). Residues M102 (left) and L33 (right) are shown in space filling mode for reference (both residues

are hot, hydrogen atoms are shown light grey for clarity). As indicated in **Figure 2.3**, red/salmon correspond to high μ-factors, whereas white indicates midrange values and cyan/blue indicate low range values. (**B**) The patch of hot surface residues corresponds to the substrate binding region (box). (**C**) The *C*-terminal core region (box) is dominated by hot residues; M102 is in contact with many hot core residues. (**D**) The *N*-terminal core region (box), though smaller, is also dominated by hot residues; L33 is in contact with many hot core residues. (**E**) Approximately 25% of the canonical hydrophilic residues are hot, including the key catalytic residue E11. (**F**) Approximately 25% of the canonical hydrophobic residues are cool, including L121, which is buried and makes many contacts with hot residues of the *C*-terminal core. Mutations that increase the μ-factor for L121 (e.g., S117V, as explained later in the text) would be expected to significantly increase the stability of the C-terminal core and increase the protein stability.

Hot Residue Cluster Analysis. The 31 residues of **T4L** with the highest μ factors, i.e., the 'hottest' residues, are indicated in the top shaded region of **Figure 2.3**. The two largest hot clusters constitute the C- and N-terminal domain cores (**Fig. 2.4C and D**, respectively). The N-terminal domain, which is known to be the less stable of the two, has a smaller core, and has significantly fewer hot residues in the core than the C-terminal domain. An uninterrupted vein of hot residues connects the two domains. All of the 31 hottest residues (**Table S2.2**), except for I9 and D159, make extensive contacts with other hot residues, and 21 of the 31 hottest residues co-localize in three clusters. Although the clustering leads to the protein interior being dominated by hot residues and the exterior being dominated by cold residues, there are some cool residues that are buried and some hot residues that are on the protein surface. For example, the most noticeable hot spot on the protein surface is the swath of contiguous hot

residues that constitute the substrate binding region (**Figure 2.4B)**.[57, 58]

Heat-map Analysis. Many multi-mutant variants of **T4L** have been reported. In this study, we focused on thirteen positions of wild-type enzyme that have been the subject of single point mutations (see Supporting Information **Section S2.1)**. Virtually all of these mutants have been characterized both in terms of thermal stability and structure.[54] We examined these mutants in detail. **Table S2.3** gives the θ_d^i and θ_m^i sets for each residue of this protein (see also **Table S2.4**). Based on the location of each residue of the ϑ_m sets, seven of the sites (positions 3, 11, 115, 117, 119, 132, 133) were assessed as MC-I mutations and therefore expected to contribute primarily to conformational ground state stabilization. For example, residue S117 rests near the center of an α-helix (σ = 4). As shown in **Figure 2.5A**, site 117 is

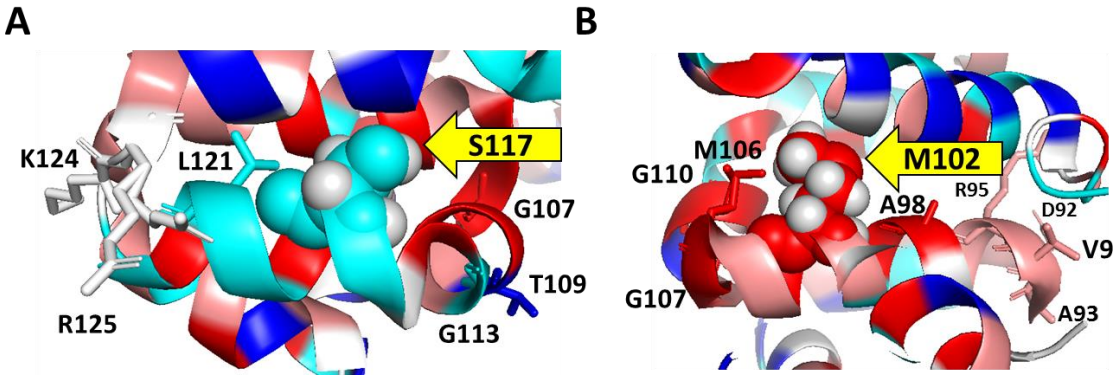


Figure 2.5. Mutation impacts the contribution to interaction free energy of multiple residues. (**A**) Mutation of S117 (spheres) changes the μ-factors calculated for θ_m–related residues (sticks). Since serine has a low intrinsic value (γ, **Fig. 2.1B**), mutation to most other residues will increase these □-factors. This set includes hot and cold residues that make many contacts with the hot residues of the C-terminal core and will therefore significantly stabilize the protein upon mutation of S in most cases. (**B**) Mutation of M102 (spheres) in most cases will decrease the μ-factors of the θ_m–related residues (sticks). This set of residues, which are hot core residues, will therefore cool down and significantly destabilize the protein.

a member of the determining set of sites 107(1), 109(4), 113(4), 117(4), 121(4), 124(1), and 125(1) (i.e., θ_m^{117}, the number in parentheses indicates the σ-index of the residue listed). Consequently, mutation at 117 will impact the interaction factors of each member of this set in this fold. Many of the residues of θ_m^{117}, specifically 107, 117, and 121, are part of the hot C-terminal core cluster residues or contact these core residues. Therefore, mutation at 117 should bear primarily on ground state stability. The S117V mutant is the most stabilizing single point mutant known for **T4L** (ignoring covalent crosslinks). Interestingly, the thermal stability of this mutant has been difficult to rationalize.[53, 54] Within the framework of this model, the scale invariant terms (γ) increase from serine to alanine to valine (**Fig. 2.1B**). Indeed, the interaction factors for the entire θ_m^{117} set will increase for S117A and for S117V. The thermal stability is expected to increase because the majority of residues are buried and interact with (hot) core residues. The S117V mutant would be expected to increase in stability to an even greater extent over S117A. Similarly, residue M102 is buried, and θ_m^{102} includes some of the hottest C-terminal core residues in one of the largest clusters of hot residues in this protein (92(1), 93(1), 94(4), 95(1), 98(4), 102(4), 106(4), 107(1), 110(4) $\in \theta_m^{102}$, **Fig. 2.5B**). Hence, mutation at this site is expected to bear primarily on ground state stability, as well. However, relative to methionine most mutations at position 102 would be expected to compromise stability (**Fig. 2.1B**). In this way, we used this model first to qualitatively assess the impact of the 13 point mutation sites, as MC-I or MC-II, and then to quantitatively assess the expected change in folding free energy for the 28 mutants distributed across these seven MC-I sites, as described below.

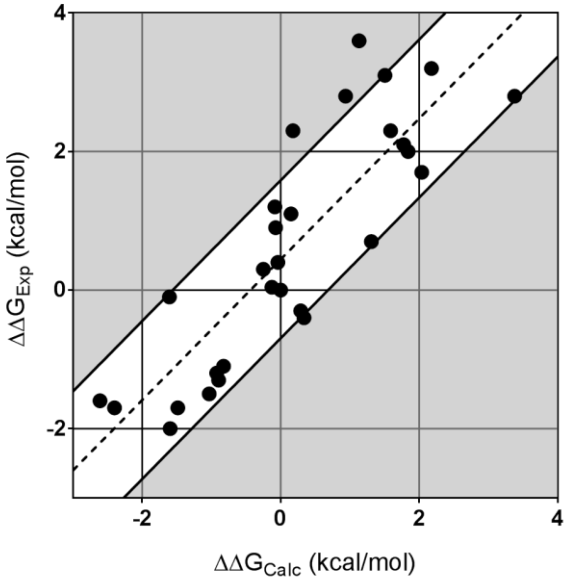


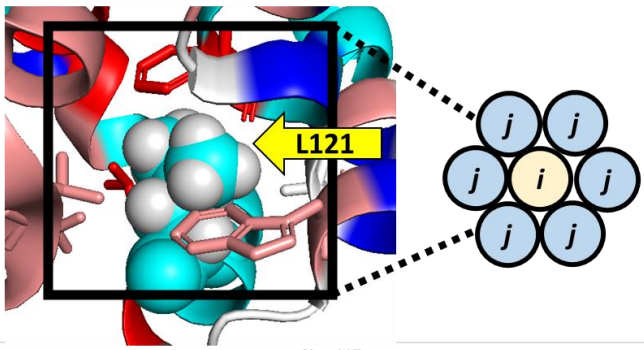
Figure 2.6*.* Comparison of computed and experimental *ΔΔG* of T4L mutants. Differences in thermal stability of well-characterized single point mutants were compared to experimentally determined values (Ref. 52). The calculation is remarkably accurate, since the only input is sequence and a list of side chain-side chain nearest neighbors. $R^2 = 0.73$ (line); y-intercept = 0.44, average uncertainty in the sums and products of $\mu < 0.01$ kcal/mol; average uncertainty attributable to estimated sidechain and backbone entropy < 0.7 kcal/mol (Ref. 58); average unassigned error (AUE) = 0.81 kcal/mol (white band); fit with Eq. 3, $\varepsilon = \tau = 1$, $\lambda = 12.5$ *kcal/mol*.

Quantitative Assessment of Mutant Thermal Stability. The change in folding free energy was calculated and compared to experimental data for the MC-I point mutants of **T4L** (**Figure 2.6**, see also **Table S2.5**). Briefly, free energy contributions are approximated by summing the products of interaction factors of residue (*i*) and its nearest neighbors (*j*) for the entire member set (θ_m) of each mutation site (i.e., $\Sigma\Sigma\mu_i\mu_j$) for both the wild-type and mutant proteins as well as changes in the backbone and sidechain entropy at the site of mutation (see

Appendix, **Eq. 3**, and **S1** for details). It is noteworthy that the μ factors alone provide a reasonable, though less accurate, prediction of $\Delta\Delta G_{calc}$ (see **Figure S2.2**). Although changes in local structure or among contacts between residues, etc., could also be assessed, for simplicity we approximate the structure of the mutants as perturbations of the wild-type protein with the same residue-residue contacts. The similarity of the crystal structures supports this approximation. **Figure 2.7** illustrates the key aspects of the analysis with S117V. The S117V mutation significantly modulates the entire θ_m^{117} (**Fig. 2.5C**). For example, L121 is one of the residues impacted by this mutation. The six nearest neighbors of the L121 side chain are V87, L91, L118, W126, A129, and F153. All but A129 are calculated to have high μ-factors, i.e., to be hot residues (**Fig. 2.7A**). In the wild-type protein L121 is cold, but the S117V mutant converts L121 to a hot residue and thereby increases the estimated stability of the protein. Taken together, the S117V mutant is estimated to significantly increase protein stability – in agreement with experiment.

A contact schematic of nearest neighbors for L121 is illustrated in **Fig. 2.7A** (see inset). The central residue corresponds to *i* and the peripheral residues correspond to its nearest neighbors *j* (c.f. **Eq. 2**). **Table S2.6** shows the entire *i,j* array for the 13 mutation sites. We use the backbone and side chain confinement penalties described by Baxa, et al.,[59] to estimate these entropy contributions to free energy (**Eq. 3**, see **S2.1** for details). These data, the computed μ factors, and the scaling factor λ (λ = 12.5 kcal/mol, **Eq. 3**) gave $\Delta\Delta G_{\text{calc}}$, which is in good agreement with experimental results (R^2 = 0.73, *AUE* = 0.81 kcal/mol, **Fig. 2.6**).

A



B

	Site 117						
$\theta_m(i)$	107	109	113	117	121	124	125
Contacts (*j*)	108	108	114	114	87	126	120
	110	107	W	132	91	90	123
	106	110	W	133	126	91	119
	109	W	W	118	129	92	128
	W	W	W	121	153	W	W
	W	W	W	129	118	W	W

Figure 2.7. Contacts (*i,j* pairs) for members of the θ_m set of residue 117. **(A)** Interaction energy of each residue is determined by summing the interaction factor products of each residue (*i*) with its nearest neighbors (*j*, set to a maximum of 6). **(B)** The impact of a single point mutation requires evaluating all the contacts for each residue of θ_m. Mutation of site 117 is illustrated, members (*i* residues) are listed in the top row and neighbors (*j* residues) are listed in the corresponding columns. Columns with less than 6 neighbors indicate solvent exposed residues.

Important correlations that govern the free energy are hidden, i.e. there are contributions that are non-negligible, non-local, and not easily identifiable by current models and rationales.[60, 61] Local interaction-based rationales, although possibly sound in terms of enthalpy, have been broadly recognized as untenable decompositions of free energy. And yet, perhaps because no alternative is at hand, this fundamental of thermodynamics is occasionally disregarded, and attempts to justify free energy-governed processes in terms of local atomic interactions are advanced. For example, in drug discovery, small molecule-protein binding free energy is sometimes rationalized in terms of individual interactions between the small molecule

and the protein.[62] Recent studies based on sound theories use all-atom simulations to estimate free energy.[15, 16, 47, 63] These computed free energies are not decomposed into individual interactions, even in cases where free energy is binned on a per-residue basis[20,[64, 65] are determined by thermodynamic cycles that swap solvent arrays,[14, 66-68] or dissected in terms of many-body correlated changes.[60, 61] These approaches aim to sample over multiple timescales in an effort to solve the scaling problem of protein energetics. Similarly, the model presented here does not attempt to decompose free energy into isolated individual interactions.

The structure of the model is reminiscent of earlier ideas on phase transition behavior of polymer-, protein-, and Ising-based statistical frameworks.[23, 69-76] The interaction factor, μ, defined in **Eq. 1**, is a function of the average per-residue effects (γ) and local protein conformation (σ). The propagation length ($\xi = 10$), corresponds to the standard hydropathic sliding window average, reflects sharp cut-offs of phase transition-like critical effects, and is consistent with extended Zimm-Bragg models. Importantly, the γ parameter includes critical factors such as solvation and excluded volume properties, among others. Although highly simplified, these ideas are in line with the energetic and dynamical properties of proteins being tuned to the energetic and dynamical properties of water, the long-range effects of solvation and liquid-vapor coexistence models, and protein-solvent fluctuations.[77-79] Taken together, this model offers a framework to approximate protein interaction free energy in terms of per-residue contributions.

In addition to considering conformation in terms compatible with dynamic solvated surfaces (σ-index), the model uses multiple fractal terms defined for each amino acid. These are high-precision terms that uniquely capture properties related to accessible surface area of each amino acid residue in proteins. Proteins have been described in terms of fractal dimensionality

and much recent effort has been devoted to using protein fractal dimension to describe protein physics.[24-26, 80-82] Additionally, proteins have been described in terms of hybrid numerical-dynamical models to assess per-residue contributions to free energy.[20,[64, 65] This study provides the first example of per-residue fractal dimension being used to describe protein energetics.

The generality of these findings requires further study. For example, although the other parameters of the model are not specific to **T4L**, additional studies are required to determine whether magnitude of the scaling parameter (λ) is specific to this protein. Nevertheless, the per-residue approximation of interaction free energy presented here appears promising, compares favorably to molecular simulations using state-of-the-art potentials and machine learning estimations, is easily visualized, and has the potential to be of considerable utility. Comparison to a recent study, where a state-of-the-art all-atom physical potentials simulation method (CHARMM36H) was used, points out the advantages of the presented model. A very good fit for the barnase mutant set was obtained through extended dynamics simulation (correlations of 0.67+),[83] and although computationally and setup intensive, it is indeed comprehensive. However, it does not enable straight forward per-residue interpretation, in contrast to the presented coarse-grained model.

Conclusion

Protein interaction free energy is the lynchpin to the understanding of many biological processes and to biotechnology applications, including protein and peptide engineering,[48, 49, 84] design and development of biologics as well as small molecule drugs,[85] among many others. Despite the temptation of describing protein stability and energetics in terms of molecular interactions such as ionic interactions and hydrogen bonding, these interactions are

often too numerous, varied, weak, and delicately balanced to allow meaningful prediction of per-residue contributions to interaction free energy. The model presented here outlines an approach to describe protein interaction energies based on sequence and per-residue contacts. This approach provides a straight forward basis to describe the individual and correlated effects of per-residue contributions to interaction free energy. As an example, T4 lysozyme was analyzed using the presented approach. Residues modeled to contribute significantly to interaction free energy of T4 lysozyme are depicted as hot in heat map renderings of the protein, with the interior of the protein dominated by hot residues and the exterior dominated by cool residues. The most notable exception to cool surface residues is the large hot patch on the surface that corresponds to the substrate binding surface of this enzyme.[57, 58] We used the model to classify mutations as primarily impacting ground state conformations (MC-I) or ground and/or excited state conformations (MC-II). In addition to these qualitative insights, differences in thermal properties, $\Delta\Delta G$, of T4L single point MC-I mutants were calculated based on the observed ground state structure, without recourse to dynamic simulations. Conveniently, the model requires insignificant computing power. The calculation of the μ values and generation of the heat maps require a few seconds of computing time on a dual processor (laptop). The agreement of this simple model with experiment is encouraging.

Although further studies on the scope and limitations studies of the model are warranted, the simplicity, directness, qualitative insights, and accuracy of the approach suggests potentially broad utility as well as complementarity to all-atom and bioinformatics-based protein modeling.[83, 86, 87] The most exciting implications of this work are that the high complexity of protein structure can be reduced to a simple network of readily visualized and interpreted interaction factors and that it may be possible to convert structure space – so abundant in the wake of the structural genomics revolution – into energy space.

Appendix

The two conjectures of the model flow from the two lines of inquiry. Firstly, *What information is contained in, and what information is excluded from, power law exponents derived from accessible surface areas?* Conjecture #1: The exponents contain the fixed contributions to compaction for each amino acid residue, including the combined effect of solvation, steric/excluded volume, dispersion and van der Waals interactions, as well as stereoelectronic, polar, and Coulombic effects. The exponents do not contain contributions that vary from protein structure to protein structure, including temperature, medium effects, or sequence. Moreover, secondary, tertiary, and quaternary structural details, destabilizing charge-charge interactions, and other fine structural details have been removed as well. Idiosyncratic fluctuations have been averaged out of these exponents. Protein backbone and side chain entropy penalty factors would be expected to be largely absent as well, because these do not correlate strongly with side chain accessible surface area. Still, much information intrinsic to the amino acid residues in a protein context should be included in these exponents. A model that aimed to use these per-residue exponents as reduced complexity descriptors of protein interaction free energy would have to include the effects that are omitted and would have to build in the effects of sequence and conformation. Secondly, *How are these intrinsic per-residue effects propagated in proteins?* Conjecture #2: The effective contribution of an individual residue can be described by normalization of its intrinsic contributions with certain residues nearby in the protein sequence as a function of the amino acid residue's assigned conformation (σ-index, see **S2.1**). A physical interpretation of this is as a hydrated surface centered on the residue of interest and corresponding to an extended motif revealed upon partial unfolding.[19, 39, 40] Normalization in this way simultaneously includes multibody interaction, local structure,

and local sequence effects and is described in more detail below. The model equation parameters are listed in **S2.1**.

Proteins are considered at a level of coarseness that removes all molecular fine structure such that they are approximated as flexible chains with identical blobs uniformly spaced along the chain. The chain represents the protein backbone, blobs represent the side chains of residues, identical spacing is a consequence of proteins being composed of α-amino acid building blocks, and proximity of blobs is indicative of the impact of compaction. A set of simple rules determines the *σ*-index of each blob. Proximity of blobs renders them nearest neighbors. As shown in **Figure 2.2**, CLNNs are within four residues along the chain (i.e., *σ*-index of 1, 2, 3, or 4) and have sidechains that are in contact. This guarantees that the *σ*-index refers to residues that are both clustered together and nearby in terms of sequence. Viewed linearly, a blob is assigned a σ-index based on the translational relationship, the number of steps within the interval determines the index. This is applicable to all residues in a protein, including extended strands, helical motifs, various turns and loops. The CLNN definition captures regular secondary structure designations, i.e., extended *σ* = 2 intervals correspond to beta strands, repeated *σ* = 3 intervals to the 3_{10}-helix motif, and repeated *σ* = 4 intervals to the canonical alpha-helix. The *σ*-index captures the essence of conformation and subsumes the many varied combinations of backbone and side chain angles of contacting residues.

Equation 1 expresses the ideas presented above in compact form. Accordingly, for each residue *i* with index *σ*, the *γ*-values of θ_d^i are summed, normalized, and then assigned as interaction factor μ_i^σ. The μ term thus captures the effective sequence, conformation, and fluctuation dependence of each amino acid residue in a given fold of a given protein. As shown

below, the interaction free energy of adjacent residues, *i* and *j*, is relatable to the product of adjacent interaction factors ($\mu_i^{\sigma}\mu_j^{\sigma}$). See also, **S2.1**.

Equation 2 indicates that the free energy is relatable to the sum of the products of interaction factors of residues *i* with nearest neighbors *j*. The sign indicates the favorability of low energy, ε is a medium parameter, λ is a scaling factor (fit as 12.5 kcal/mol, see **Fig. 2.6**), $\tau_{i,j}$ is an efficiency parameter of the *i,j*-interaction (e.g., polar/non-polar compatibility). See also, **S2.1**.

Equation 3, given below, describes how the change in free energy upon mutation (ΔΔG) was computed. λ is a scaling parameter; both ε and $\tau_{i,j}$ are set to 1; *T* is the melting temperature in Kelvin of the wild type protein; ΔT_m is the change in melting temperature of the mutant. Since only the native state is analyzed, S_i^{bb} and S_i^{sc} are protein backbone (bb) and side chain (sc) estimates of entropy change from non-native to native state. **Eq. 3** captures the net favorable interactions between residues *i* and *j* (as $\mu_i\mu_j$) and penalties of folding for wild type and mutant (as S_i^{bb} and S_i^{sc}). We use backbone and side chain entropy values determined from evaluation of native and non-native state ensembles to approximate these confinement penalties.[59] For simplicity we set the number of neighbors (*j*) to a maximum of 6. Surface residues with less than 6 neighbors are assigned phantom solvent neighbors. Although formally based on pairwise nearest neighbor interactions, **Eq. 3** includes implicit multibody effects captured in μ_i^{σ} (**Eq. 1**). Even for mutants that adopt the same fold as the wild type protein, there may be significant differences in many μ terms, and in this way multibody effects dominate the interaction free energy. See also, **S2.1**.

$$\Delta\Delta G = \left[-\lambda\varepsilon \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\,\mu_i\mu_j - (T+\Delta T_m)\sum_i^n\left(\Delta S_i^{bb}+\Delta S_i^{sc}\right)\right]_{Mutant} - \left[-\lambda\varepsilon \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\,\mu_i\mu_j - T\sum_i^n\left(\Delta S_i^{bb}+\Delta S_i^{sc}\right)\right]_{Wild\ Type}$$

Eq. 3

Acknowledgements

The authors wish to thank Prof. J. C. Phillips (Rutgers University) for introducing them to Ref. 31, Prof. Phillips and Dr. Doug Allan (Corning) for stimulating conversations, and Prof. M. A. Moret (Universidade Estadual de Feira de Santana, Brazil) for sharing the amino acid gamma expontents and standard deviations (*personal communication to LJW*). This work was partially funded through support from Corning Inc. and by NIH grants P41EB002503, R01EB020036, and T32GM008339.

Supporting Information

Supporting Information Content:

S2.1. Parameters for Equations 1-3.

Each parameter for each equation is described below.

Equation 1. For each residue *i* with index *σ*, the *γ*-values of θ_a^i are summed, normalized, and then assigned as interaction factor μ_i^σ. The μ term thus captures the effective sequence, conformation, and fluctuation dependence of each amino acid residue in a given fold of a given protein. In **Eq. 2**, the interaction energy of adjacent residues, *i* and *j*, is relatable to the product of adjacent interaction factors ($\mu_i^\sigma \mu_j^\sigma$).

$$\mu_i^\sigma = (2\xi_\sigma + 1)^{-1} \sum_{n=-\xi_\sigma}^{\xi_\sigma} \gamma_{i+\sigma n} \ , \ \xi_\sigma = \left\lfloor \frac{\xi}{\sigma} \right\rfloor \quad \textbf{Eq. 1}$$

γ-Accessible Surface Area Fractional Exponent Calculations. The accessible solvent area calculations were reported by Moret.[31] They found that the relative accessible surface area for each amino acid type as a function of peptide segment length obeyed a power law with a fractional exponent. The exponent value was specific to the amino acid: 1/16 < *γ* < 1/4. The power law behavior was estimated to emerge for segments longer than ~8 residues.

σ-Index and Closest-Linked Nearest Neighbors (CLNNs). A detailed procedure of *σ*-index assignment, which is determined from CLNNs, is given here. The protein pdb file coordinates were used to determine *candidate CLNNs* as follows and in this order: 1. Phantom methyl groups were added to glycine residues (with alanyl residue stereochemistry) for the purpose of sigma assignment only. 2. For each non-H side-chain atom of each residue, *i*, the shortest distances between residues in the interval *i+/-4* were determined. 3. Residue pairs (*i*, *i+/-n where n = 1, 2, 3 or 4*) within 6.5 Å were designated *candidate clNNs = n*. The distance, 6.5 Å,

was selected because this cut-off, coupled with the following steps, the distance that reliably recapitulated the canonical α-helix and β-strand secondary motifs.

The *candidate CLNNs* were used to determine CLNNs as follows and in this order: Step 1. **Smaller span in same direction wins** – If a residue were to have more than one type of *candidate clNN* designation in the same direction (n vs n' in the C-terminal direction or n vs n' in the N-terminal direction, where n < n') the larger *candidate CLNNs* designation (n') would be eliminated (e.g. if there were contacts between residues *i, i+2* and *i, i+4* then the *i, i+4* would be ignored and the *candidate CLNNs* would be 2; similarly, if there were contacts between *i, i+4* but there were also contacts between residues *i+2* and *i+3* then the *i, i+4* contact would be ignored and the *candidate CLNNs* for residue *i+2* and *i+3* would be 1 and no value would be assigned to residue *i* or *i+4* based on this rule). [*Exception: if both* 3 *and* 4 candidate CLNN *designations were evident then both would be retained and the* candidate *CLNN temporarily would be 3/4.*] Step 2. **Larger span in opposite direction wins** – If a residue were to have more than one type of *candidate clNN* designation in opposite directions (n vs n') the smaller *candidate CLNN* designation would be eliminated (e.g., if there were contacts between residues *i, i-2* and *i, i+4* then the *i-2* contact would be ignored and the *candidate clNN* would be 4). [*Exception: if both* 3 *and* 4 candidate *CLNN designations were evident then both would be retained and the* candidate clNN *would be 3/4.*] Step 3. **clNN designations must be reciprocal** – If a residue were to have a *candidate CLNN* of *n* but residue *i+n* did not have the reciprocal *candidate CLNN* of *n* then the *candidate CLNN* would be removed (e.g., if residue 34 were to have a contact with residue 38 but because of one of the preceding rules residue 38 were to ignore the contact to residue 34 then the *candidate CLNN* = 4 for residue 34 would be removed). Step 4. **Unassigned and in an interval** – If a residue had no *candidate clNN* designation but were to fall between residues that follow the reciprocal rule (Step #3, above) then that residue would be assigned the same

candidate CLNN designation as the reciprocal residues (e.g., if residue *i* and *i*+4 were to satisfy the above rules such that their *candidate CLNN = 4* and residues *i+2* and *i+3* were not in contact with any residues then the *candidate CLNN* for residues *i+2* and *i+3* would be 4 also). Step 5. **Unassigned and outside an interval** – If a residue had no *candidate CLNN* designations after applying the above rules the *candidate CLNN* designation would be 1. Step 6. **Blocks of 3/4 become 4** – After applying the above rules, all *candidate CLNN* designations of 3/4 would be reassigned as 4. Step 7.

<u>*σ*-Index assignment.</u> The above steps convert possible *candidate CLNN* designations to single *candidate CLNN* designations of 1, 2, 3, or 4 for each residue of the protein. These were designated the CLNN and used as the sigma assignments. The assignments match well the β-strand/β-sheet and α-helix motifs assigned by other methods, and assign turns, loops, etc. as *σ* = 1, 2, 3, or 4 as well. **Block Check**: As a final check of the *σ*-index assignment, each *σ*-index assignment was verified as being part of a clNN block of at least n+1 members with the first residue in the block having an *i+n* contact and the last residue of the block having an *i-n* contact. [*Exception: residues assigned n = 1 may be in blocks of only 1 member.*] For example, a block of *σ* = 4 would be composed of at least 5 consecutive residues of *σ* = 4. In this way a string of *σ* = n may be isolated and short (n+1 residues only), long and uninterrupted (n+x, where x>>1), long but composed of multiple blocks, etc.

<u>ξ Parameter.</u> A key concept explored in this model is that the interaction factor of residue *i* depends on a subset of the nearby residues termed the determining, or hidden symmetry, set (θ_d^i). This set serves as a proxy for the hidden correlations that impact residue *i*, whatever the fine-structural details might be. A simple relationship defines this set for residue *i* of index *σ.* Specifically, $i + n\sigma_i \in \theta_d^i$, where $n = -\xi_\sigma, \ldots -2, -1, 0, 1, 2, \ldots \xi_\sigma$, and $\xi_\sigma = \lfloor \xi / \sigma_i \rfloor$ (**Eq. 1**). We term ξ

the propagation length. The propagation length corresponds to a type of sliding window. θ_d^i is a comb filter of fixed band width. The width (ξ = 10), which will be discussed in greater detail elsewhere, matches the standard sliding window width of the Kyte-Doolittle hydropathic analysis. The normalization length aims to capture the critical role that residues proximal in sequence play in phase transition-like behavior.[24, 25, 77, 78] For example, the θ_d^{117} of T4 lysozyme (for which σ = 4, ξ_σ = 2, and n = -2, -1, 0, 1, 2) contains five residues (θ_d^{117} = 109, 113, 117, 121, 125). Mutation of any of these five residues will impact the interaction factor of residue 117, and hence the computed interaction energies attributable to residue 117. Not all elements of θ_d^i will necessarily have the same σ-index in the observed protein structure as that assigned to i. This is attributable to the periodic relationship of the set, i.e., the translational symmetry of the blobs defined by σ as an extended, if hypothetical, motif. Returning to the example above, residue 117 adopts σ = 4. The elements of θ_d^{117} are independent of whether all the residues in this set, i.e., 109, 113, 121, and 125, adopt σ = 4 in the native fold. This can be considered as the model describing certain folding fluctuations as making important contributions.

Equation 2 indicates that the free energy is relatable to the sum of the products of interaction factors of residues i with nearest neighbors j ($\mu_i^\sigma \mu_j^\sigma$). The sign indicates the favorability of low energy, ε is a medium parameter, λ is a scaling factor with units of energy, $\tau_{i,j}$ is an efficiency parameter of the i,j-interaction (e.g., polar/non-polar compatibility). For this study, $\varepsilon = \tau = 1$. The value for λ (fit using **Eq. 3**) was estimated as 12.5 kcal/mol. This is the only parameter that was specifically fit to this protein; further studies will be required to determine the generality of this value for other proteins.

$$\Delta G \sim -\varepsilon\lambda \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\,\mu_i\mu_j \qquad \textbf{Eq. 2}$$

Nearest Neighbors. A sidechain-sidechain contact matrix was generated for each residue (*i*) using a 5.5 Å cut off. For residues with more than 6 neighbors (*j*), the closest 6 were selected. Surface residues with less than 6 neighbors are assigned phantom solvent as neighbors. See also **Table S2.6**.

Per-residue Interaction Free Energy: Total vs. Partial. Eq. 1 gives the interaction factor μ, from which the per-residue contribution to interaction free energy may be estimated (via Eq. 2 or, for single point mutants, by Eq. 3). For example, given two residues in contact where $\mu_i = \mu_j$ = 0.170 ($\varepsilon = \tau = 1$; λ = 12.5 kcal/mol) the estimated interaction energy would be -0.36 kcal/mol, with -0.18 kcal/mol attributable to each residue. For $\mu_i = \mu_j$ = 0.150, the per-residue contribution would be -0.14 kcal/mol; and for $\mu_i = \mu_j$ = 0.130, the per-residue contribution would be -0.11 kcal/mol. For a change in interaction free energy for μ_i = 0.130 going to μ_i = 0.150 (e.g., because of mutation at another site, where the central residue (*i*) has six neighbors μ_i = 0.150), the approximate total interaction free energy change attributable to residue *i* would be: -0.11 kcal/mol (from: (-0.84 kcal/mol - (-0.73 kcal/mol)) = -0.11 kcal/mol).

Equation 3 describes how the change in free energy upon mutation ($\Delta\Delta G$) was computed. As indicated above, λ is a scaling parameter with units of energy; both ε and $\tau_{i,j}$ are set to 1; T is the melting temperature in Kelvin of the wild type protein; ΔT_m is the change in melting temperature of the mutant. Since only the native state is analyzed, S_i^{bb} and S_i^{sc} are protein backbone (bb) and side chain (sc) estimates of entropy change from non-native to native state. **Eq. 3** aims to capture the favorable interaction free energy between residues *i* and *j* as $\mu_i\mu_j$ and

the penalties of folding for wild type and mutant as S_i^{bb} and S_i^{sc}. Although formally based on pairwise nearest neighbor interactions, **Eq. 3** includes implicit multibody effects captured in μ_i^{σ} (**Eq. 1**), and in this way multibody effects dominate the interaction free energy.

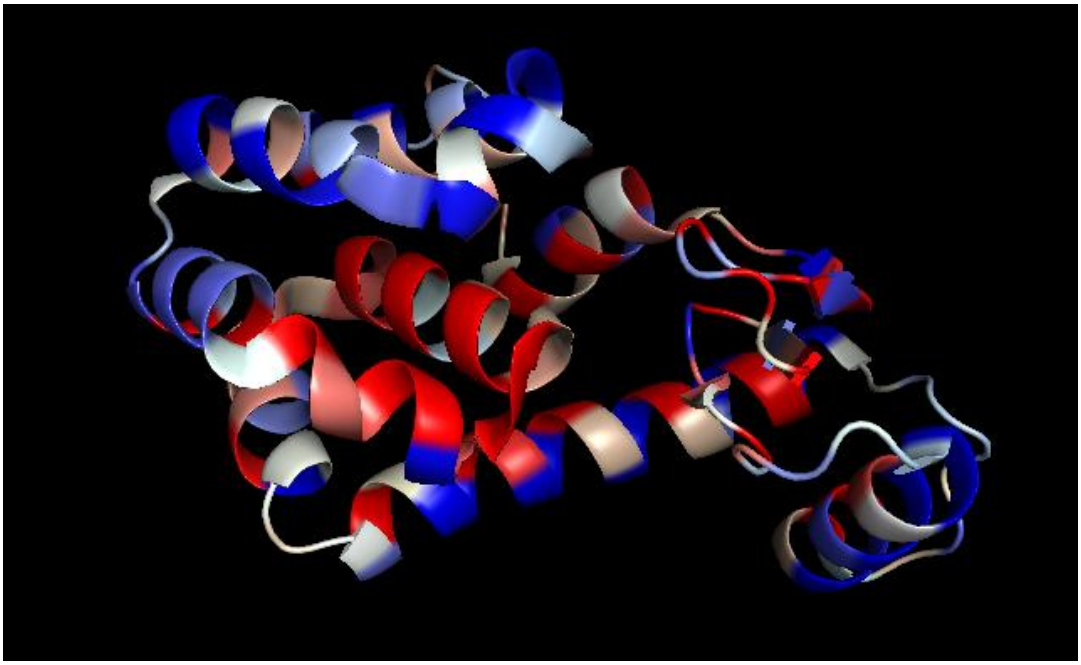
$$\Delta\Delta G = \left[-\lambda\varepsilon \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\, \mu_i \mu_j - (T + \Delta T_m) \sum_i^n \left(\Delta S_i^{bb} + \Delta S_i^{sc}\right)\right]_{Mutant} - \left[-\lambda\varepsilon \sum_{i,j\ (i\neq j)}^{n} \tau_{i,j}\, \mu_i \mu_j - T \sum_i^n \left(\Delta S_i^{bb} + \Delta S_i^{sc}\right)\right]_{Wild\ Type}$$

Eq. 3

Backbone and Sidechain Entropy Estimates. S_i^{bb} and S_i^{sc} are protein backbone (bb) and side chain (sc) estimates. Backbone and side chain entropy values were approximated using published estimates of native and non-native state ensemble confinement penalties.[59] Specifically, the S_i^{bb} values used were based on σ-index by analogy to regular secondary motifs (cal/mol·K): for helix (σ = 4 and 3), S_i^{bb} = 1.6 cal/mol·K; for strand (σ = 2), S_i^{bb} = 1.1 cal/mol·K; for coil (σ = 1), S_i^{bb} = 1.2 cal/mol·K; except for glycine (G: S_i^{bb} = 1.3 cal/mol·K) and proline (P: S_i^{bb} = 0.4 cal/mol·K) and the residue, X, that precedes proline (X: S_i^{bb} = 0.8 cal/mol·K. The S_i^{sc} values used were based on the amino acid identity (cal/mol·K): A = 0; C = 0.14; D = 0.28; E = 0.42; F = 0.28; G = 0; H = 0.28; I = 0.28; K = 0.56; L = 0.28; M = 0.42; N = 0.28; P = 0.14; Q = 0.42; R = 0.56; S = 0.14; T = 0.14; V = 0.14; W = 0.28; Y = 0.28.

Figure S2.1. Heat maps of T4L, myoglobin, ubiquitin, barnase, staphylococcal nuclease, ribonuclease A (2 renderings: residue side chains shown as lines in bottom image). Hydrogen atoms, ligands, and in some images foreground residues are omitted for clarity.

A. Heat map of T4 Lysozyme, PDB ID: 3FA0



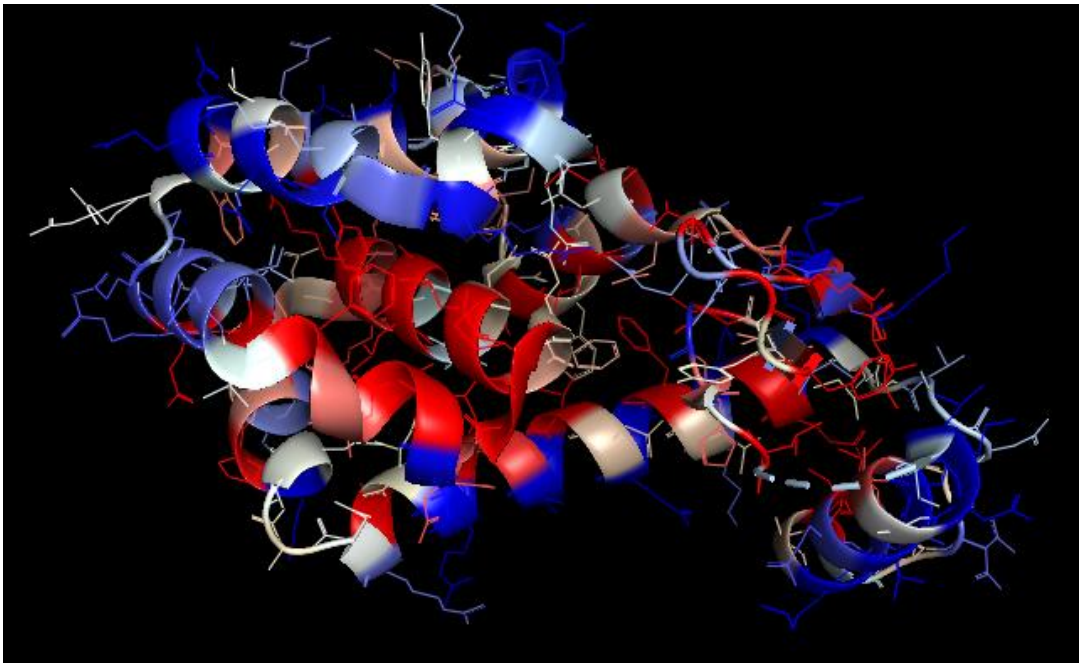
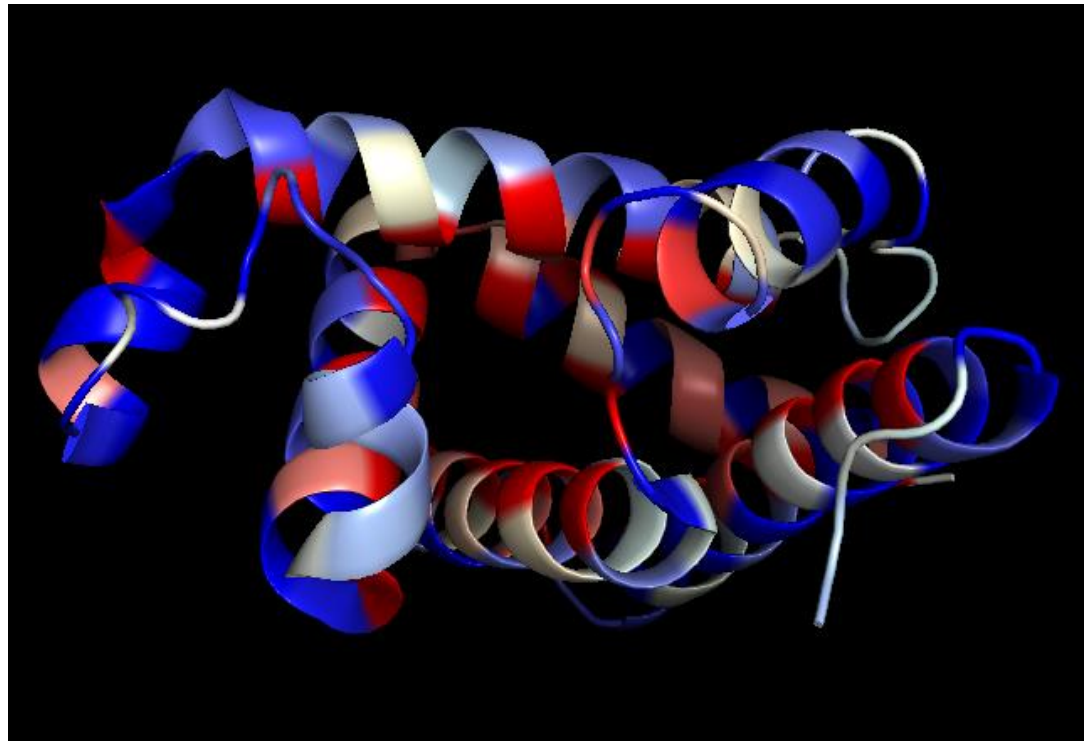


Figure S2.1 B. Heat map of myoglobin, PDB ID: 1MBN



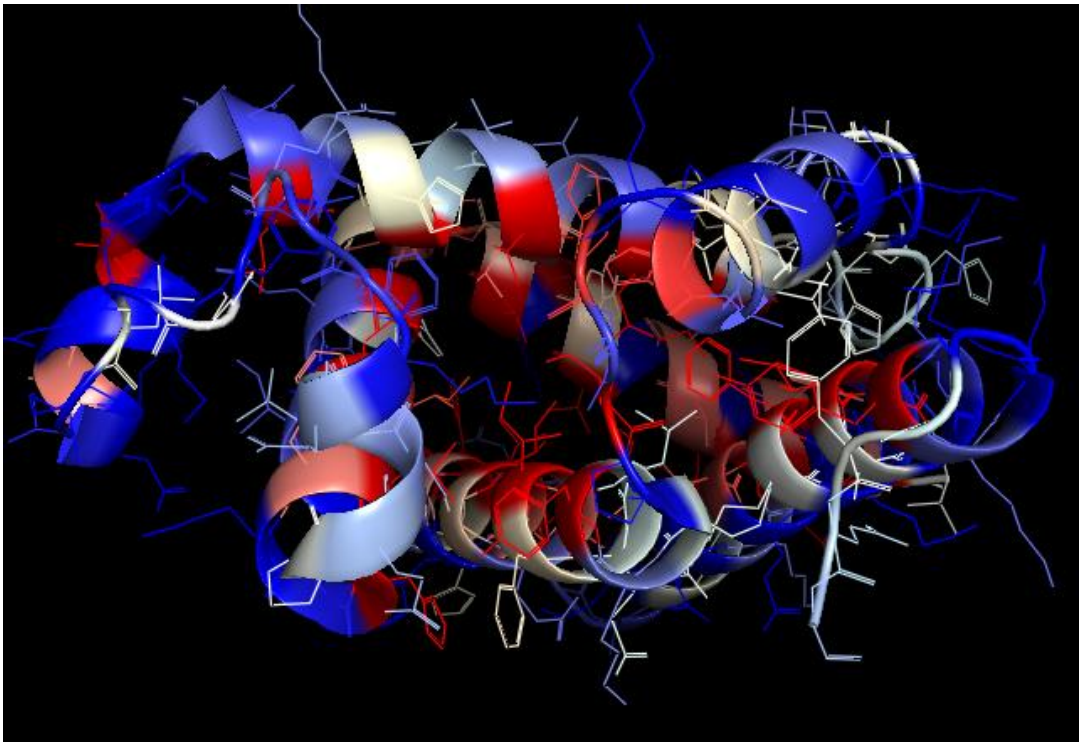
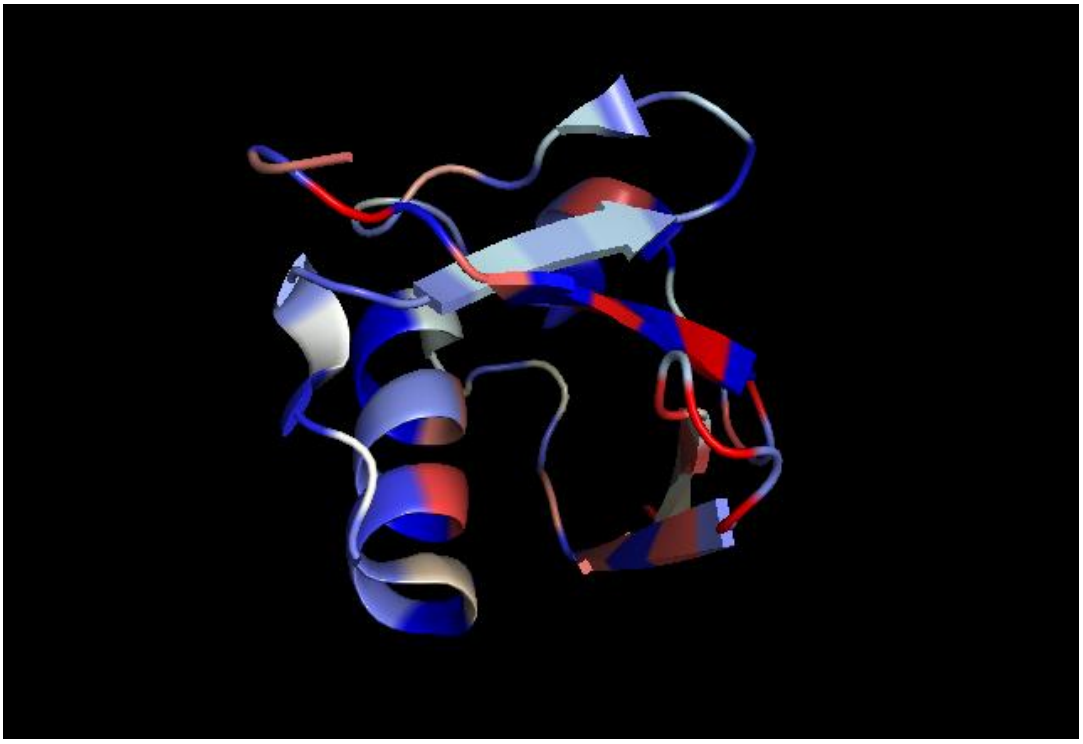


Figure S2.1 C. Heat map of ubiquitin, PDB ID: 1D3Z



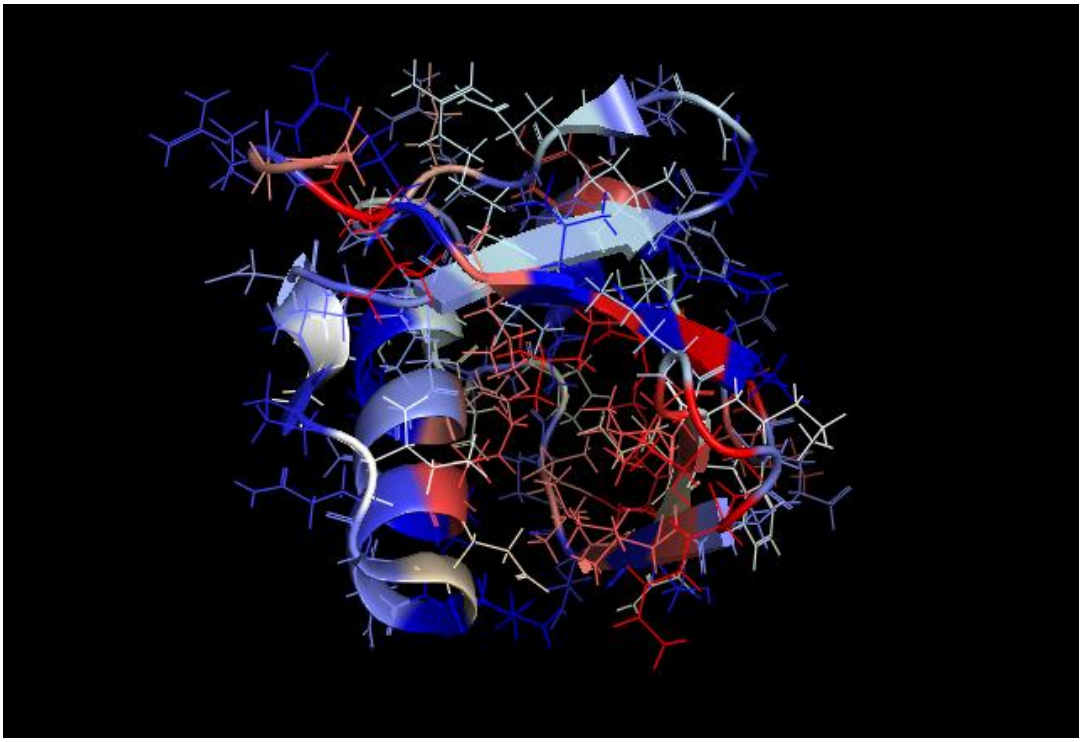
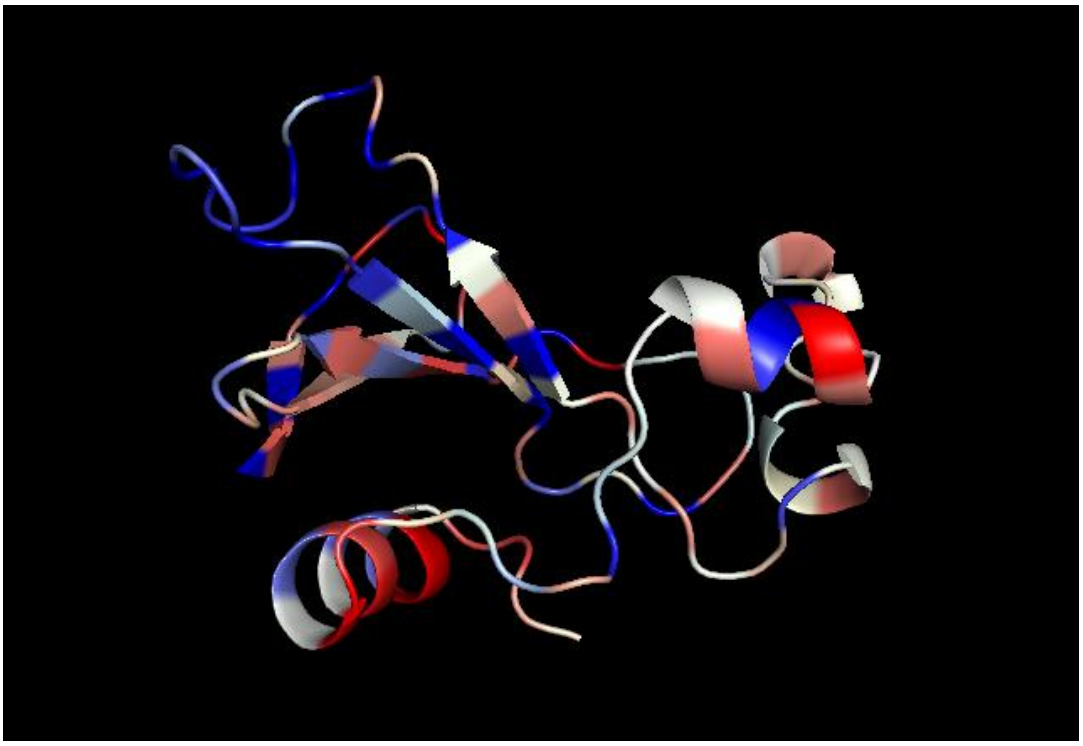


Figure S2.1. D. Heat map of barnase, PDB ID: 1BNI



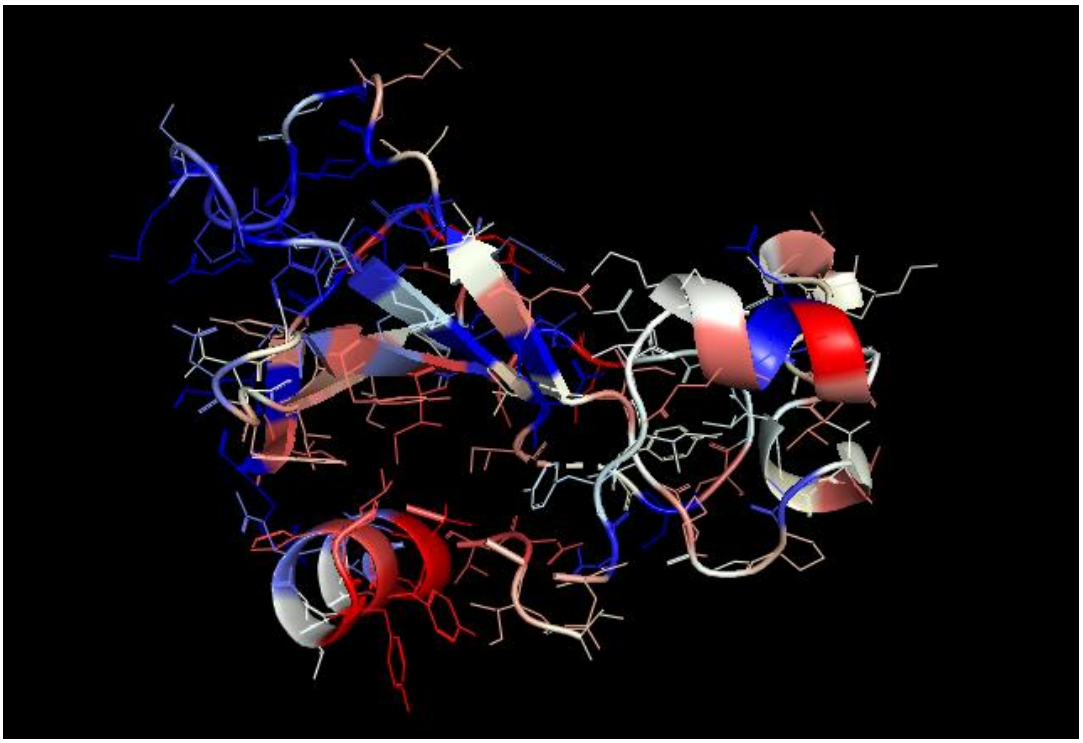
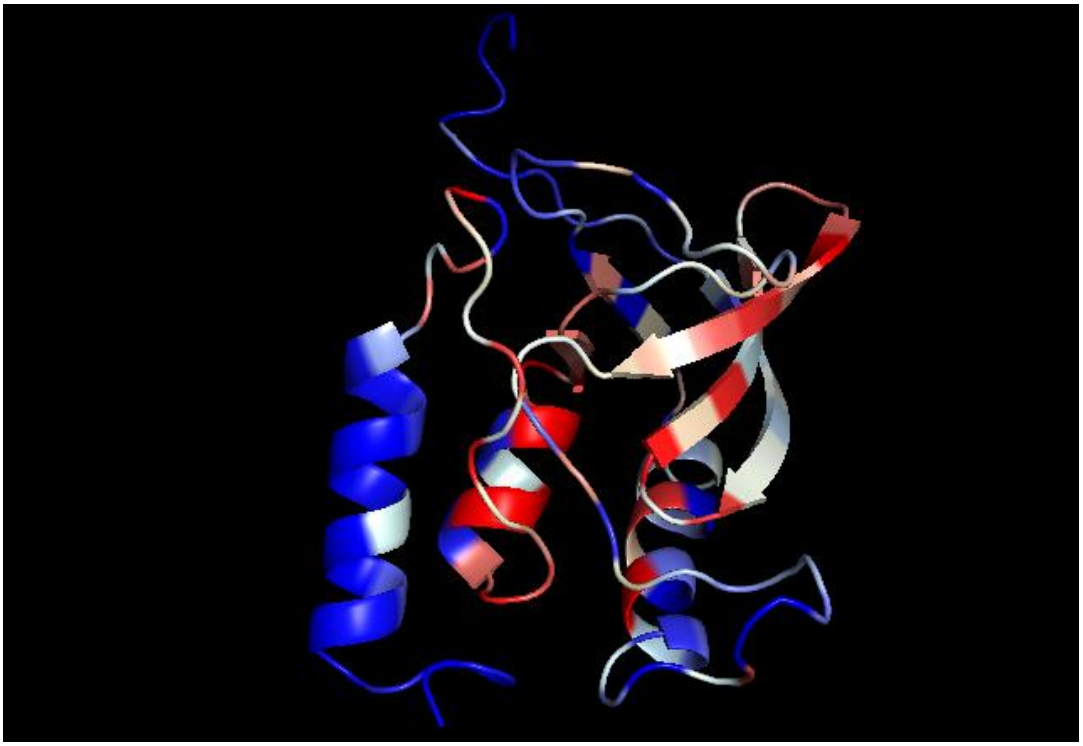


Figure S2.1. E. Heat maps of staphylococcal nuclease, PDB ID: 2SNS



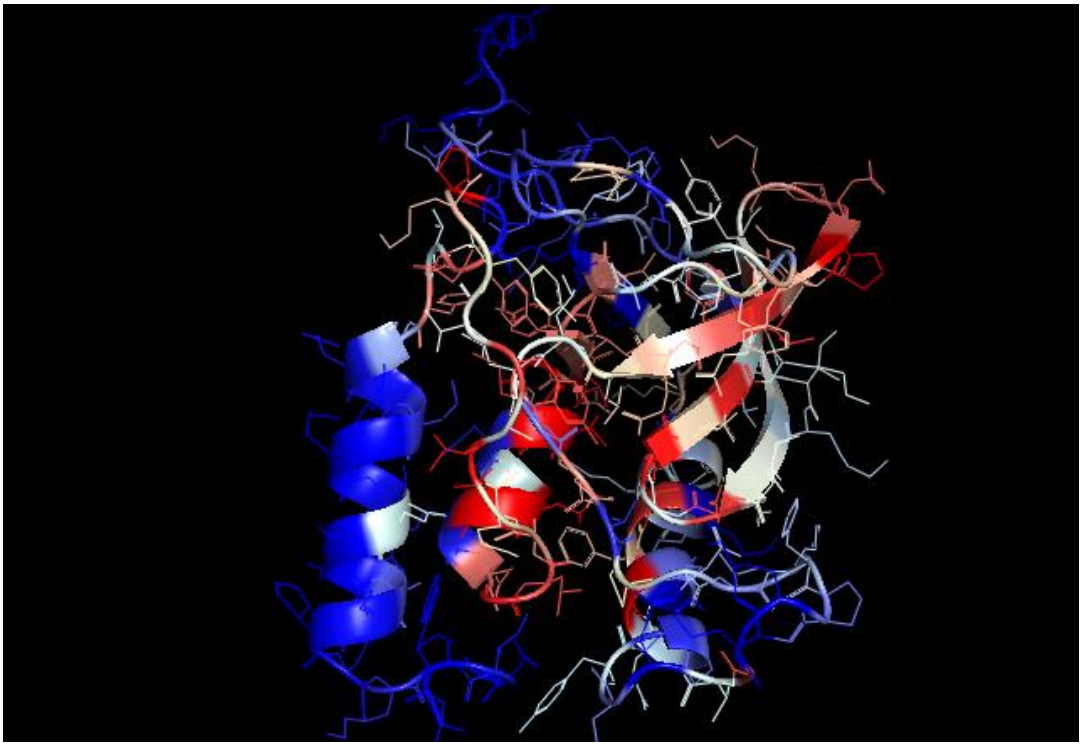
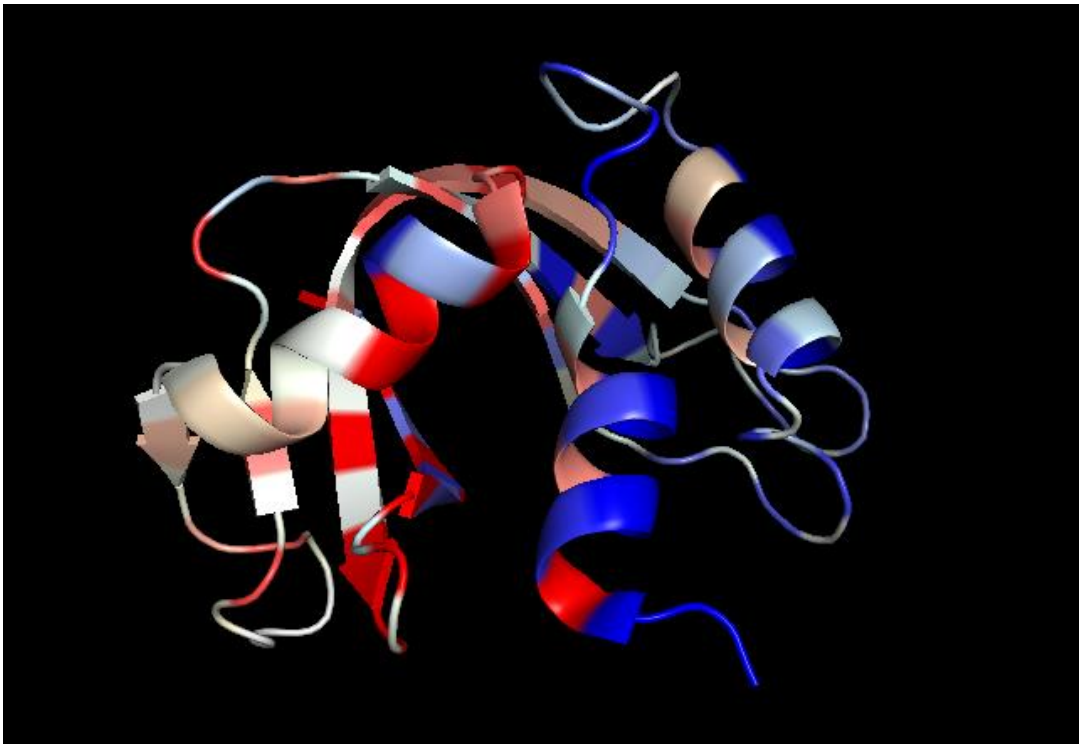


Figure S2.1. F. Heat maps of ribonuclease A, PDB ID: 1FSE



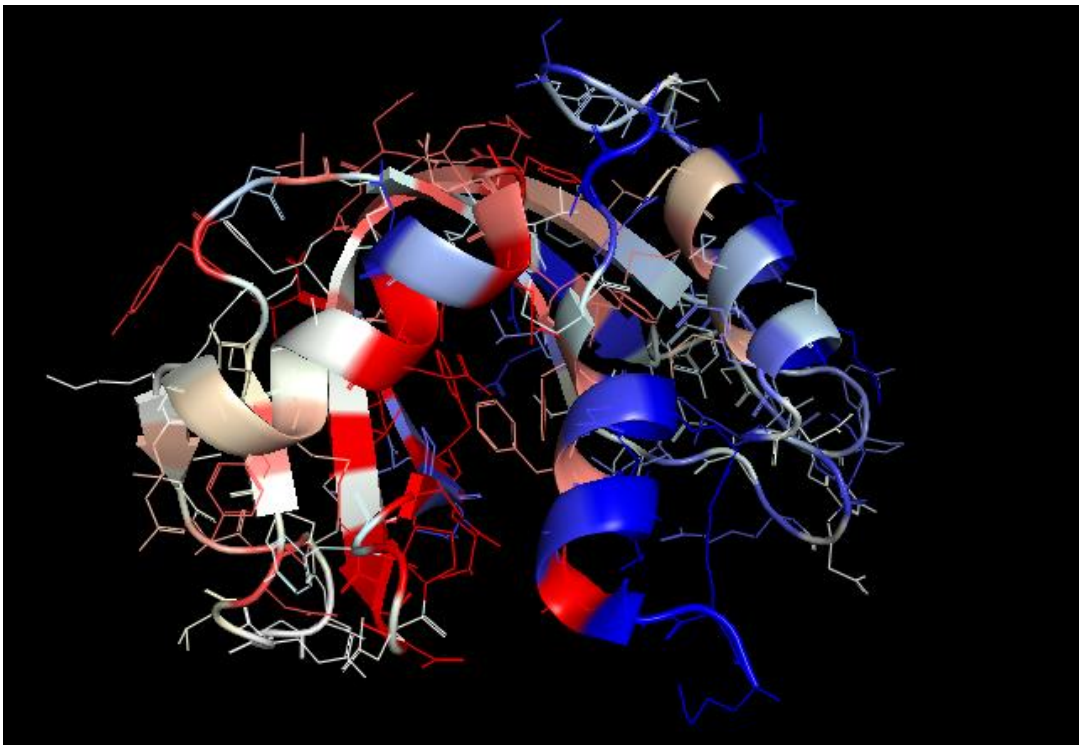


Figure S2.2. Comparison of ΔΔG values calculated with μ_i, μ_j contributions only ($\Delta\Delta G_{Calc|\mu i\mu j}$) with experimental values ($\Delta\Delta G_{Exp}$). $\Delta\Delta G_{Calc|\mu i\mu j}$ was calculated using **Eq. 3** without the entropy and temperature containing terms using (c.f. **Eq. 2**). Linear regression yielded a best fit line (dashed) with the equation y= 1.18x + 0.67 and R^2=0.65. The average unassigned error (white band) is 0.97 kcal/mol.

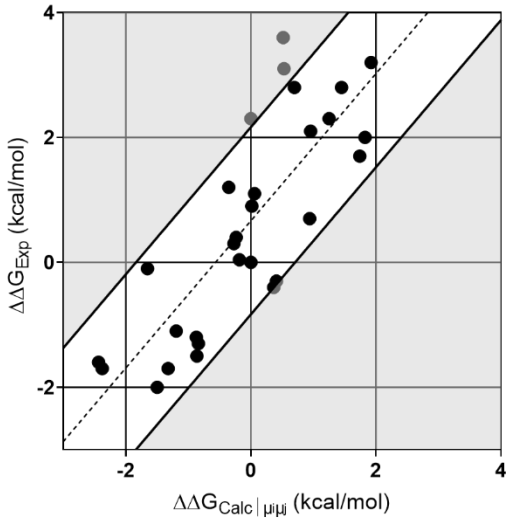


Table S2.1. Calculated μ-factors for T4 Lysozyme sequence. The last four columns give the μ-factor value for a given σ assignment.

Res #	AA	γ *	σ = 1	σ = 2	σ = 4	σ = 3
1	M	99	150	155	138	150
2	N	113	150	146	140	95
3	I	222	154	161	171	205
4	F	218	148	136	151	160
5	E	94	152	165	153	91
6	M	221	146	128	125	204
7	L	197	151	172	178	145
8	R	78	155	138	130	113
9	I	222	150	161	167	207
10	D	87	147	133	144	134
11	E	94	147	159	156	109
12	G	156	146	129	122	196
13	L	197	148	164	174	133
14	R	78	148	139	140	116
15	L	197	149	164	143	196
16	K	69	150	132	122	134
17	I	222	151	176	200	121
18	Y	222	149	126	123	197
19	K	69	155	178	148	128
20	D	87	152	133	138	142
21	T	135	155	172	200	187
22	E	94	160	143	137	137
23	G	156	162	181	159	157
24	Y	222	159	141	146	193
25	Y	222	159	169	200	128
26	T	135	154	143	148	155
27	I	222	157	162	134	179
28	G	156	151	146	152	136
29	I	222	150	160	179	138
30	G	156	152	136	124	176
31	H	152	155	168	159	142
32	L	197	156	142	158	148
33	L	197	155	162	184	179
34	T	135	152	143	137	138
35	K	69	146	157	142	140
36	S	100	145	141	133	160
37	P	121	143	144	158	135
38	S	100	135	135	149	133
39	L	197	136	138	115	139
40	N	113	136	141	116	135
41	A	157	136	132	145	133
42	A	157	132	133	162	139
43	K	69	128	129	116	124
44	S	100	130	138	92	121
45	E	94	129	121	128	146
46	L	197	133	140	184	121
47	D	87	140	137	124	134
48	K	69	145	151	103	165
49	A	157	146	138	152	136
50	I	222	140	148	209	139
51	G	156	139	128	112	147
52	R	78	136	146	94	132
53	N	113	136	128	138	130

Table S2.1. Calculated μ-factors for T4 Lysozyme sequence. *Continued*.

Res #	AA	γ *	σ = 1	σ = 2	σ = 4	σ = 3
54	C	246	137	141	196	147
55	N	113	136	128	130	135
56	G	156	141	149	93	125
57	V	238	142	139	133	162
58	I	222	143	142	196	138
59	T	135	145	141	156	129
60	K	69	141	143	102	167
61	D	87	142	148	122	128
62	E	94	139	131	169	131
63	A	157	143	148	172	157
64	E	94	145	138	104	140
65	K	69	144	160	131	137
66	L	197	143	123	151	156
67	F	218	143	163	197	135
68	N	113	142	129	88	137
69	Q	105	141	160	115	154
70	D	87	138	116	151	131
71	V	238	140	158	210	129
72	D	87	143	124	90	160
73	A	157	142	156	120	141
74	A	157	144	133	164	126
75	V	238	143	148	192	166
76	R	78	145	136	111	138
77	G	156	147	163	120	133
78	I	222	148	138	149	171
79	L	197	146	152	196	138
80	R	78	146	137	132	129
81	N	113	151	160	116	171
82	A	157	144	137	151	154
83	K	69	147	153	188	108
84	L	197	151	151	132	181
85	K	69	148	145	116	165
86	P	121	140	143	168	97
87	V	238	148	146	156	158
88	Y	222	148	151	132	189
89	D	87	147	150	134	97
90	S	100	148	151	155	154
91	L	197	150	142	156	193
92	D	87	155	164	161	102
93	A	157	159	154	134	170
94	V	238	166	169	167	205
95	R	78	161	157	190	123
96	R	78	169	171	165	157
97	C	246	170	165	142	226
98	A	157	163	169	187	128
99	L	197	159	155	173	136
100	I	222	163	163	140	214
101	N	113	169	169	151	138
102	M	221	167	168	199	156
103	V	238	171	165	181	208
104	F	218	173	180	154	148
105	Q	105	169	163	151	164
106	M	221	170	169	195	193
107	G	156	171	165	193	153
108	E	94	169	179	161	168
109	T	135	165	150	122	186
110	G	156	166	185	203	141
111	V	238	165	150	169	171
112	A	157	165	175	161	183
113	G	156	159	149	139	140

Table S2.1. Calculated μ-factors for T4 Lysozyme sequence. *Continued*.

114	F	218	151	161	179	154
115	T	135	144	135	142	159
116	N	113	148	157	131	120
117	S	100	141	133	133	164
118	L	197	138	145	170	140
119	R	78	141	133	129	111
120	M	221	142	151	131	173
121	L	197	146	143	138	143
122	Q	105	140	147	170	123
123	Q	105	142	139	129	155
124	K	69	142	147	122	148
125	R	78	135	131	146	123
126	W	174	134	136	158	134
127	D	87	132	126	115	143
128	E	94	135	142	119	118
129	A	157	137	137	141	145
130	A	157	138	134	153	147
131	V	238	133	139	144	123
132	N	113	130	126	98	128
133	L	197	131	132	123	138
134	A	157	131	127	159	126
135	K	69	131	130	147	129
136	S	100	135	135	107	139
137	R	78	130	129	123	137
138	W	174	130	126	156	114
139	Y	222	137	143	144	138
140	N	113	140	138	103	157
141	Q	105	139	141	139	124
142	T	135	134	136	169	135
143	P	121	139	139	123	143
144	N	113	133	133	108	139
145	R	78	132	133	143	118
146	A	157	136	133	153	140
147	K	69	138	139	136	151
148	R	78	142	140	119	123
149	V	238	138	140	155	153
150	I	222	135	138	153	139
151	T	135	140	140	109	114
152	T	135	137	131	128	169
153	F	218	137	144	178	129
154	R	78	138	131	134	110
155	T	135	140	146	107	177
156	G	156	143	133	132	132
157	T	135	142	155	203	117
158	W	174	147	130	128	181
159	D	87	152	167	119	144
160	A	157	146	137	149	126
161	Y	222	139	155	192	167
162	K	38	140	123	97	125

* Gamma values for residues at the N- and C- termini are multiplied by 0.45 and 0.55 respectively to account for the additional ionic character at these positions.

List S2.1. MC-1 and MC-II T4 Lysozyme Single Point Mutation Sites

[R96X mutations were not included in MC-I because the known extensive polar contacts that are compromised for many mutants (see Ref. 53) is inconsistent with the □ = 1 approximation (Eq. 2 and 3).]

MC-I: 3, 11, 115, 117, 119, 132, 133
MC-II: 86, 102, 105, 131, 157

Table S2.2. Hottest (red colored) residues in T4 Lysozyme (ranked in order of highest to lowest μ values).

Residue	AA	μx10^3
71	V	210
50	I	209
110	G	203
102	M	199
67	F	197
79	L	196
106	M	195
75	V	192
98	A	187
46	L	184
103	V	181
23	G	181
29	I	179
114	F	179
153	F	178
19	K	178
7	L	178
17	I	176
99	L	173
63	A	172
21	T	172
107	G	171
3	I	171
122	Q	170
118	L	170
62	E	169
111	V	169
142	T	169
86	P	168
159	D	167
9	I	167

Table S2.3. Per-residue θ_d and θ_m sets for T4L (based on PDB ID 3FA0).

Residue	AA	θ_d	θ_m
1	M	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,	1, 2, 5, 9,
2	N	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,	1, 2, 6, 10,
3	I	3, 7, 11,	1, 2, 3, 7, 11, 13,
4	F	4, 8, 12,	1, 2, 4, 8, 12, 14,
5	E	1, 5, 9, 13,	1, 2, 5, 9, 13, 15,
6	M	2, 6, 10, 14,	1, 2, 6, 10, 14, 16,
7	L	3, 7, 11, 15,	1, 2, 3, 7, 11, 13, 15, 17,
8	R	4, 8, 12, 16,	1, 2, 4, 8, 12, 14, 16, 18,
9	I	1, 5, 9, 13, 17,	1, 2, 5, 9, 13, 15, 17, 19,
10	D	2, 6, 10, 14, 18,	1, 2, 6, 10, 14, 16, 18, 20,
11	E	3, 7, 11, 15, 19,	1, 2, 3, 7, 11, 13, 15, 17, 19, 21,
12	G	4, 8, 12, 16, 20,	2, 4, 8, 12, 14, 16, 18, 20, 22,
13	L	3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23,	5, 9, 13, 15, 17, 19, 21, 23,
14	R	4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24,	6, 10, 14, 16, 18, 20, 22, 24,
15	L	5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25,	7, 11, 13, 15, 17, 19, 21, 23, 25,
16	K	6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26,	8, 12, 14, 16, 18, 20, 22, 24, 25,
17	I	7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27,	9, 13, 15, 17, 19, 21, 23, 25,
18	Y	8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28,	10, 14, 16, 18, 20, 22, 24, 25, 26,
19	K	9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29,	11, 13, 15, 17, 19, 21, 23, 25, 27,
20	D	10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30,	12, 14, 16, 18, 20, 22, 24, 25, 28,
21	T	11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31,	13, 15, 17, 19, 21, 23, 25, 29,
22	E	12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32,	14, 16, 18, 20, 22, 24, 25, 26, 30, 32,
23	G	13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33,	13, 15, 17, 19, 21, 23, 25, 27, 31, 33,
24	Y	14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34,	14, 16, 18, 20, 22, 24, 25, 28, 32, 34,
25	Y	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,	15, 17, 19, 21, 23, 25, 29, 33, 35,
26	T	18, 22, 26, 30, 34,	16, 18, 20, 22, 24, 25, 26, 30, 32, 34, 36,
27	I	19, 23, 27, 31, 35,	17, 19, 21, 23, 25, 27, 31, 33, 35, 37,
28	G	20, 24, 28, 32, 36,	18, 20, 22, 24, 25, 28, 32, 34, 36, 38,
29	I	21, 25, 29, 33, 37,	19, 21, 23, 25, 29, 33, 35, 37,
30	G	22, 26, 30, 34, 38,	20, 22, 24, 25, 26, 30, 32, 34, 36, 38,
31	H	23, 27, 31, 35, 39,	21, 23, 25, 27, 31, 33, 35, 37, 39,
32	L	22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42,	22, 24, 25, 28, 32, 34, 36, 38, 40,
33	L	23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43,	23, 25, 29, 33, 35, 37, 41,
34	T	24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44,	24, 25, 26, 30, 32, 34, 36, 38, 42,
35	K	25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45,	25, 27, 31, 33, 35, 37, 39, 43,
36	S	26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46,	28, 32, 34, 36, 38, 40, 44,
37	P	27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47,	29, 33, 35, 37, 41, 45,
38	S	28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48,	30, 32, 34, 36, 38, 42, 46,
39	L	31, 35, 39, 43, 47,	31, 33, 35, 37, 39, 43, 47,
40	N	32, 36, 40, 44, 48,	32, 34, 36, 38, 40, 44, 48,
41	A	33, 37, 41, 45, 49,	33, 35, 37, 41, 45, 49, 51,
42	A	34, 38, 42, 46, 50,	32, 34, 36, 38, 42, 46, 50, 51, 52,
43	K	35, 39, 43, 47, 51,	33, 35, 37, 39, 43, 47, 51, 53,
44	S	36, 40, 44, 48, 52,	34, 36, 38, 40, 44, 48, 51, 52, 54,
45	E	37, 41, 45, 49, 53,	35, 37, 41, 45, 49, 51, 53, 55,
46	L	38, 42, 46, 50, 54,	36, 38, 42, 46, 50, 51, 52, 54, 56,
47	D	39, 43, 47, 51, 55,	37, 39, 43, 47, 51, 53, 55, 57,
48	K	40, 44, 48, 52, 56,	38, 40, 44, 48, 51, 52, 54, 56, 58,
49	A	41, 45, 49, 53, 57,	41, 45, 49, 51, 53, 55, 57, 58, 59,
50	I	42, 46, 50, 54, 58,	42, 46, 50, 51, 52, 54, 56, 58, 59,
51	G	41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61,	43, 47, 51, 53, 55, 57, 58, 59,
52	R	42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62,	44, 48, 51, 52, 54, 56, 58, 59, 60,
53	N	43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63,	45, 49, 51, 53, 55, 57, 58, 59, 61,

Table S2.3. Per-residue θ_d and θ_m sets for T4L sequence and the structure indicated observed in PDB ID: 3FA0. *Continued*

Residue	AA	θ_d	θ_m
54	C	44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64,	46, 50, 51, 52, 54, 56, 58, 59, 62,
55	N	45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65,	47, 51, 53, 55, 57, 58, 59, 63,
56	G	46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66,	48, 51, 52, 54, 56, 58, 59, 60, 64,
57	V	47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67,	49, 51, 53, 55, 57, 58, 59, 61, 65,
58	I	48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68,	50, 51, 52, 54, 56, 58, 59, 62, 66,
59	T	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69,	51, 53, 55, 57, 58, 59, 63, 67,
60	K	52, 56, 60, 64, 68,	51, 52, 54, 56, 58, 59, 60, 64, 68,
61	D	53, 57, 61, 65, 69,	51, 53, 55, 57, 58, 59, 61, 65, 69,
62	E	54, 58, 62, 66, 70,	52, 54, 56, 58, 59, 62, 66, 70,
63	A	55, 59, 63, 67, 71,	53, 55, 57, 58, 59, 63, 67, 71,
64	E	56, 60, 64, 68, 72,	54, 56, 58, 59, 60, 64, 68, 72,
65	K	57, 61, 65, 69, 73,	55, 57, 58, 59, 61, 65, 69, 73,
66	L	58, 62, 66, 70, 74,	56, 58, 59, 62, 66, 70, 74,
67	F	59, 63, 67, 71, 75,	57, 58, 59, 63, 67, 71, 75,
68	N	60, 64, 68, 72, 76,	58, 59, 60, 64, 68, 72, 76,
69	Q	61, 65, 69, 73, 77,	59, 61, 65, 69, 73, 77,
70	D	62, 66, 70, 74, 78,	62, 66, 70, 74, 78,
71	V	63, 67, 71, 75, 79,	63, 67, 71, 75, 79,
72	D	64, 68, 72, 76, 80,	64, 68, 72, 76, 80, 82,
73	A	65, 69, 73, 77, 81,	65, 69, 73, 77, 81, 82, 83,
74	A	66, 70, 74, 78, 82,	66, 70, 74, 78, 82, 83, 84,
75	V	67, 71, 75, 79, 83,	67, 71, 75, 79, 82, 83, 84,
76	R	68, 72, 76, 80, 84,	68, 72, 76, 80, 82, 83, 84,
77	G	69, 73, 77, 81, 85,	69, 73, 77, 81, 82, 83, 84, 85,
78	I	70, 74, 78, 82, 86,	70, 74, 78, 82, 83, 84, 86,
79	L	71, 75, 79, 83, 87,	71, 75, 79, 82, 83, 84, 87,
80	R	72, 76, 80, 84, 88,	72, 76, 80, 82, 83, 84, 88,
81	N	73, 77, 81, 85, 89,	73, 77, 81, 82, 83, 84, 85, 89,
82	A	72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92,	74, 78, 82, 83, 84, 86, 90, 92,
83	K	73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93,	75, 79, 82, 83, 84, 87, 91, 92, 93,
84	L	74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94,	76, 80, 82, 83, 84, 88, 92, 93, 94,
85	K	77, 81, 85, 89, 93,	77, 81, 82, 83, 84, 85, 89, 92, 93, 94, 95,
86	P	78, 82, 86, 90, 94,	78, 82, 83, 84, 86, 90, 92, 93, 94, 95,
87	V	79, 83, 87, 91, 95,	79, 82, 83, 84, 87, 91, 92, 93, 94, 95,
88	Y	80, 84, 88, 92, 96,	80, 82, 83, 84, 88, 92, 93, 94, 95, 96,
89	D	81, 85, 89, 93, 97,	81, 82, 83, 84, 85, 89, 92, 93, 94, 95, 97,
90	S	82, 86, 90, 94, 98,	82, 83, 84, 86, 90, 92, 93, 94, 95, 98,
91	L	83, 87, 91, 95, 99,	82, 83, 84, 87, 91, 92, 93, 94, 95, 99,
92	D	82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102,	82, 83, 84, 88, 92, 93, 94, 95, 96, 100,
93	A	83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103,	83, 84, 85, 89, 92, 93, 94, 95, 97, 101,
94	V	84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104,	84, 86, 90, 92, 93, 94, 95, 98, 102,
95	R	85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105,	87, 91, 92, 93, 94, 95, 99, 103,
96	R	88, 92, 96, 100, 104,	88, 92, 93, 94, 95, 96, 100, 104,
97	C	89, 93, 97, 101, 105,	89, 92, 93, 94, 95, 97, 101, 105, 107,
98	A	90, 94, 98, 102, 106,	90, 92, 93, 94, 95, 98, 102, 106, 107,
99	L	91, 95, 99, 103, 107,	91, 92, 93, 94, 95, 99, 103, 107,
100	I	92, 96, 100, 104, 108,	92, 93, 94, 95, 96, 100, 104, 107, 108,
101	N	93, 97, 101, 105, 109,	92, 93, 94, 95, 97, 101, 105, 107, 109,
102	M	94, 98, 102, 106, 110,	92, 93, 94, 95, 98, 102, 106, 107, 110,
103	V	95, 99, 103, 107, 111,	93, 94, 95, 99, 103, 107, 111,
104	F	96, 100, 104, 108, 112,	94, 95, 96, 100, 104, 107, 108, 112,
105	Q	97, 101, 105, 109, 113,	95, 97, 101, 105, 107, 109, 113,
106	M	98, 102, 106, 110, 114,	98, 102, 106, 107, 110, 114,
107	G	97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117,	99, 103, 107, 111, 115,
108	E	100, 104, 108, 112, 116,	100, 104, 107, 108, 112, 116,
109	T	101, 105, 109, 113, 117,	101, 105, 107, 109, 113, 117,
110	G	102, 106, 110, 114, 118,	102, 106, 107, 110, 114, 118,
111	V	103, 107, 111, 115, 119,	103, 107, 111, 115, 119,
112	A	104, 108, 112, 116, 120,	104, 107, 108, 112, 116, 120,
113	G	105, 109, 113, 117, 121,	105, 107, 109, 113, 117, 121,

Table S2.3. Per-residue θ_d and θ_m sets for T4L sequence and the structure indicated observed in PDB ID: 3FA0. *Continued*

Residue	AA	θ_d	θ_m
114	F	106, 110, 114, 118, 122,	106, 107, 110, 114, 118, 122, 124,
115	T	107, 111, 115, 119, 123,	107, 111, 115, 119, 123, 124,
116	N	108, 112, 116, 120, 124,	107, 108, 112, 116, 120, 124,
117	S	109, 113, 117, 121, 125,	107, 109, 113, 117, 121, 124, 125,
118	L	110, 114, 118, 122, 126,	110, 114, 118, 122, 124, 126,
119	R	111, 115, 119, 123, 127,	111, 115, 119, 123, 124, 127,
120	M	112, 116, 120, 124, 128,	112, 116, 120, 124, 128,
121	L	113, 117, 121, 125, 129,	113, 117, 121, 124, 125, 129,
122	Q	114, 118, 122, 126, 130,	114, 118, 122, 124, 126, 130,
123	Q	115, 119, 123, 127, 131,	115, 119, 123, 124, 127, 131,
124	K	114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134,	116, 120, 124, 128, 132, 134,
125	R	117, 121, 125, 129, 133,	117, 121, 124, 125, 129, 133, 134, 135,
126	W	118, 122, 126, 130, 134,	118, 122, 124, 126, 130, 134, 135, 136,
127	D	119, 123, 127, 131, 135,	119, 123, 124, 127, 131, 134, 135, 136,
128	E	120, 124, 128, 132, 136,	120, 124, 128, 132, 134, 135, 136,
129	A	121, 125, 129, 133, 137,	121, 124, 125, 129, 133, 134, 135, 136, 137,
130	A	122, 126, 130, 134, 138,	122, 124, 126, 130, 134, 135, 136, 138,
131	V	123, 127, 131, 135, 139,	123, 124, 127, 131, 134, 135, 136, 139,
132	N	124, 128, 132, 136, 140,	124, 128, 132, 134, 135, 136, 140,
133	L	125, 129, 133, 137, 141,	124, 125, 129, 133, 134, 135, 136, 137, 141,
134	A	124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144,	124, 126, 130, 134, 135, 136, 138, 142,
135	K	125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145,	127, 131, 134, 135, 136, 139, 143,
136	S	126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146,	128, 132, 134, 135, 136, 140, 144,
137	R	129, 133, 137, 141, 145,	129, 133, 134, 135, 136, 137, 141, 145,
138	W	130, 134, 138, 142, 146,	130, 134, 135, 136, 138, 142, 146,
139	Y	131, 135, 139, 143, 147,	131, 134, 135, 136, 139, 143, 147,
140	N	132, 136, 140, 144, 148,	132, 134, 135, 136, 140, 144, 148,
141	Q	133, 137, 141, 145, 149,	133, 134, 135, 136, 137, 141, 145, 149,
142	T	134, 138, 142, 146, 150,	134, 135, 136, 138, 142, 146, 150,
143	P	135, 139, 143, 147, 151,	134, 135, 136, 139, 143, 147, 151,
144	N	136, 140, 144, 148, 152,	134, 135, 136, 140, 144, 148, 152,
145	R	137, 141, 145, 149, 153,	135, 136, 137, 141, 145, 149, 153,
146	A	138, 142, 146, 150, 154,	136, 138, 142, 146, 150, 154, 156,
147	K	139, 143, 147, 151, 155,	139, 143, 147, 151, 155, 156, 157,
148	R	140, 144, 148, 152, 156,	140, 144, 148, 152, 156, 158,
149	V	141, 145, 149, 153, 157,	141, 145, 149, 153, 156, 157, 159,
150	I	142, 146, 150, 154, 158,	142, 146, 150, 154, 156, 158, 160,
151	T	143, 147, 151, 155, 159,	143, 147, 151, 155, 156, 157, 159, 160, 161,
152	T	144, 148, 152, 156, 160,	144, 148, 152, 156, 158, 160, 161, 162,
153	F	145, 149, 153, 157, 161,	145, 149, 153, 156, 157, 159, 160, 161, 162,
154	R	146, 150, 154, 158, 162,	146, 150, 154, 156, 158, 160, 161, 162,
155	T	147, 151, 155, 159,	147, 151, 155, 156, 157, 159, 160, 161, 162,
156	G	146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162,	148, 152, 156, 158, 160, 161, 162,
157	T	147, 149, 151, 153, 155, 157, 159, 161,	149, 153, 156, 157, 159, 160, 161, 162,
158	W	148, 150, 152, 154, 156, 158, 160, 162,	150, 154, 156, 158, 160, 161, 162,
159	D	149, 151, 153, 155, 157, 159, 161,	151, 155, 156, 157, 159, 160, 161, 162,
160	A	150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162,	152, 156, 158, 160, 161, 162,
161	Y	151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162,	153, 156, 157, 159, 160, 161, 162,
162	K	152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162,	154, 156, 158, 160, 161, 162,

Table S2.4. A. Calculated μ values and color-coding gradient (show right) for T4L sequence and the structure indicated observed in PDB ID: 3FA0.

Color Ranges	Low	High
Red	166	
Salmon	151	165.9999
White	141	150.9999
Cyan	126	140.9999
Blue		125.9999

Residue	AA	ux10^3	sigma-index
1	M	150	1
2	N	150	1
3	I	171	4
4	F	151	4
5	E	153	4
6	M	125	4
7	L	178	4
8	R	130	4
9	I	167	4
10	D	144	4
11	E	156	4
12	G	122	4
13	L	164	2
14	R	139	2
15	L	164	2
16	K	132	2
17	I	176	2
18	Y	126	2
19	K	178	2
20	D	133	2
21	T	172	2
22	E	143	2
23	G	181	2
24	Y	141	2
25	Y	159	1
26	T	148	4
27	I	134	4
28	G	152	4
29	I	179	4
30	G	124	4
31	H	159	4
32	L	142	2
33	L	162	2
34	T	143	2
35	K	157	2
36	S	141	2
37	P	144	2
38	S	135	2
39	L	115	4
40	N	116	4
41	A	145	4
42	A	162	4

Table S2.4. A. Calculated μ values and color-coding gradient. *Continued*.

Res	AA	μx10^3	sigma index	Res	AA	μx10^3	sigma index
43	K	116	4	87	V	156	4
44	S	92	4	88	Y	132	4
45	E	128	4	89	D	134	4
46	L	184	4	90	S	155	4
47	D	124	4	91	L	156	4
48	K	103	4	92	D	155	1
49	A	152	4	93	A	159	1
50	I	209	4	94	V	166	1
51	G	139	1	95	R	161	1
52	R	146	2	96	R	165	4
53	N	128	2	97	C	142	4
54	C	141	2	98	A	187	4
55	N	128	2	99	L	173	4
56	G	149	2	100	I	140	4
57	V	139	2	101	N	151	4
58	I	143	1	102	M	199	4
59	T	145	1	103	V	181	4
60	K	102	4	104	F	154	4
61	D	122	4	105	Q	151	4
62	E	169	4	106	M	195	4
63	A	172	4	107	G	171	1
64	E	104	4	108	E	161	4
65	K	131	4	109	T	122	4
66	L	151	4	110	G	203	4
67	F	197	4	111	V	169	4
68	N	88	4	112	A	161	4
69	Q	115	4	113	G	139	4
70	D	151	4	114	F	179	4
71	V	210	4	115	T	142	4
72	D	90	4	116	N	131	4
73	A	120	4	117	S	133	4
74	A	164	4	118	L	170	4
75	V	192	4	119	R	129	4
76	R	111	4	120	M	131	4
77	G	120	4	121	L	138	4
78	I	149	4	122	Q	170	4
79	L	196	4	123	Q	129	4
80	R	132	4	124	K	142	1
81	N	116	4	125	R	146	4
82	A	144	1	126	W	158	4
83	K	147	1	127	D	115	4
84	L	151	1	128	E	119	4
85	K	116	4	129	A	141	4
86	P	168	4	130	A	153	4

Table S2.4. A. Calculated μ values and color-coding gradient. *Continued*.

Residue	AA	μx10^3	sigma-index
131	V	144	4
132	N	98	4
133	L	123	4
134	A	131	1
135	K	131	1
136	S	135	1
137	R	123	4
138	W	156	4
139	Y	144	4
140	N	103	4
141	Q	139	4
142	T	169	4
143	P	123	4
144	N	108	4
145	R	143	4
146	A	153	4
147	K	136	4
148	R	119	4
149	V	155	4
150	I	153	4
151	T	109	4
152	T	128	4
153	F	178	4
154	R	134	4
155	T	107	4
156	G	143	1
157	T	155	2
158	W	130	2
159	D	167	2
160	A	146	1
161	Y	139	1
162	K	140	1

Table S2.4. B. Calculated μ values and color-coding gradient. *Continued*. Pymol commands for residue color assignments from **Table S2.4 (A)** and used in **Figures 2.4, 2.5 and 2.7**. (For example, 'color blue, resi 6+12+…+155' etc.)

Blue	6+12+18+30+39+40+43+44+47+48+60+61+64+68+69+72+73+76+77+81+85+109+127+128+132+133+137+140+143+144+148+151+155
Cyan	8+14+16+20+27+36+38+45+51+53+54+55+57+65+80+88+89+100+113+116+117+119+120+121+123+134+135+136+141+147+152+154+158+161+162
White	1+2+4+10+22+24+26+32+34+37+41+52+56+58+59+78+82+83+97+115+124+125+129+131+139+145+156+160
Salmon	5+11+13+15+25+28+31+33+35+42+49+66+70+74+84+87+90+91+92+93+94+95+96+101+104+105+108+112+126+130+138+146+149+150+157
Red	3+7+9+17+19+21+23+29+46+50+62+63+67+71+75+79+86+98+99+102+103+106+107+110+111+114+118+122+142+153+159

Table S2.5. Calculated and experimental thermal stability data for MC-I mutants. Experimental data was obtained from cited reference 52.

Mutant	Calc ΔΔG (kcal/mol)	Exp ΔΔG (kcal/mol)	Exp ΔT_{Mutant}
I3A	1.3	0.7	-1.8
I3C (S-H)	-0.09	1.2	-3.7
I3D	2.2	3.2	-8.5
I3E	1.8	2	-5.7
I3F	0.15	1.1	-3
I3G	1.8	2.1	-5.8
I3L	0.34	-0.4	0.9
I3M	-0.07	0.9	-2.3
I3P	3.4	2.8	-7.3
I3S	2.0	1.7	-4.6
I3T	1.6	2.3	-6
I3V	-0.04	0.4	-1.2
I3W	0.94	2.8	-8
I3Y	0.18	2.3	-5.9
E11A	-0.83	-1.1	2.6
E11F	-2.4	-1.7	4.3
E11M	-2.6	-1.6	4.1
T115E	0.29	-0.3	0.7
S117I	-1.5	-1.7	4.2
S117V	-1.6	-2	5.1
R119E	-0.13	0.04	-0.1
R119M	-1.6	-0.1	0.3
N132F	-0.90	-1.3	3.3
N132I	-0.92	-1.2	3
N132M	-1.0	-1.5	3.6
L133A	1.1	3.6	-10.55
L133F	-0.25	0.3	-0.8
L133G	1.5	3.1	-11.7

Table S2.6. Contact arrays for MC-I mutant sites used in **Figure 2.6**. Each θ_m set residue (*i*) was limited to its six nearest side chain contacts (*j*). Residues with less than six contacts were assigned a solvent neighbor, **W**, with μ=110.

Site 3						
$\theta_m(i)$	1	2	3	7	11	13
Contacts (*j*)	158	4	7	145	145	8
	5	5	100	101	104	29
	161	3	4	104	29	15
	6	W	97	67	7	63
	9	W	6	3	105	60
	W	W	71	71	30	28

Site 11										
$\theta_m(i)$	1	2	3	7	11	13	15	17	19	21
Contacts (*j*)	158	4	7	145	145	8	13	25	25	142
	5	5	100	101	104	29	63	43	23	22
	161	3	4	104	29	15	58	33	W	20
	6	W	97	67	7	63	60	39	W	W
	9	W	6	3	105	60	57	27	W	W
	W	W	71	71	30	W	W	56	W	W

Site 115						
$\theta_m(i)$	107	111	115	119	123	124
Contacts (*j*)	108	84	116	123	125	126
	110	102	83	125	119	90
	W	103	W	W	120	91
	W	106	W	W	W	W
	W	99	W	W	W	W
	W	118	W	W	W	W

Table S2.6. Contact arrays for MC-I mutant sites used in **Figure 6**. *Continued.*

Site 117							
$\theta_m(i)$	107	109	113	117	121	124	125
Contacts (*j*)	108	108	W	114	87	126	120
	110	W	W	132	91	90	123
	W	W	W	133	126	91	119
	W	W	W	W	129	W	128
	W	W	W	W	153	W	W
	W	W	W	W	118	W	W

Site 119						
$\theta_m(i)$	111	115	119	123	124	127
Contacts (*j*)	84	116	123	125	126	126
	102	83	125	119	90	154
	103	W	W	120	91	W
	106	W	W	W	W	W
	99	W	W	W	W	W
	118	W	W	W	W	W

Site 132							
$\theta_m(i)$	124	128	132	134	135	136	140
Contacts (*j*)	126	120	117	139	131	114	141
	90	125	120	150	132	138	139
	91	W	116	W	W	139	W
	W	W	135	W	W	W	W
	W	W	W	W	W	W	W
	W	W	W	W	W	W	W

Site 133									
$\theta_m(i)$	124	125	129	133	134	135	136	137	141
Contacts (*j*)	126	120	120	117	139	131	114	141	22
	90	123	121	138	150	132	138	22	137
	91	119	133	150	W	W	139	W	142
	W	128	W	153	W	W	W	W	140
	W	W	W	102	W	W	W	W	W
	W	W	W	114	W	W	W	W	W

References

1. Bowman, G. R.; Voelz, V. A.; Pande, V. S., Taming the complexity of protein folding. *Current Opinion in Structural Biology* **2011,** *21* (1), 4-11.
2. Brenner, S., Life's code script. *Nature* **2012,** *482*, 461.
3. Carlson, R., Estimating the biotech sector's contribution to the US economy. *Nature biotechnology* **2016,** *34* (3), 247.
4. Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S., Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In *Protein Crystallography: Methods and Protocols*, Wlodawer, A.; Dauter, Z.; Jaskolski, M., Eds. Springer New York: New York, NY, 2017; pp 627-641.
5. Baldwin, R. L., Energetics of Protein Folding. *Journal of Molecular Biology* **2007,** *371* (2), 283-301.
6. Tanford, C., How protein chemists learned about the hydrophobic factor. *Protein Science* **1997,** *6* (6), 1358-1366.
7. Fersht, A., *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan: 1999.
8. Richards, F. M., Areas, Volumes, Packing, and Protein Structure. *Annual Review of Biophysics and Bioengineering* **1977,** *6* (1), 151-176.
9. Anfinsen, C. B., Principles that Govern the Folding of Protein Chains. *Science* **1973,** *181* (4096), 223-230.
10. Tanford, C.; Reynolds, J., *Nature's robots: a history of proteins*. OUP Oxford: 2003.
11. Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. A., Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2016,** *2* (11).
12. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W., Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010,** *330* (6002), 341-346.
13. MacCallum, J. L.; Perez, A.; Dill, K. A., Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences* **2015,** *112* (22), 6985-6990.
14. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015,** *137* (7), 2695-2703.
15. Lenselink, E. B.; Louvel, J.; Forti, A. F.; van Veldhoven, J. P. D.; de Vries, H.; Mulder-Krieger, T.; McRobb, F. M.; Negri, A.; Goose, J.; Abel, R.; van Vlijmen, H. W. T.; Wang, L.; Harder, E.; Sherman, W.; Ijzerman, A. P.; Beuming, T., Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. *ACS Omega* **2016,** *1* (2), 293-304.
16. Ford, M. C.; Babaoglu, K., Examining the Feasibility of Using Free Energy Perturbation (FEP+) in Predicting Protein Stability. *Journal of Chemical Information and Modeling* **2017,** *57* (6), 1276-1285.
17. Piana, S.; Klepeis, J. L.; Shaw, D. E., Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology* **2014,** *24*, 98-105.

18. Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K., Challenges in protein-folding simulations. *Nature Physics* **2010,** *6*, 751.
19. Skinner, J. J.; Yu, W.; Gichana, E. K.; Baxa, M. C.; Hinshaw, J. R.; Freed, K. F.; Sosnick, T. R., Benchmarking all-atom simulations using hydrogen exchange. *Proceedings of the National Academy of Sciences* **2014,** *111* (45), 15975-15980.
20. Gu, J.; Hilser, V. J., Predicting the Energetics of Conformational Fluctuations in Proteins from Sequence: A Strategy for Profiling the Proteome. *Structure (London, England : 1993)* **2008,** *16* (11), 1627-1637.
21. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450*, 964.
22. Zwier, M. C.; Chong, L. T., Reaching biological timescales with all-atom molecular dynamics simulations. *Current Opinion in Pharmacology* **2010,** *10* (6), 745-752.
23. Lesne, A.; Laguës, M., *Scale invariance: From phase transitions to turbulence*. Springer Science & Business Media: 2011.
24. Phillips, J. C., Scaling and self-organized criticality in proteins I. *Proceedings of the National Academy of Sciences* **2009,** *106* (9), 3107-3112.
25. Phillips, J. C., Scaling and self-organized criticality in proteins II. *Proceedings of the National Academy of Sciences* **2009,** *106* (9), 3113-3118.
26. Reuveni, S.; Granek, R.; Klafter, J., Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure. *Proceedings of the National Academy of Sciences* **2010,** *107* (31), 13696-13700.
27. Banerji, A.; Ghosh, I., Fractal symmetry of protein interior: what have we learned? *Cellular and Molecular Life Sciences* **2011,** *68* (16), 2711-2737.
28. Koga, N.; Tatsumi-Koga, R.; Liu, G.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D., Principles for designing ideal protein structures. *Nature* **2012,** *491*, 222.
29. Lehmann, M.; Pasamontes, L.; Lassen, S. F.; Wyss, M., The consensus concept for thermostability engineering of proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **2000,** *1543* (2), 408-415.
30. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M., AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* **2008,** *36* (Database issue), D202-D205.
31. Moret, M. A.; Zebende, G. F., Amino acid hydrophobicity and accessible surface area. *Physical Review E* **2007,** *75* (1), 011920.
32. Creamer, T. P.; Srinivasan, R.; Rose, G. D., Modeling unfolded states of peptides and proteins. *Biochemistry* **1995,** *34* (50), 16245-16250.
33. Creamer, T. P.; Srinivasan, R.; Rose, G. D., Modeling Unfolded States of Proteins and Peptides. II. Backbone Solvent Accessibility. *Biochemistry* **1997,** *36* (10), 2832-2835.
34. Ramachandran, G. N.; Sasisekharan, V., Conformation of Polypeptides and Proteins**The literature survey for this review was completed in September 1967, with the journals which were then available in Madras and the preprinta which the authors had received.††By the authors' request, the publishers have left certain matters of usage and spelling in the form in which they wrote them. In *Advances in Protein Chemistry*, Anfinsen, C. B.; Anson, M. L.; Edsall, J. T.; Richards, F. M., Eds. Academic Press: 1968; Vol. 23, pp 283-437.
35. Ramakrishnan, C.; Ramachandran, G. N., Stereochemical Criteria for Polypeptide and Protein Chain Conformations: II. Allowed Conformations for a Pair of Peptide Units. *Biophysical Journal* **1965,** *5* (6), 909-933.
36. Hollingsworth, S. A.; Karplus, P. A., A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomolecular concepts* **2010,** *1* (3-4), 271-283.

37. Zhang, T.; Faraggi, E.; Zhou, Y., Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins* **2010,** *78* (16), 3353-3362.
38. DuBay, K. H.; Geissler, P. L., Calculation of Proteins' Total Side-Chain Torsional Entropy and Its Influence on Protein–Ligand Interactions. *Journal of Molecular Biology* **2009,** *391* (2), 484-497.
39. Englander, S. W.; Mayne, L.; Kan, Z.-Y.; Hu, W., Protein Folding—How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry. *Annual Review of Biophysics* **2016,** *45* (1), 135-152.
40. Best, R. B.; Hummer, G.; Eaton, W. A., Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences* **2013,** *110* (44), 17874-17879.
41. Meirovitch, H.; Cheluvaraja, S.; White, R. P., Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding. *Current protein & peptide science* **2009,** *10* (3), 229-243.
42. De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A., Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry* **2016,** *59* (9), 4035-4061.
43. Sanchez-Ruiz, J. M., Protein kinetic stability. *Biophys Chem* **2010,** *148* (1), 1-15.
44. Karshikoff, A.; Nilsson, L.; Ladenstein, R., Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS Journal* **2015,** *282* (20), 3899-3917.
45. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.
46. Pohorille, A.; Jarzynski, C.; Chipot, C., Good Practices in Free-Energy Calculations. *The Journal of Physical Chemistry B* **2010,** *114* (32), 10235-10253.
47. Nascimento, É. C. M.; Oliva, M.; Świderek, K.; Martins, J. B. L.; Andrés, J., Binding Analysis of Some Classical Acetylcholinesterase Inhibitors: Insights for a Rational Design Using Free Energy Perturbation Method Calculations with QM/MM MD Simulations. *Journal of Chemical Information and Modeling* **2017,** *57* (4), 958-976.
48. Bryan, P. N.; Orban, J., Proteins that switch folds. *Current Opinion in Structural Biology* **2010,** *20* (4), 482-488.
49. Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N., A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences* **2009,** *106* (50), 21149-21154.
50. Matthews, B. W., Studies on Protein Stability With T4 Lysozyme. In *Advances in Protein Chemistry*, Anfinsen, C. B.; Richards, F. M.; Edsall, J. T.; Eisenberg, D. S., Eds. Academic Press: 1995; Vol. 46, pp 249-278.
51. Xu, J.; Baase, W. A.; Baldwin, E.; Matthews, B. W., The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Science* **1998,** *7* (1), 158-177.
52. Rennell, D.; Bouvier, S. E.; Hardy, L. W.; Poteete, A. R., Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology* **1991,** *222* (1), 67-88.
53. Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W., A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences* **1995,** *92* (2), 452-456.
54. Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W., Lessons from the lysozyme of phage T4. *Protein Science* **2010,** *19* (4), 631-641.
55. Mooers, B. H. M.; Tronrud, D. E.; Matthews, B. W., Evaluation at atomic resolution of the role of strain in destabilizing the temperature-sensitive T4 lysozyme mutant Arg 96 → His. *Protein science : a publication of the Protein Society* **2009,** *18* (5), 863-870.

56. Kyte, J.; Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **1982,** *157* (1), 105-132.
57. McHaourab, H. S.; Oh, K. J.; Fang, C. J.; Hubbell, W. L., Conformation of T4 Lysozyme in Solution. Hinge-Bending Motion and the Substrate-Induced Conformational Transition Studied by Site-Directed Spin Labeling. *Biochemistry* **1997,** *36* (2), 307-316.
58. Kuroki, R.; Weaver, L. H.; Matthews, B. W., Structural basis of the conversion of T4 lysozyme into a transglycosidase by reengineering the active site. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (16), 8949-8954.
59. Baxa, M. C.; Haddadian, E. J.; Jumper, J. M.; Freed, K. F.; Sosnick, T. R., Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proceedings of the National Academy of Sciences* **2014,** *111* (43), 15396-15401.
60. Mark, A. E.; van Gunsteren, W. F., Decomposition of the Free Energy of a System in Terms of Specific Interactions: Implications for Theoretical and Experimental Studies. *Journal of Molecular Biology* **1994,** *240* (2), 167-176.
61. Dill, K. A., Additivity Principles in Biochemistry. *Journal of Biological Chemistry* **1997,** *272* (2), 701-704.
62. Geschwindner, S.; Ulander, J.; Johansson, P., Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip? *Journal of Medicinal Chemistry* **2015,** *58* (16), 6321-6335.
63. Clark, A. J.; Gindin, T.; Zhang, B.; Wang, L.; Abel, R.; Murret, C. S.; Xu, F.; Bao, A.; Lu, N. J.; Zhou, T.; Kwong, P. D.; Shapiro, L.; Honig, B.; Friesner, R. A., Free Energy Perturbation Calculation of Relative Binding Free Energy between Broadly Neutralizing Antibodies and the gp120 Glycoprotein of HIV-1. *Journal of Molecular Biology* **2017,** *429* (7), 930-947.
64. Roy, A.; Perez, A.; Dill, Ken A.; MacCallum, Justin L., Computing the Relative Stabilities and the Per-Residue Components in Protein Conformational Changes. *Structure* **2014,** *22* (1), 168-175.
65. Yu, W.; Baxa, M. C.; Gagnon, I.; Freed, K. F.; Sosnick, T. R., Cooperative folding near the downhill limit determined with amino acid resolution by hydrogen exchange. *Proceedings of the National Academy of Sciences of the United States of America* **2016,** *113* (17), 4747-4752.
66. Murphy, R. B.; Repasky, M. P.; Greenwood, J. R.; Tubert-Brohman, I.; Jerome, S.; Annabhimoju, R.; Boyles, N. A.; Schmitz, C. D.; Abel, R.; Farid, R.; Friesner, R. A., WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand–Receptor Docking. *Journal of Medicinal Chemistry* **2016,** *59* (9), 4364-4384.
67. Woods, C. J.; Malaisree, M.; Michel, J.; Long, B.; McIntosh-Smith, S.; Mulholland, A. J., Rapid decomposition and visualisation of protein-ligand binding free energies by residue and by water. *Faraday Discussions* **2014,** *169* (0), 477-499.
68. Guimarães, C. R. W.; Mathiowetz, A. M., Addressing Limitations with the MM-GB/SA Scoring Procedure using the WaterMap Method and Free Energy Perturbation Calculations. *Journal of Chemical Information and Modeling* **2010,** *50* (4), 547-559.
69. Flory, P. J.; Garrett, R. R., Phase Transitions in Collagen and Gelatin Systems1. *Journal of the American Chemical Society* **1958,** *80* (18), 4836-4845.
70. Tran, H. T.; Pappu, R. V., Toward an Accurate Theoretical Framework for Describing Ensembles for Proteins under Strongly Denaturing Conditions. *Biophysical Journal* **2006,** *91* (5), 1868-1886.
71. Zimm, B. H.; Bragg, J. K., Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *The Journal of Chemical Physics* **1959,** *31* (2), 526-535.
72. Lifson, S.; Roig, A., On the Theory of Helix—Coil Transition in Polypeptides. *The Journal of Chemical Physics* **1961,** *34* (6), 1963-1974.

73. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics* **1999,** *35* (2), 133-152.
74. Selinger, J. V., *Introduction to the theory of soft matter: from ideal gases to liquid crystals*. Springer: 2015.
75. Lazaridis, T.; Karplus, M., Thermodynamics of protein folding: a microscopic view. *Biophys Chem* **2002,** *100* (1), 367-395.
76. Murza, A.; Kubelka, J., Beyond the nearest-neighbor Zimm–Bragg model for helix-coil transition in peptides. *Biopolymers* **2009,** *91* (2), 120-131.
77. Lum, K.; Chandler, D.; Weeks, J. D., Hydrophobicity at Small and Large Length Scales. *The Journal of Physical Chemistry B* **1999,** *103* (22), 4570-4577.
78. Berne, B. J.; Weeks, J. D.; Zhou, R., Dewetting and Hydrophobic Interaction in Physical and Biological Systems. *Annual review of physical chemistry* **2009,** *60*, 85-103.
79. Chandler, D., Interfaces and the driving force of hydrophobic assembly. *Nature* **2005,** *437*, 640.
80. Senet, P.; Maisuradze, G. G.; Foulie, C.; Delarue, P.; Scheraga, H. A., How main-chains of proteins explore the free-energy landscape in native states. *Proceedings of the National Academy of Sciences* **2008,** *105* (50), 19708-19713.
81. Neusius, T.; Sokolov, I. M.; Smith, J. C., Subdiffusion in time-averaged, confined random walks. *Physical Review E* **2009,** *80* (1), 011109.
82. Meroz, Y.; Ovchinnikov, V.; Karplus, M., Coexisting origins of subdiffusion in internal dynamics of proteins. *Physical Review E* **2017,** *95* (6), 062403.
83. Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L., Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angewandte Chemie (International Ed. in English)* **2016,** *55* (26), 7364-7368.
84. Rooklin, D.; Modell, A. E.; Li, H.; Berdan, V.; Arora, P. S.; Zhang, Y., Targeting Unoccupied Surfaces on Protein–Protein Interfaces. *Journal of the American Chemical Society* **2017,** *139* (44), 15560-15563.
85. Sivaprakasam, P.; Han, X.; Civiello, R. L.; Jacutin-Porte, S.; Kish, K.; Pokross, M.; Lewis, H. A.; Ahmed, N.; Szapiel, N.; Newitt, J. A.; Baldwin, E. T.; Xiao, H.; Krause, C. M.; Park, H.; Nophsker, M.; Lippy, J. S.; Burton, C. R.; Langley, D. R.; Macor, J. E.; Dubowchik, G. M., Discovery of new acylaminopyridines as GSK-3 inhibitors by a structure guided in-depth exploration of chemical space around a pyrrolopyridinone core. *Bioorganic & Medicinal Chemistry Letters* **2015,** *25* (9), 1856-1863.
86. Gumbart, J. C.; Roux, B.; Chipot, C., Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles. *Journal of Chemical Theory and Computation* **2013,** *9* (8), 3789-3798.
87. Duan, L. L.; Zhu, T.; Li, Y. C.; Zhang, Q. G.; Zhang, J. Z. H., Effect of polarization on HIV-1protease and fluoro-substituted inhibitors binding energies by large scale molecular dynamics simulations. *Scientific Reports* **2017,** *7*, 42223.

Chapter 3: Using the Hidden Symmetry Model to Predict Peptide-Antibody Affinities and Design Novel Peptide Antigens

Note: This is a draft manuscript (tentatively titled "HSyM-Guided Engineering of the Immunodominant p53 Transactivation Domain Putative Peptide Antigen for Improved Binding to its Anti-p53 Monoclonal Antibody") that we intend to submit for peer reviewed publication in April 2021.

Zachary Fritz[1], Rene Schloss[1], Martin Yarmush[1], Lawrence Williams[2]

[1]Department of Biomedical Engineering, Rutgers, The State University of New Jersey

[2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey

Contribution

As lead author, I (Zachary Fritz) planned and conducted all experiments, analyzed the data, made all the figures, and wrote the first manuscript draft.

Introduction

We applied the Hidden Symmetry model (HSyM) in a novel engineering strategy to enhance the binding of a specific peptide antigen to its specific monoclonal antibody (mAb). The *p53* gene is the most commonly mutated in cancer. Many mutations compromise the critically important regulatory and tumor suppression roles that the p53 protein serves. Accumulation and overexpression of the mutant protein appears to lead to autoimmune targeting of p53. Clinical data show that detection of anti-p53 and related autoantibodies in human patients represents a modality of cancer detection that can precede other methods of diagnosis by approximately 2-5 years[1, 2]. Typically these peptide segments that correspond to the antigenic portions of the protein are taken as is and used to analyze serum samples for the autoantibodies. The detection of autoantibodies directed against fragments of this and other key proteins represents a potentially powerful strategy for the early detection and characterization of cancer and other disease states [2, 3].

We focused on the decameric segment of residues 46-55 from the p53 transactivation domain. This decamer is the immunodominant antigen and putative epitope of mAb Ab28. Several peptide constructs were designed to bind similarly or more tightly than the parent peptide. Each contained the full decapeptide epitope and differed only in the flanking regions. Binding enhancements of over 500-fold were observed by surface plasmon resonance measurements and confirmed in immunoassay simulations with immobilized peptide constructs and mAb in BSA, PBS solutions. Moreover, guided by our model, comparison of computed and experimental

data show excellent correlation between the predicted and observed binding ($R^2 > 0.75$) and led us to reevaluate the putative epitope and to suggest that the epitope may well be the undecapeptide region of residues 45-55.

Even though synthetic peptide antigens are the lynchpin to autoantibody detection strategies, there are many challenges to cancer detection by way of peptide-based autoantibody detection. Most of these challenges are traceable to the diversity of the patient population immune response, including the variability in immunogenicity of different regions of a given protein antigen and the specific antigenic regions of the protein antigen. Much progress has been made in characterizing the autoantibody targets, selecting representative peptide antigens, and combining peptide antigens into panels to be used to identify disease. Since the affinity of the detecting peptide epitopes for the polyclonal response poses a serious challenge, we focused on the problem of improving affinity of an antibody to a selected antigenic peptide.

Model-Guided Peptide Engineering

Our engineering strategy is based primarily on a method that we recently reported to estimate and rationalize per-residue interaction free energy contributions in proteins, and we sought to extend it to polypeptides and to engineer peptide-protein binding [4]. This coarse-grain model assumes small dynamical fluctuations and describes the interaction energy of each residue (μ) as a simple function, $\mu(\gamma, \sigma, \phi)$. The molecular architecture of each amino acid residues is approximated as a blobs of a specific intrinsic interaction free energy (γ). Protein backbone conformation is coarsely approximated as one of four structural motifs and is dependent on whether a residue (i) makes a contact with the first, second, third, or fourth residue in the sequence (σ). The effective contribution to binding free energy is determined by averaging the intrinsic contributions of a specific subset of residues along the peptide backbone within a small

window (ϕ). The effective free energy contribution of each amino acid residue (μ) is therefore a function of γ, σ, and ϕ, and the interaction free energy between residues *i* and *j* is approximated as the product of their interaction factors ($G \sim -\mu_i\mu_j$), with two residues of $\mu = 200$ interacting constituting a free energy gain of approximately -0.5 kcal/mol. The free energy of binding of a ligand and its receptor, for example, would be approximated as the sum of these interactions, i.e., $G \sim -\Sigma\mu_i\mu_j$. The elaborated equations are detailed in ref [4]. The model allows fast (< 1 sec), single-state calculations of proteins, protein-protein, and protein-peptide complexes with minimal computing capability. We benchmarked HSyM by accurately calculated the thermostability (ΔΔG) of 28 mutations of T4 lysozyme ($R^2 > 0.7$) [4].

The design rationale is summarized in **Figure 3.1.** Since no structural information regarding the antibody or the antibody-antigen structure is available, all peptides in the epitope region were assigned $\sigma = 4$. This assignment corresponds to that calculated for residues 46-55 in the WT p53 transactivation domain, which adopts an approximate alpha helix conformation in this region (PDB ID: 2L14) [5]. It is noteworthy that within the framework of HSyM it is possible for both a series of loops and/or turns, as well as an extended alpha helix, to be assigned $\sigma = 4$, though the network of intra-peptide contacts would be very different for each arrangement. Regardless of these details, for the purposes of this study, all peptides were assumed to bind with a peptide conformational assignment that matched the parent protein structure.

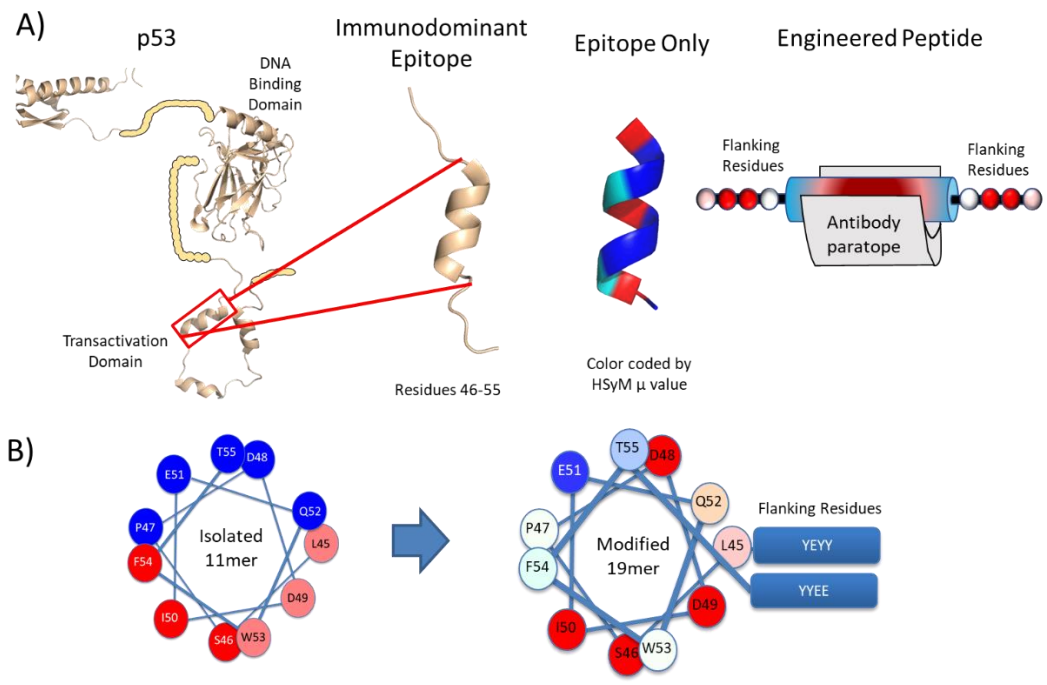


Figure 3.1: Peptide Design Methodology. **(A)** A 10mer linear epitope (residues 46-55) from the Transactivation Domain of human p53 (PDB: 2L14) was input into the model in both whole protein and isolated peptide forms to obtain per-residue interaction energy contribution (μ) values. Color codes for each residue are based on the computed μ values. Colors range from dark blue to white to dark red. Blue is below the average, which is white, and red is above the average. The darker the color the further the value for that residue is from the average. Flanking residues were added to the recapitulate or increase the whole protein μ profile to induce tighter antibody binding. **(B)** An example of modifying the later expanded 11mer epitope with flanking residues to alter its μ profile. The peptides are represented as α-helical wheels throughout this manuscript, due to the secondary structure of the 2L14 structure.

The μ values calculated for the transactivation domain that corresponds to the 46-55 epitope compared to those calculated for the isolated 46-55 decapeptide are shown in **Figure 3.2(A)**. The μ values of the segment in the WT protein are similar, though the μ values of the *N*-terminal

region are relatively high and the central isoleucine (I50) of the WT epitope has the highest value of this portion of the protein. The μ values for the isolated decapeptide are more polarized, with most residues having μ values lower than the WT epitope and with S46, I50, and F54 having μ values higher than the WT epitope **Figure 3.2(B)**. The higher value residues would be expected to play a more important role in binding to the antibody paratope, and yet our calculations also indicate that the decapeptide has a lower average μ value profile compared to the WT epitope profile. These data suggested to us that careful design of additional flanking residues could increase the μ value profile of the decapeptide, even though these flanking regions are not part of the epitope. This highlights the contextual importance of the native whole protein sequence and structure in determining residue interaction energies, and consequently, antibody binding affinity. In contrast, adding one residue, Leu45, to the conserved epitope (**Figure 3.2(C)**) significantly increases the μ values of 2 additional residues in the peptide.

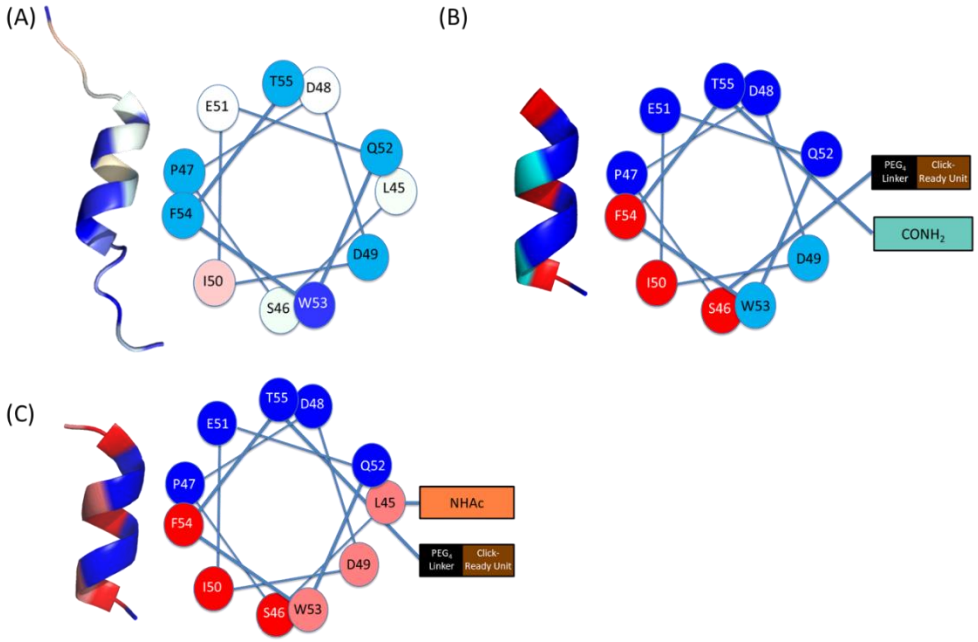


Figure 3.2: Epitope Variations and μ Profiles. The epitope sequence (45-55) is represented as a helical wheel, with color-coding based on μ value (see **Figure 3.1**). The PEG linker and Click

Ready Region (black and brown boxes) of the peptide constructs are used for microbead functionalization (see Materials and Methods). **(A)** The μ profile for the wildtype, whole p53 protein epitope. **(B)** The μ profile for the 10mer peptide construct, which lacks the L45 position. **(C)** The μ profile for the conserved 11mer epitope peptide construct.

Within the framework of our model there is a theoretical maximum that a residue can contribute to binding free energy (i.e., a maximum μ value). Of course, flanking regions dominated by non-polar residues would render the construct prone to non-specific binding, aggregation, and low solubility. Our modeling indicated that the 46-55 decapeptide segment is a good candidate for engineered binding affinity. By adding flanking segments that would not be dominated by non-polar residues, our design aimed to tune the μ values of the decapeptide epitope to be near or slightly above that of the parent WT p53 transactivation domain, with the expectation that the binding affinity would be near or above the WT decapeptide. Accordingly, we prepared the immunodominant decapeptide and 18 various polypeptides that have up to four flanking residues on either or both N and C termini. The flanking sequences were either permutations of the WT p53 flanking residues (42-45, 56-59) or non-native residues. As shown below, we came to wonder whether the true epitope was the undecapeptide, residues 45-55 in the Transactivation domain, and prepared an additional five peptides that contained the undecapeptide epitope. These 24 peptide constructs are shown in **Table 3.1**.

As shown in **Figure 3.3**, each residue of the epitope was assigned four or five contacts to the antibody, since helical peptide:Ab complexes tend to form four or five contacts per-residue (c.f. PDBIDs: 1MVU, 4HPO, 2AP2) [6, 7]. Four antibody contacts were assigned to those residues that were calculated to have μ values lower than the average of the set for the WT whole protein epitope. Five antibody contacts were assigned for those calculated μ values greater than the

average. It seems unlikely that all residues in a helix would make contacts with the antibody paratope. Although this approach could add noise to our estimated binding energies, and weaken the correlation of predicted and observed binding affinity, it avoids the risk of neglecting key residues.

No.	Sequence	Average K_D [M]	Affinity Improvement Relative to 10mer	Experimental ΔG (kcal/mol)	HSyM Calculated ΔG (kcal/mol)
1	SPDDIEQWFT	3.27E-06	-	-7.48	-10.08
2	"..."-EDPG	2.10E-06	2	-7.75	-9.79
3	"..."-PGED	3.52E-06	1	-7.44	-9.86
4	"..."-RNLL	1.84E-06	2	-7.83	-10.28
5	**"..."**-LNRL	1.41E-06	2	-7.98	-10.41
6	"..."-LLNR	7.62E-07	4	-8.35	-10.47
7	"..."-KNFV	1.69E-06	2	-7.87	-10.47
8	"..."-NVFK	6.93E-07	5	-8.4	-10.6
9	"..."-VFNK	7.40E-07	4	-8.36	-10.74
10	VFNK-"..."	6.80E-08	48	-9.78	-11.44
11	KFVN-"..."	7.20E-08	45	-9.74	-11.68
12	NKVF-"..."	1.60E-08	204	-10.63	-12.06
13	LMLD-"..."	3.53E-07	9	-8.8	-11.84
14	LLDM-"..."	2.65E-08	124	-10.34	-12.16
15	DLML-"..."	9.08E-09	360	-10.97	-12.2
16	DLML-"..."-EDPG	1.39E-08	235	-10.72	-11.75
17	KFVN-"..."-ANRA	1.01E-07	33	-9.55	-11.49
18	ARNA-"..."-NVFK	7.01E-08	47	-9.76	-11.57
19	RFKV-"..."-AANN	5.25E-08	62	-9.93	-11.77
20	LSPDDIEQWFT	6.27E-09	522	-11.19	-11.65
21	KKNN-LSPDDIEQWFT	8.69E-09	376	-11	-10.94
22	DDLM-LSPDDIEQWFT	6.99E-09	468	-11.13	-11.95
23	YYEY-LSPDDIEQWFT	7.68E-09	426	-11.07	-12.64
24	YYEY-LSPDDIEQWFT-YYEE	1.27E-08	258	-10.77	-12.76

Table 3.1: Peptide constructs and associated HSyM and experimental (SPR) antibody affinity values. Peptides 2-19 were based off of Peptide 1, with its sequence abbreviated ("…") for these entries. The "HSyM Calculated ΔG" values are based on the expanded 11mer epitope.

The μ_j values assigned to the hypothetical paratope were based on the epitope residues (μ_i) in the WT protein. The assignment was based on averaging the proximal μ values calculated in the protein context according to the equation: $\mu_j = (1/7)[\Sigma(\mu_{i-7} + \mu_{i-4} + \mu_{i-3} + \mu_i + \mu_{i+3} + \mu_{i+4} + \mu_{i+7})]$. Mirroring the μ values in this way places emphasis on the WT antigen and allows an approximation of the interaction energy in terms of $G \sim -\mu_i\mu_j$, for each residue of the epitope, *i*, and paratope contact *j*. This rationale is consistent with the view that Ab binding is a balance of maximizing binding affinity and selectivity. Since we alter only the non-epitope region, the binding affinity, ΔG_{int}, was approximated using the following equation:

$$\Delta G_{Binding,Calc} = -\lambda \sum_{i,j=1}^{n} \mu_i \mu_{Ab,j}$$

where μ_i refers to the interaction factors of the peptide epitope residues, $\mu_{Ab,j}$ is the hypothetical antibody μ values, *n* is the length of the epitope, and λ is a scaling factor, previously determined to be 1.25E-5 kcal/mol [4]. Entropy contributions to binding to the constant region of the epitope are take to be approximately constant and are ignored. These model calculated binding energies were fitted against energies that were derived from the experimental SPR data using the equation:

$$\Delta G_{Binding,Exp} = RTln(K_D)$$

where R is the ideal gas constant, T is the temperature (298 K), and K_D is the experimentally determined disassociation constant.

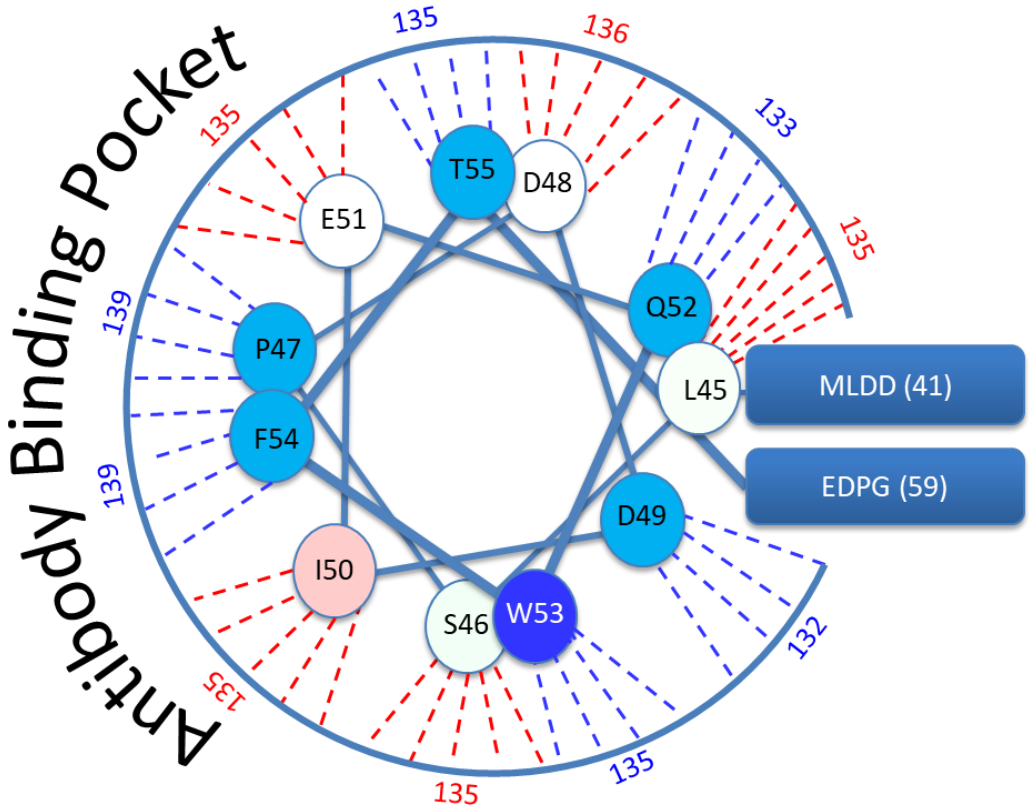


Figure 3.3: Proposed Antibody-Peptide Binding Scheme Used in Affinity Calculations. The hypothetical antibody μ values are shown on the outside of the semicircle and were determined by averaging the μ values of the WT full protein epitope's (shown in helical wheel format) helical neighbors. Residues were assigned antibody contacts based on the WT epitope's μ values, with residues with higher-than-average μ values (in white, off-white, and salmon) assigned 5 and the rest assigned 4.

Materials and Methods

Peptide Constructs

Soluble peptide constructs were prepared and analyzed by SPR. These constructs were also immobilized on beads for limit of detection studies (**Table 3.2**). *Soluble Peptide Constructs:* **Figure 3.4** depicts the construct design. The initial 20 peptides contained a 10mer conserved epitope corresponding to residues 46-55 of wildtype human p53, the epitope specified for the monoclonal antibody used in all experiments (Abcam, ab28); the later 5 peptides would have an expanded 11mer epitope (residues 45-55). Peptide constructs contained a PEG_4 spacer and a terminal ε-azido lysine and were synthesized by Genscript (Piscataway, NJ; >75% purity). With the exception of the simple 10-mer epitope peptide (Peptide 1) and the three 14-mers, Peptides 10-12, which had the PEG_4 and an acetylated azido-lysine on the N-terminus, all peptides were *N*-acetylated, had a *C*-terminus tethered with PEG_4 linker terminating in the Click-ready azido-lysine unit. Soluble peptide constructs were used as is for SPR analysis. *Immobilized Peptide Constructs:* Peptides were conjugated at 20 μM to 6.5 μm carboxylated magnetic microbeads (Luminex) that had been amidated with dibenzocyclooctyne (DBCO)-PEG_4-amine linker. Click coupling of the engineered soluble peptide construct to the functionalized bead gave the immobilized peptide construct ready for immunoassay analysis.

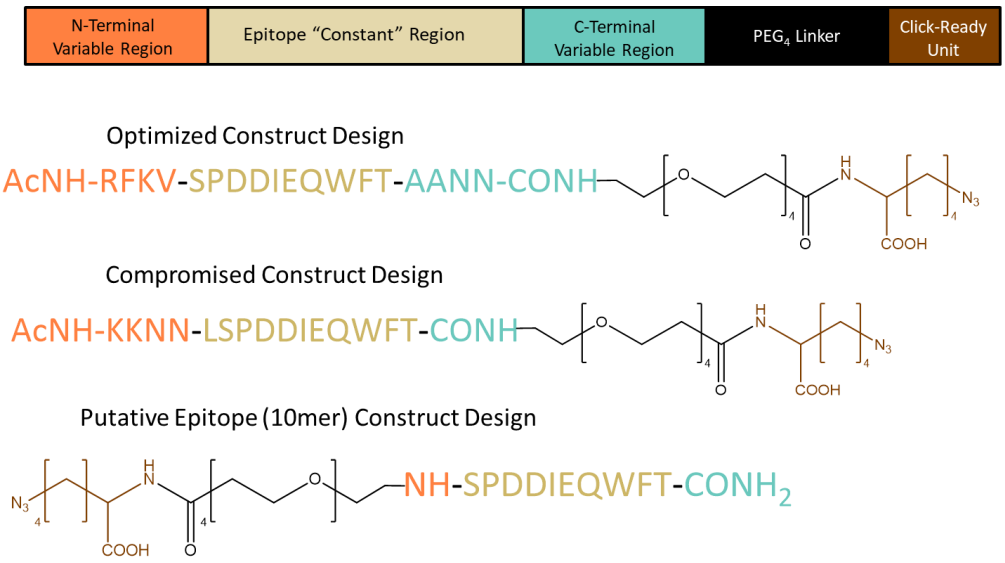


Figure 3.4: Peptide Construct Design. All but Peptides 1, 10, 11, and 12 followed the construct design shown in the top box; for these exception peptides, the PEG linker and Click Ready Unit were located on the N-terminus instead of the C.

SPR and LOD Assays

Surface plasmon resonance (SPR) experiments were conducted with a Biacore T200 system (Cytiva) at a working temperature of 25°C. Monoclonal anti-p53 antibody (ab28, Abcam) specific to the putative conserved epitope was conjugated to a CM5 sensor chip at a 2500 RU target response level with amine coupling. Peptides were prepared at varying concentrations in a running buffer of 1.02x PBS with 2% DMSO and 0.05% Tween 20; 5 M NaCl was used as the regeneration buffer. Three replicate measurements were taken for each peptide.

The Biacore's evaluation software was used to determine each peptide's disassociation constant (K_D) using either kinetic 1:1 binding model fitting to determine k_d and k_a or steady state curve fitting (fitting the SPR response at equilibrium vs the peptide concentration). Multiple best fit K_D measurements for each peptide were averaged to obtain the final affinity data. The highest affinity peptides required kinetic model curve fitting of their sensograms. This is likely due to peptide aggregation. See **Supplementary Information** or sensograms and fitting curves for each peptide. The highest concentration sensograms were removed from the analysis to improve the

fit (see **Supplementary Information**). In some cases, an imperfect fit of a 1:1 binding model was observed. Limit of detection (LOD) assays were performed as follows: Anti-p53 monoclonal antibody (ab28, Abcam) standards were prepared at concentrations ranging from 0.1 to 100 ng/ml in a 1:3 dilution of fetal bovine sera and PBS and plated in triplicate with the peptide-functionalized microbeads. A BioRad multiplex immunoassay kit and a biotinylated anti-mouse IgG detection antibody (Abcam) were used to perform the assay, and a Bio-Plex 200 system was used for fluorescent signal reading. The limit of detection (LOD) of each peptide was determined by multiplying the standard deviation of the blank's signal by three and dividing by the slope of the standard curve, with 3 slope values determined and averaged.

Clinical Sera Assays

Ten clinical sera samples obtained from our collaborators—consisting of 5 samples from patients with colorectal cancer and 5 from healthy controls—were diluted 100x in PBS and tested via Bio-Plex assays to assess the diagnostic capability of the peptide antigens to detect cancer autoantibodies. Normal pooled human sera (Innovative Research, Novi, MI) was also diluted 100x in PBS and used as a negative control. A positive result was defined for each peptide as a fluorescent signal greater than the negative control's signal plus three times its standard deviation. Results for these assays are found in the **Supplementary Information.**

Results and Discussion

As shown in **Table 3.1**, the initial set of peptides designed, peptides 1-19, range from slightly lower to much greater binding free energy than the putative decapeptide epitope. A 20th initially designed peptide, a C-terminal modified 14mer (C-flanking sequence: PEDG) proved to have very high nonspecific binding in the SPR measurements thus making it impossible to obtain an accurate affinity value.

Figure 3.5 (C) shows the correlation of calculated and experimental binding energy for those peptides with only C-terminal flanking regions is good (R^2 = 0.72). Interestingly, even though the design proved effective at generating peptides with much improved binding energy, there was no correlation between the calculated and experimental binding energy for those peptides with only N-terminal flanking regions (R^2 = 0.02). The combined fit for the entire set was poor (R^2 = 0.24). Pleasingly, several peptides were designed to have superior binding, as observed, and several had over 100-fold improvement over the decapeptide alone.

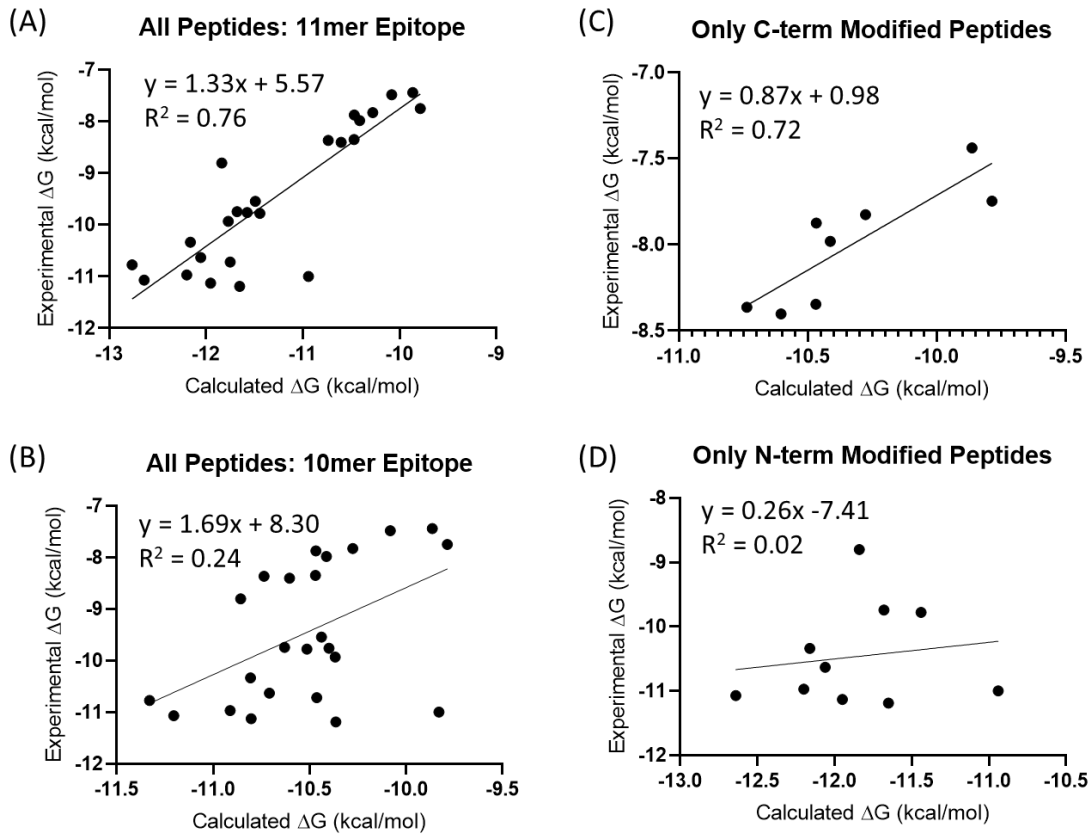


Figure 3.5: Comparison of Model Calculated and Experimental Peptide Affinities. **(A)** Linear regression fit for all 24 peptides, with model affinity calculations based off of the 11mer (residues 45-55) conserved epitope. **(B)** Linear regression fit for all 24 peptides, with model affinity calculations based off of the 10mer epitope (residues 46-55) and no residues in position

45 included in the calculations. **(C)** Linear regression fit for the 8 peptides that only had C-terminal flanking sequences added (does not include the unmodified peptides 1 and 20). **(D)** Linear regression fit for the 10 peptides with only N-terminal flanking sequences (includes peptide 20).

The lack of correlation for the tight binding peptides and the good correlation of the weaker binding peptides prompted us to exam the data in more detail and to design peptides 20-24. The binding data show no correlation with calculated peptide helicity (see **Supporting Information** for details). Alternative modes of peptide:antibody binding other than that depicted in **Figure 3.3** were studied in detail. Partial binding modes, consistent with a helix setting on a protein surface, were unsuccessful at improving the correlation. Interestingly, while these alternatives did not improve the correlation certain combinations could compromise the correlation (see **Supporting Information** for details). Remarkably, however, there was a significant improvement in the correlation when the decapeptide epitope, which spans residues 46-55, was expanded to the undecapeptide epitope, i.e., when residue 45 was included in the set of 46-55. The correlation for the peptides with C-terminal flanking residues remained unchanged, of course, though if the undecapeptide was included in the correlation the fit improved (R^2 = 0.86). **Figure 3.5 (A)** shows the calculated and the experimentally determined binding energies for all 24 peptides, which indicates an excellent correlation between computed and observed binding energy (R^2 = 0.76). This improvement was also noted when the expanded epitope (residues 45-55) was evaluated for alternative modes of peptide:antibody binding (c.f. **Figure 3.3**). Equally good correlations were observed for partial binding modes consistent with a helix setting on a protein surface, provided residue 45 was included in the interface. Alternative binding modes that lacked this residue abolished the correlation in binding (see **Supporting Information** for details).

Each of five additional peptides included L45, and indeed, were found to have enhanced affinities for the antibody. And although the unmodified undecapeptide performed best, the affinity differences between these five peptides are insignificant. There is an effective ceiling evident from the experimental data for the peptides with the highest affinities. To significantly increase the binding affinity of the undecapeptide our model indicates that the flanking regions would have to be non-polar amino acid residues. Practically, however, flanking regions dominated by non-polar residues would be expected to attenuate solubility, increase the likelihood of aggregation, and promote non-specific binding. This appears to be the case for the constructs with the most non-polar flanking regions (peptides 16, 22,23 24). For example, Peptide 24 has the highest overall μ profile and calculated affinity but has a somewhat lower experimental binding affinity. This is likely due to aggregation, as suggested by the non-one-to-one SPR binding profile of higher peptide concentrations for this and similar peptides.

The HSyM model predicted that the five additional peptides would have very high but similar affinities, assuming that the peptides remained aggregation-free, maintained specific binding, and had sufficient solubility. Surprisingly, Peptide 21 bound much more tightly than expected - equivalent to the other tightest binding peptides - and significantly reduced the binding energy correlation. Indeed, ignoring this peptide the correlation is improved (R^2 = 0.85, compared to R^2 = 0.76).

The monoclonal antibody immunoassays corroborated the SPR measurements. The peptides with the lowest and highest affinities determined by SPR had the highest and lowest limits of detection (LOD), respectively. **Table 3.2** shows the LODs for select peptides. The data are consistent with the SPR results: the engineered peptides had well over 100-fold LOD improvement over the dodecapeptide. Peptides that housed a non-polar residue in position 45 had the lowest LODs. Fully functionalized beads showed slow deterioration in function over the

course of several months, as evinced by steadily increasing LODs. Evidently, this bead-peptide construct combination slowly degrades upon storage under these conditions.

These data suggest that the correct epitope for this antibody should be taken to be the undecapeptide, residues 45-55 in the p53 Transactivation domain. It should be added that the model indicated that the N-terminal region (residues 43-46) can make a significant contribution to binding free energy. This suggests that the model may be useful for identifying potentially important portions of epitopes used to detect autoantibodies and predict where an epitope sequence might be better terminated, and optimized. This sort of strategy may improve detection of relevant polyclonal autoantibodies. The HSyM model is based on deep learning information and represents a method that is complementary to modern epitope predictors that use machine-learning methods [8, 9].

Peptide	Sequence	Type	Antibody LOD (ng/mL)	Std Dev
1	SPDDIEQWFT	Native	2.52	0.8
10	KFVN-10mer	Optimized	0.30	0.18
11	VFNK-10mer	Optimized	0.31	0.21
12	NKVF-10mer	Optimized	0.04	0.02
16	DLM-11mer-EDPG	Native	0.06	0.02
20	LSPDDIEQWFT	Native	0.01	6.00E-04
21	KKNN-11mer	(De)Optimized	0.009	0.003
22	DDLM-11mer	Native	0.02	0.02
23	YYEY-11mer	Optimized	0.02	0.02

Table 3.2: Immunoassay Limits of Detection for Selected Peptides

These modifications, in addition to a more strenuous determination of which WT residues are likely to make up the conserved epitope, are expected to improve the model's ability to design high affinity peptides going forward. To fully demonstrate the applicability of the model for improving peptide-based diagnostic assays, a case-control study focusing on the detection of

p53 autoantibodies in colorectal cancer sera samples is planned. The model will be used to expand and enhance additional p53 linear epitopes, and the modified peptide assay will be compared to its WT peptide and whole protein counterparts. The design of a panel of engineered peptides for this and other studies is outlined below:

1. Select 4-5 known, immunodominant linear epitopes of approximately ten residues.
2. Use the HSyM model to evaluate the WT proteins. Examine the flanking residues that appear in the WT sequence of each putative epitope. Flanking residues with high gamma values, and hence would have the potential to remotely impact binding should be considered for inclusion in an expanded antigen/epitope.
3. Engineer flanking residues to balance polar and non-polar residues and to match the μ-profile of the peptide construct to, or above, the WT values. The number and location of flanking residues will likely be dependent on the coarse conformation of the antigenic peptide.

Conclusion

Engineered peptides, proteins, and small molecules have many biomedical applications, including as vaccines [10, 11], therapeutics [12, 13], scaffolds for tissue engineering [14, 15], drug-delivery carrier and targeting systems [14, 16], as well as in clinical and investigatory immunoassays [1, 17]. Peptides offer several potential advantages over proteins, such as greater long-term stability, ease of synthesis, as well as potentially better intracellular access, greater target specificity and selectivity, and lower risk of toxicity or side effects [12, 18]. Most applications depend on the peptide binding to a protein target. Phage display and microarray libraries and other high-throughput screening methods, [19, 20] as well as in silico docking [19,

21, 22] and all-atom simulations [23, 24], are powerful approaches by which peptides with improved properties can be designed [19, 25].

The coarse-grained HSyM protein energetics model was able to accurately predict the relative binding affinities of a suite of 24 peptides based on a single linear p53 epitope for a monoclonal antibody with good reliability *without any structural data on the peptide, antibody, or peptide:antibody complex* ($R^2 > 0.75$). This method is computationally inexpensive and fast and stands in contrast to current *in silico* methods for peptide engineering, which rely on computationally expensive all-atom molecular dynamics and docking simulations or machine learning methods. Furthermore, the major peptide engineering premise of this work—the hypothesis that adding flanking residues can affect the binding interaction energies of interior residues to remotely control binding affinity — has been difficult to realize with other methods. While some studies have added flanking sequences to peptides to increase their affinity for a target, these results were largely attributed to secondary interactions between the flanking residues and the target [26, 27] and not to the increased binding affinity of the conserved residues.

Our modeling and experimental affinity data also revealed surprising information about this particular peptide-antibody system which may have implications for other intermolecular systems. Contrary to the epitope information provided by the monoclonal antibody vendor, our SPR data indicated that Leu45 should be included in the conserved epitope sequence. Peptides with Leu, Met, or Phe in this position had the highest affinities. When Leu45 was included as a conserved epitope residue in HSyM's calculations, the model's accuracy significantly increased, further corroborating this finding, and strongly suggest that the observed effects are not well-attributed to non-specific binding. We also noted that evaluation of the WT p53 sequence indicated that Leu45 had the potential to contribute significantly to binding. While the exact

epitope cannot be assigned with certainty, we strongly suspect this residue is part of the epitope for this mAb. Along these lines, we also note that the data suggest that the binding epitope may consist of only five consecutive residues, perhaps representing a single helical turn. Specifically, when L45-D49 is assigned contacts to a putative antibody paratope an excellent fit of the experimental data is evident (R^2 = 0.77), since these five residues appear to be most responsible for the antibody binding activity. Even ignoring the likelihood of L45 being part of the epitope, model-guided remote-controlled binding was realized, and several peptides with >100-fold increase in binding and LOD improvements were successfully designed, prepared, and characterized.

We engineered a peptide derived from a single, immunodominant epitope of the tumor suppressor protein p53 and a corresponding monoclonal antibody specific to that epitope. This system was chosen to correspond to the observed presence of autoantibodies specific to p53 in the sera of patients with a wide variety of cancers [28, 29]. Twenty-four peptide variants of the flanking residues around the conserved epitope we prepared and evaluated. The binding affinities were measured using surface plasmon resonance and confirmed by limit of detection immunoassays. Compared to the putative decapeptide epitope, several variants had significantly improved binding affinities and limit of detection (enhanced by factors of >300). The correlation of calculated and observed binding energies was poor for peptides with segments that flanked the N-terminus of the putative epitope (R^2 = 0.02), whereas excellent correlations were noted for peptides with segments that flanked the C-terminus of the putative region epitope (R^2 = 0.72). Importantly, when we considered the epitope to span the sequence 45-55, instead of 46-55, our calculations for all 24 peptides of the study correlated well with experiment (R^2 > 0.75). These findings are remarkable since the study was guided by the folded parent protein alone – no structural data on the antibody-peptide complex structure are

available. Taken together, this study suggests that the HSyM model complements powerful computational and experimental tools and is a useful engineering tool for visualizing and approximating interaction energy.

Acknowledgements

The authors would like to thank Drs. Nilgun Tumer and Xiao-ping Li (Rutgers University, School of Environmental and Biological Sciences) for their assistance with the SPR (Biacore T200) measurements, and Drs. Hans Wandall (University of Copenhagen, Denmark) and Usha Menon (University College London, UK) for providing clinical sera samples and additional guidance. This work was partially funded by NIH grants T32GM008339 and 1S10OD026750-01 ("A Biacore 8K System.").

Supplementary Information

S3.1 Error Analysis

It is important to consider the limitations and constraints placed on the model in order to determine ways to improve it. First and foremost, having sequence and structural data of the antibody paratope or especially the antibody-peptide complex would certainly improve the model's accuracy, as the number and strength of the antibody-antigen contacts could be better estimated. Most peptide design and affinity prediction tools, such as the PepSpec algorithm and docking methods, require at least a target structure or homology model [19, 30]. The fact that the HSyM model was still able to provide relatively accurate affinity predictions even without this information is encouraging for future studies. Furthermore, our model subsumes enthalpic and entropic contributions from the solvent, even though these factors have proven critical for influencing the affinity of protein-ligand interactions [31]. Much like our previous work with T4 lysozyme [4], adding entropy terms (especially those that account for solvent displacement) to the model's calculations should improve its correlation with experimental data. Peptide solubility must also be taken into account; while it is tempting to design very "hot" peptides with hydrophobic, high μ value flanking sequences, this will tend to cause aggregation that will reduce apparent affinity. The majority (17 out of 24) of our designed peptides had poor aqueous solubility and had to first be dissolved in DMSO prior to SPR testing or coupling to microbeads for immunoassays. Self-aggregation might explain why some "hot" peptides, such as the 19mer, performed worse than expected. Finally, based on a limited number of structures of the p53 epitope, we assumed that our peptides maintained a helical structure upon binding to the antibody. This of course may not be the case, or more likely, the terminal regions of the

peptides may assume flexible conformations. Adding a term to estimate total peptide helicity or assigning an extended structure to the flanking residues could address these issues.

S3.2 Clinical Cancer Sera Immunoassays

For the sera sample immunoassays, 10 commercially derived sera samples were obtained from our collaborators and tested, with 5 derived from patients with colorectal cancer and 5 from cancer-negative subjects. These were compared to a negative control of pooled normal human sera. Using a positivity cutoff of the negative control's signal plus three times its standard deviation, the sensitivity and specificity of each peptide was calculated and is displayed in **Table S3.1**. Of interest is the observation that only the two peptides with a native WT N-flanking sequence (N Native 15mer and Native 18mer) were able to give positive results for Sample B, but also gave false positive results for Sample G, lowering their specificity. This suggests that for some polyclonal autoantibodies the full epitope might include the other N-flanking residues in addition to Leu45. These results should be regarded as a proof-of-concept, as a full case-control study would require not only a larger sample size, but additional optimized peptides corresponding to other p53 epitopes in order to recapitulate the polyclonal autoantibody response.

Peptide	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Neg 1	Neg 2	Neg 3	Neg 4	Neg 5	Sens	Spec
1	0/3	1/3	0/3	1/4	0/3	0/3	0/3	0/3	0/3	0/3	13%	100%
12	0/3	2/4	2/4	3/3	2/4	0/4	0/3	0/3	0/3	0/3	50%	100%
16	3/3	1/4	0/3	3/3	3/3	1/4	0/3	0/3	0/3	0/3	63%	94%
20	0/3	0/3	0/3	3/3	3/3	0/3	0/3	0/3	0/3	0/3	40%	100%
22	3/3	0/3	0/3	3/3	3/3	3/3	0/3	0/3	0/3	0/3	60%	80%

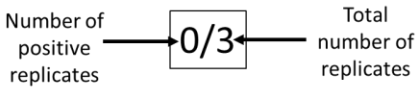


Table S3.1: Results of Cancer Clinical Sera Immunoassays for Selected Peptides

Figure S3.1: Alternate Peptide Surface Affinity Correlations

For the following figures only the epitope residues enclosed in the red box (in the case of L45 this includes any residue at that position) were used to determine the model-calculated ΔG of binding. The number of antibody contacts per residue and antibody μ values are assigned according to **Figure 3.3.**

Figure S3.1A: 11mer: All Residues

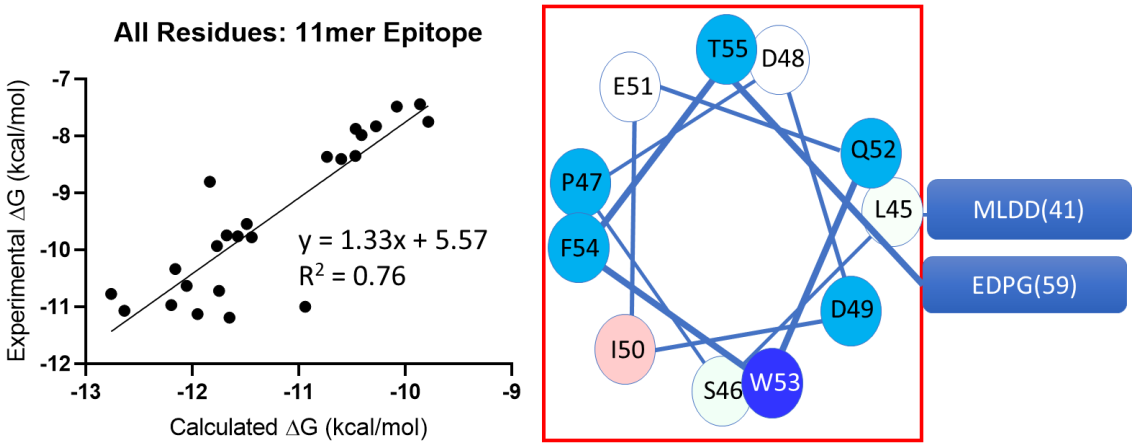


Figure S3.1B: 10mer: All Residues

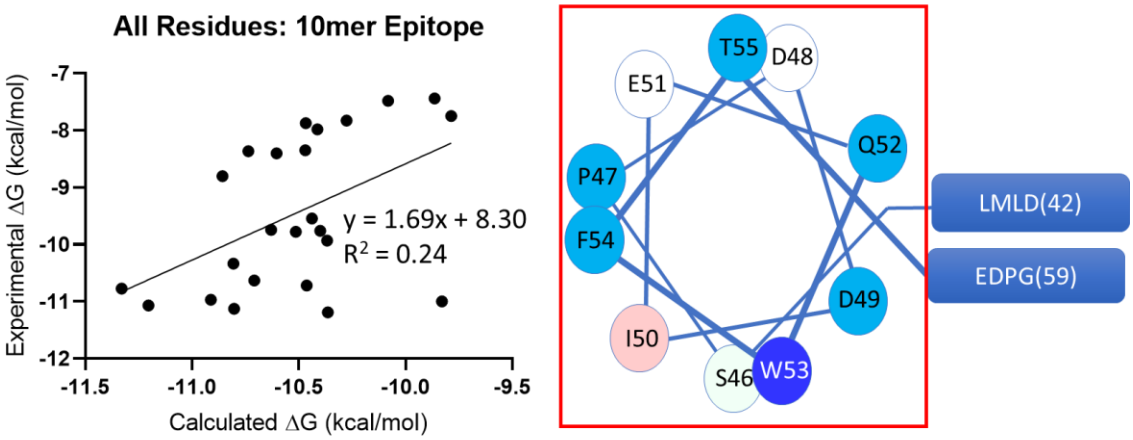


Figure S3.1C: Right Surface Plots

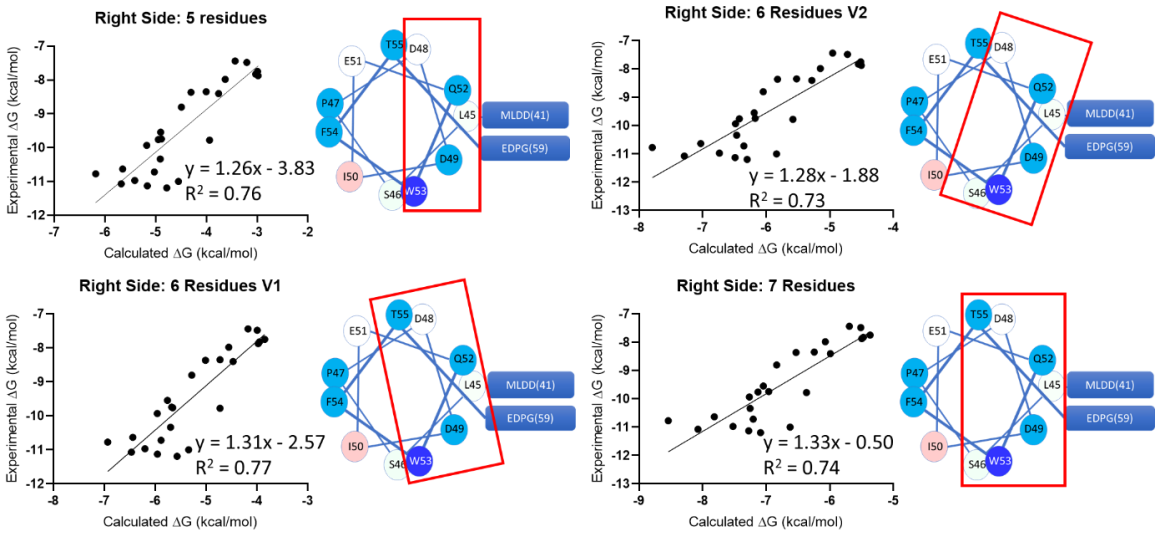


Figure S3.1D: Top Surface Plots

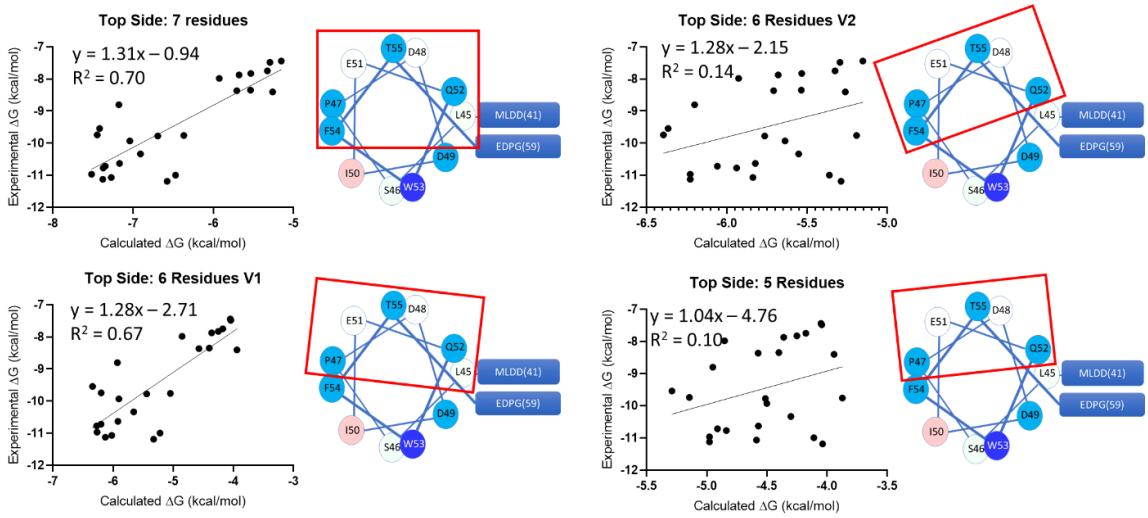


Figure S3.1E: Bottom Surface Plots

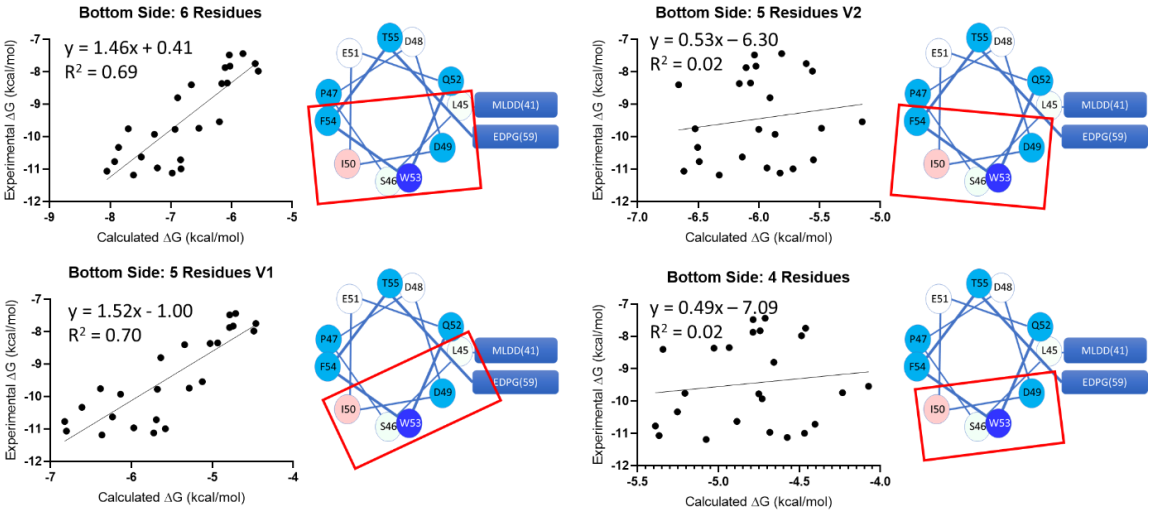


Figure S3.1F: Left Surface Plots

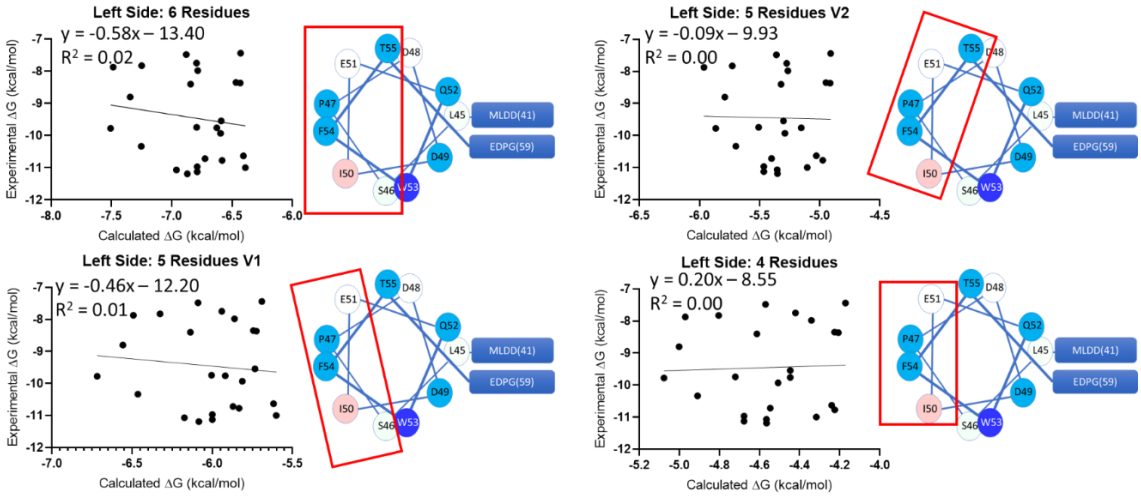
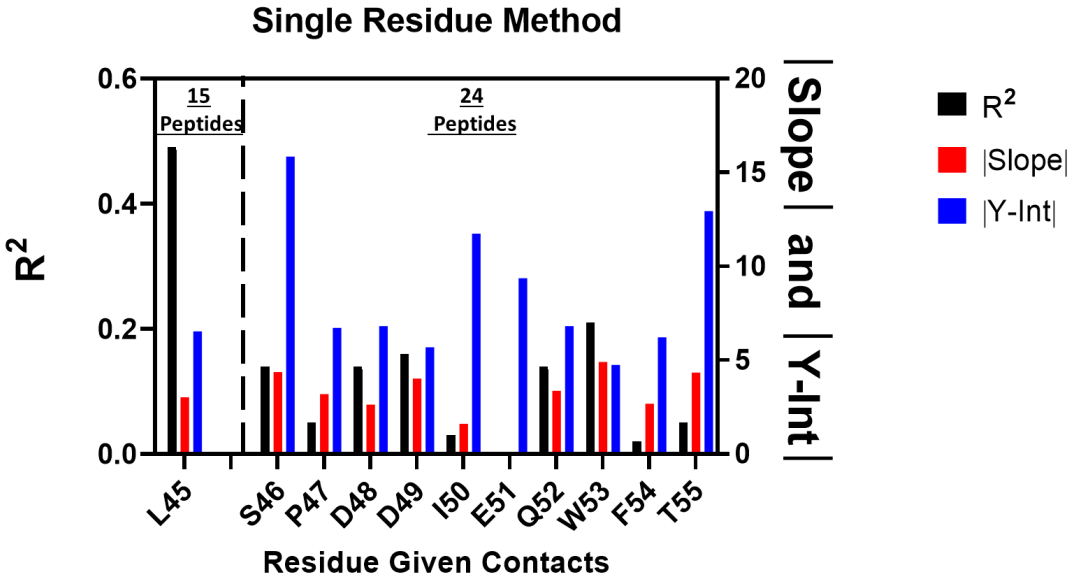
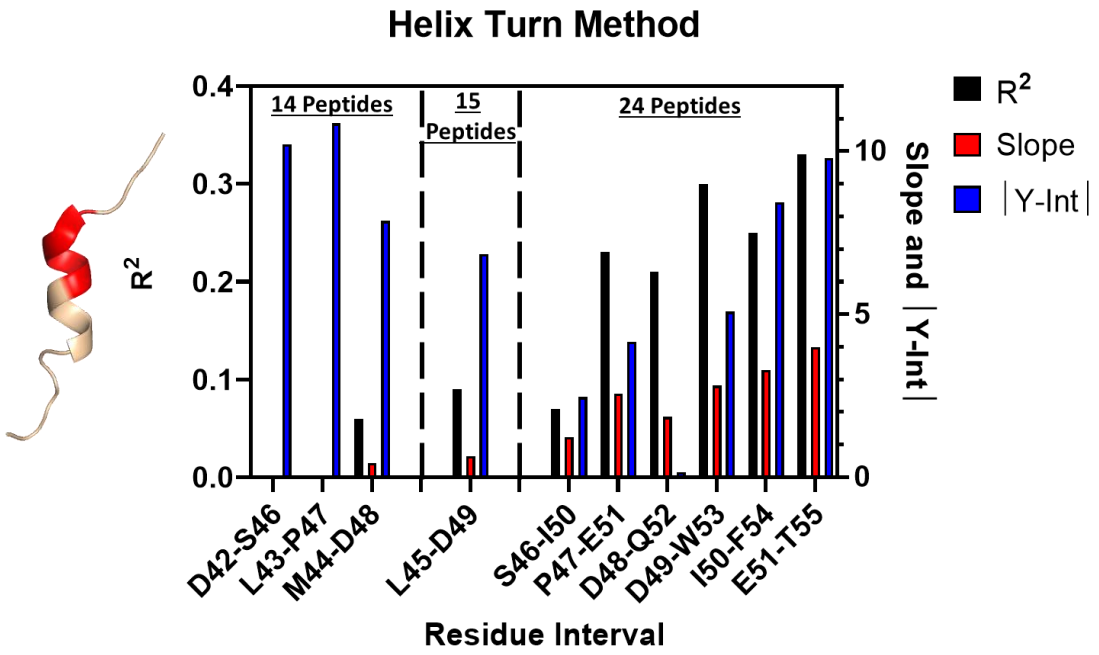


Figure S3.2: Single Residue Peptide Affinity Correlations



In this method only one residue in the epitope is assigned antibody contacts (see **Figure 3.3**) and used to determine the model-calculated ΔG. For the L45 residue, only the 15 peptides with an amino acid in that position were used to obtain the coefficient of determination, slope, and y-intercept of the regression line comparing the model and experimental affinity values.

Figure S3.3: Helix Turn Method



In the helix-turn method, only five consecutive residues (constituting a single helical turn, as in the residues colored red in the left Pymol image) were used to determine the model-calculated ΔG of binding. For the N-flanking residues, only those peptides with amino acids at all of those positions were used to obtain the coefficient of determination, slope, and y-intercept of the regression line comparing the model and experimental affinity values.

Table S3.2: Excluded Residue Method

Excluded Residue	R^2	Slope	Y-Intercept
L45	0.24	1.69	8.31
S46	0.79	1.33	3.55
P47	0.76	1.37	4.80
D48	0.76	1.48	5.76
D49	0.74	1.41	5.09
I50	0.78	1.33	3.67
E51	0.75	1.33	4.15
Q52	0.77	1.46	5.83
W53	0.71	1.36	4.60
F54	0.74	1.33	3.92
T55	0.75	1.30	4.19

This method excludes one residue from the 11mer epitope when determining the model-calculated model-calculated ΔG of binding, with hypothetical antibody contacts being assigned to all other residues as in Figure 4. The results for the L45 residue include peptides that do not have any residue at this position; if these peptides are excluded the R^2, slope, and y-intercept of the linear regression line change to 0.01, 0.18, and -8.43, respectively.

Figure S3.4: Peptide Helicity Affinity Correlations Using Web-Based Secondary Structure Predictors

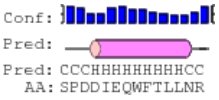
In the following figures various metrics of peptide helicity obtained from web-based secondary structure predictors fitted against the SPR peptide affinity data to determine if the peptide's affinity for the monoclonal antibody is correlated with their helicity.

Figure S3.4A: PEP2D

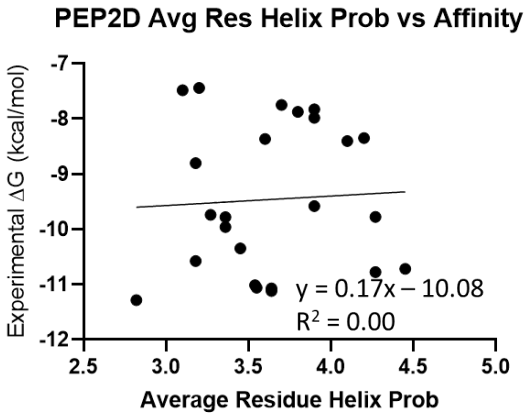
Site: http://crdd.osdd.net/raghava/pep2d/ [32]

Example Output:

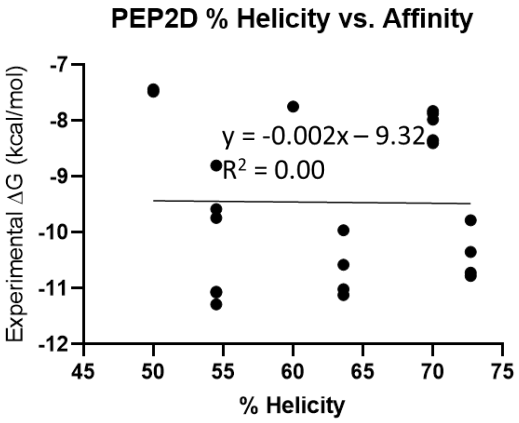
Secondary structure of peptide (**COpt3**)




```
Sequence   : SPDDIEQWFTLLNR
Structure  : CCCHHHHHHHHCC
Helix Prob : 00247866543300
Sheet Prob : 00000000000000
Coil  Prob : 87420001012359
```



The average residue helix probability for each peptide is calculated from the 11mer epitope or 10mer for peptides without the L45 position.



The helix percentage is calculated based on the number of residues within the epitope assigned an "H" configuration.

Figure S3.4B: Agadir

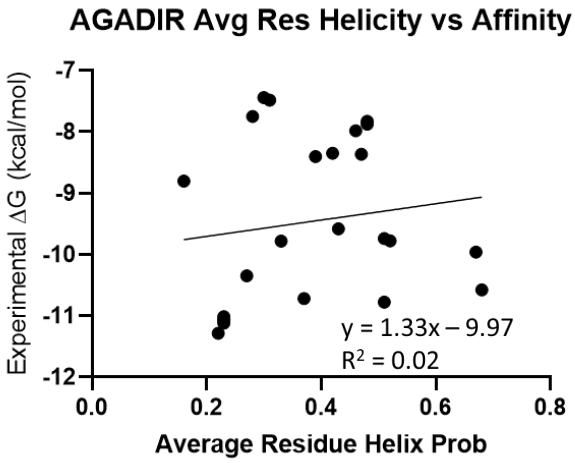
Site: http://agadir.crg.es/protected/academic/calculation3.jsp [33-35]

Example Output:

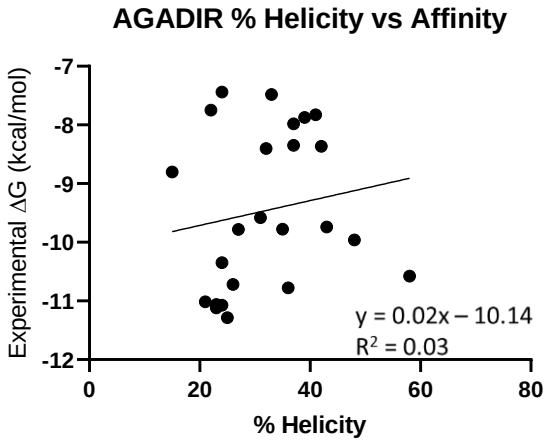
```
pH                  7
Temperature         298.15
Ionic Strength      0.1

Nterm               free
Cterm               amidated

Peptide 1       SPDDIEQWFT
res,  aa,  Hel,  Ncap,  Ccap,Hstaple,Schellm, CaH,  13Ca, JaN
-0.000784
01,   S,   0.0,  0.16,  0.00,  0.00,  0.00,  0.13 , -2.20,  6.61
02,   P,   0.2,  0.02,  0.00,  0.00,  0.00,  0.04 ,  0.00,  6.00
03,   D,   0.2,  0.25,  0.00,  0.00,  0.00, -0.00 , -0.00,  6.60
04,   D,   0.4,  0.14,  0.00,  0.00,  0.00, -0.00 ,  0.01,  6.60
05,   I,   0.6,  0.01,  0.00,  0.00,  0.00, -0.00 ,  0.02,  7.23
06,   E,   0.6,  0.01,  0.03,  0.00,  0.00, -0.09 ,  0.01,  6.07
07,   Q,   0.5,  0.00,  0.04,  0.00,  0.00, -0.11 ,  0.01,  6.39
08,   W,   0.4,  0.00,  0.14,  0.00,  0.00, -0.07 ,  0.01,  6.71
09,   F,   0.1,  0.00,  0.33,  0.00,  0.00, -0.10 ,  0.01,  7.08
10,   T,   0.1,  0.00,  0.00,  0.00,  0.00, -0.14 ,  0.00,  7.46
11,   U,   0.0,  0.00,  0.05,  0.00,  0.00,  0.00 ,  0.00,  0.00
Percentage helix        0.33
```

Input Parameters:

- Acetylated N-terminus (where applicable)
- Amidated C-terminus (where applicable)
- pH = 7.0
- Temperature = 298.15 K



The average residue helix probability for each peptide is calculated from the 11mer epitope or 10mer for peptides without the L45 position.

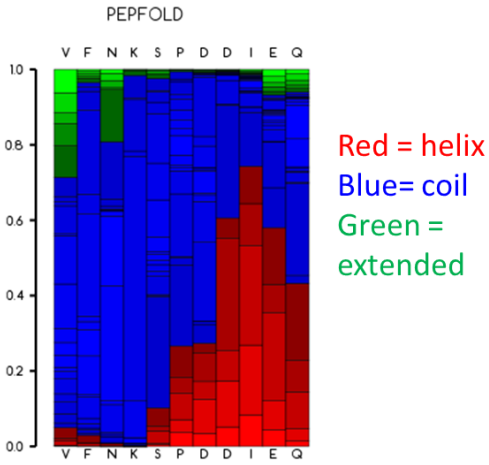


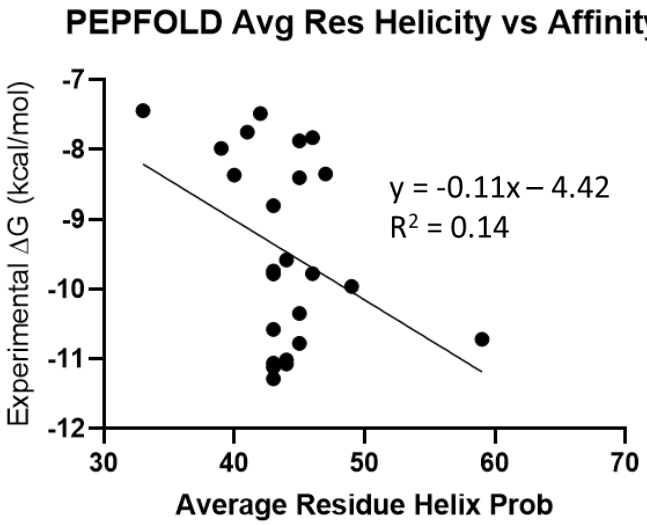
Used the outputted "percentage helix" value for each peptide.

Figure S3.4C: PEP-FOLD 3

Site: https://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/ [36-38]

Example Output:





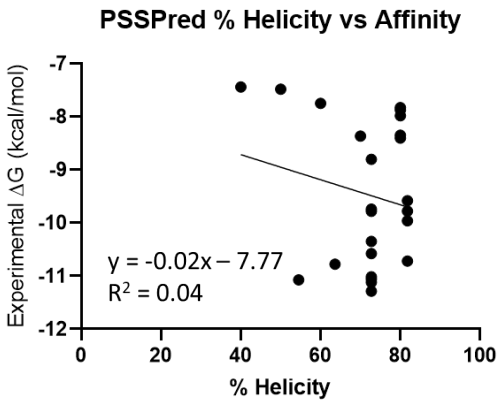
The residue helical probabilities of S46-Q52 (these were the only residues consistently displayed for all peptides) were extracted with a graph-reading program and averaged.

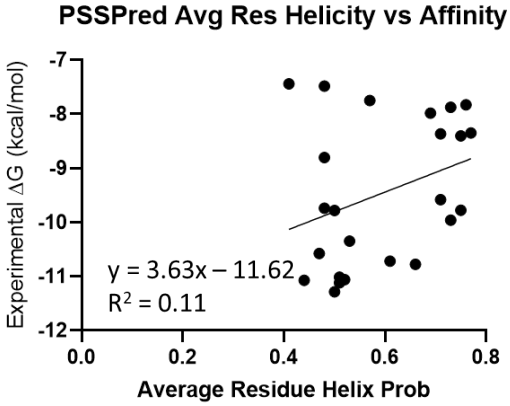
Figure S3.4D: PSSPred

Site: https://zhanglab.ccmb.med.umich.edu/PSSpred/ [39]

Example Output:

```
11   coil  helix  beta

   1 L C  0.941  0.043  0.008
   2 S C  0.840  0.133  0.020
   3 P H  0.473  0.529  0.014
   4 D H  0.394  0.596  0.017
   5 D H  0.227  0.754  0.014
   6 I H  0.192  0.773  0.022
   7 E H  0.163  0.781  0.026
   8 Q H  0.185  0.725  0.051
   9 W H  0.261  0.658  0.048
  10 F H  0.436  0.491  0.043
  11 T C  0.991  0.004  0.014
```





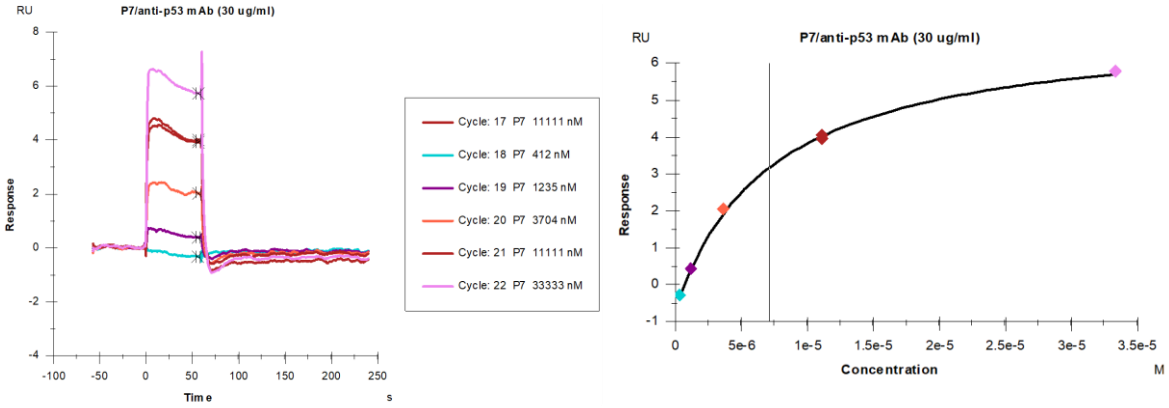
The helix percentage was calculated based on the number of residues within the epitope assigned an “H” config

The average residue helix probability for each peptide is calculated from the 11mer epitope or 10mer for peptides without the L45 position.

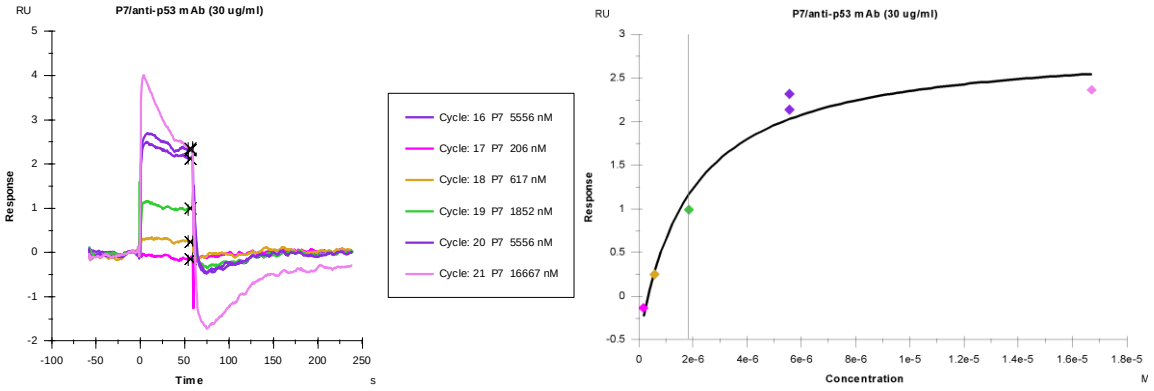
Figure S3.5: Peptide SPR Sensograms

Steady State and Kinetic curve fitting was done using the Biacore T200 Evaluation software and a 1:1 binding model. For Kinetic analysis only the four lowest concentration peptide sensograms were used to minimize fitting inaccuracies due to suspected peptide aggregation.

Peptide 1

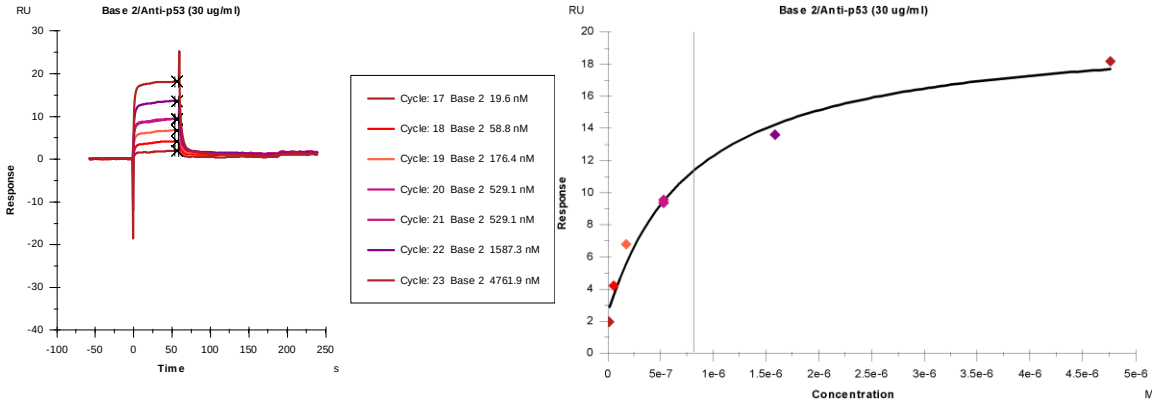


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
7.16E-06	7.806	-0.7259	0.00625



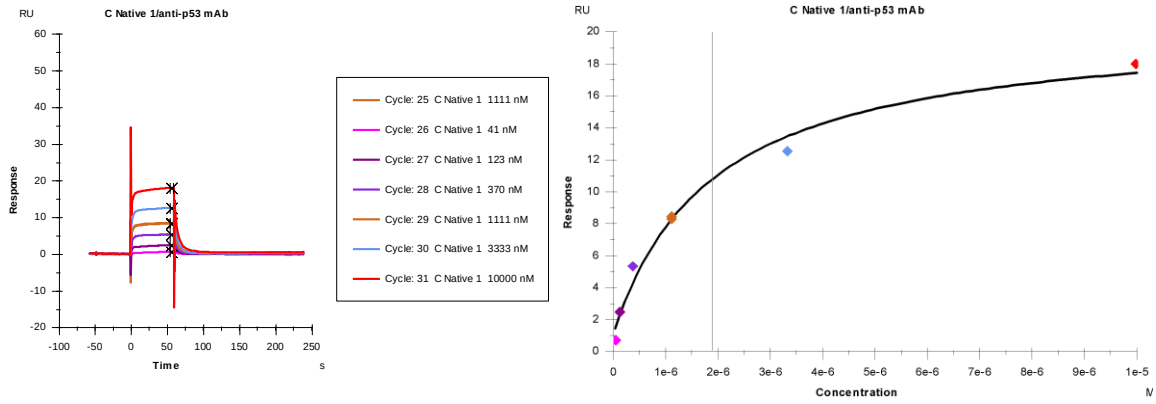
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.84E-06	3.459	-0.5662	0.0556

Peptide 1



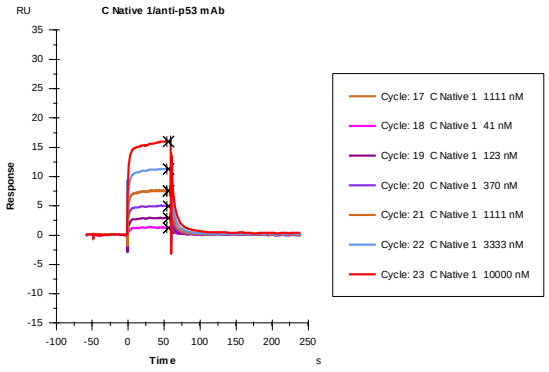
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
8.14E-07	17.83	2.47	0.777

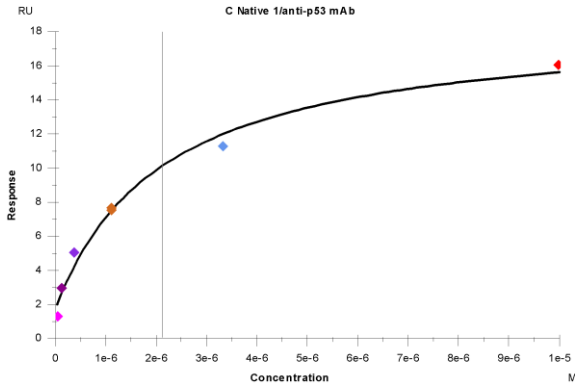
Peptide 2



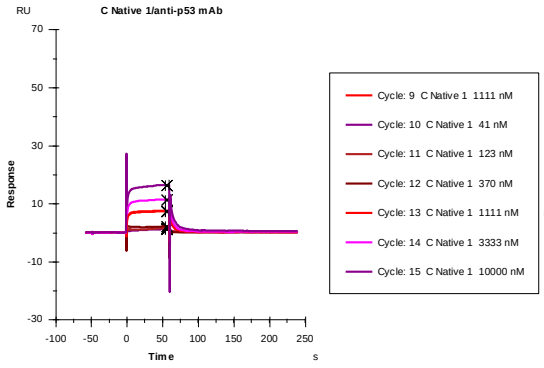
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.89E-06	19.48	1.037	0.756

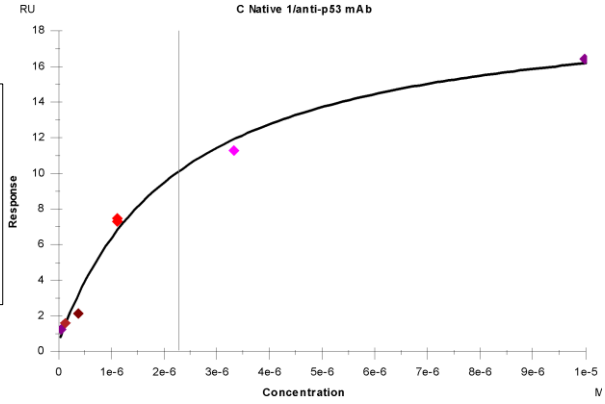
Peptide 2





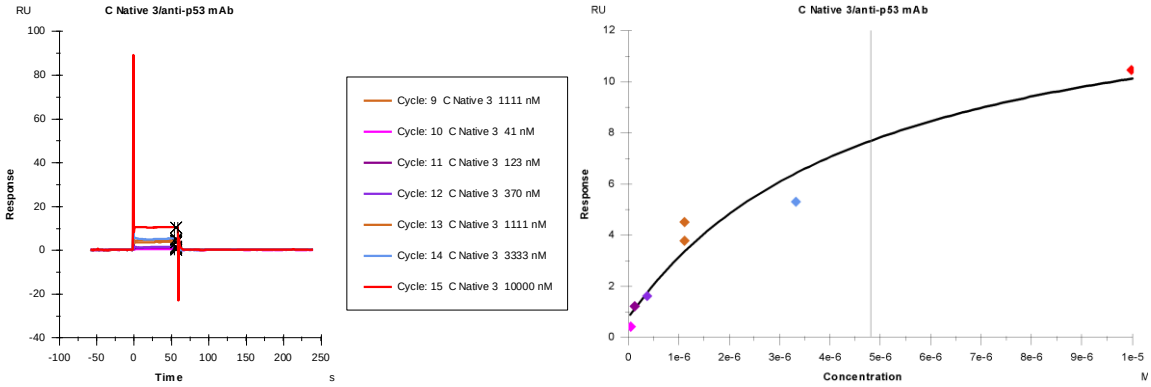
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
2.13E-06	16.89	1.69	0.495



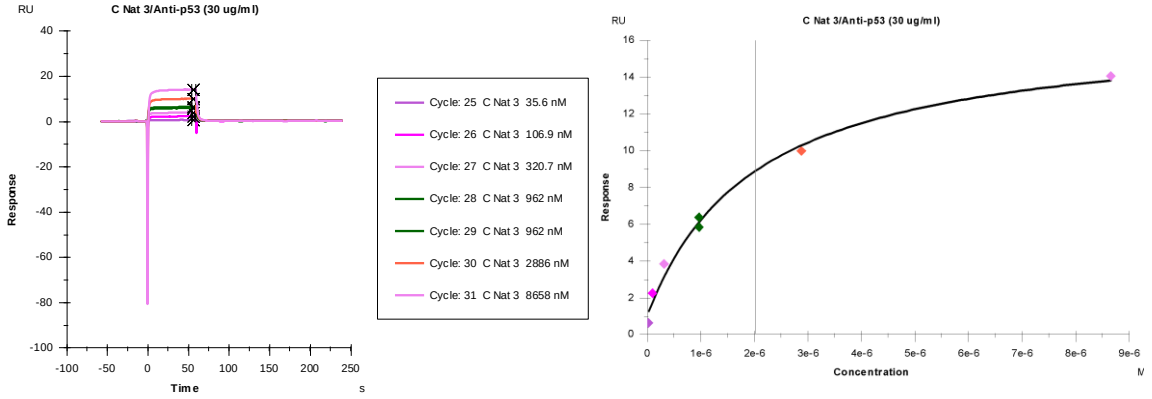


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
2.28E-06	19.31	0.4631	0.61

Peptide 3

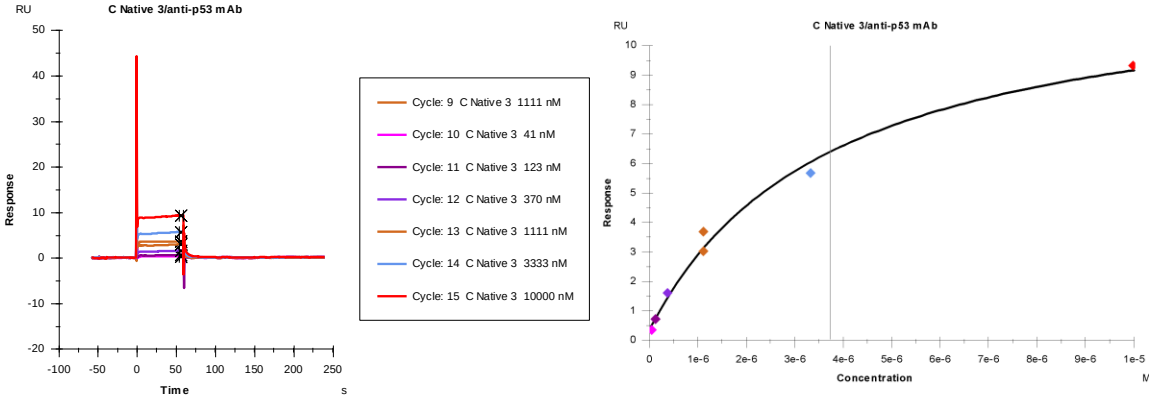


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
4.81E-06	13.87	0.7682	0.772



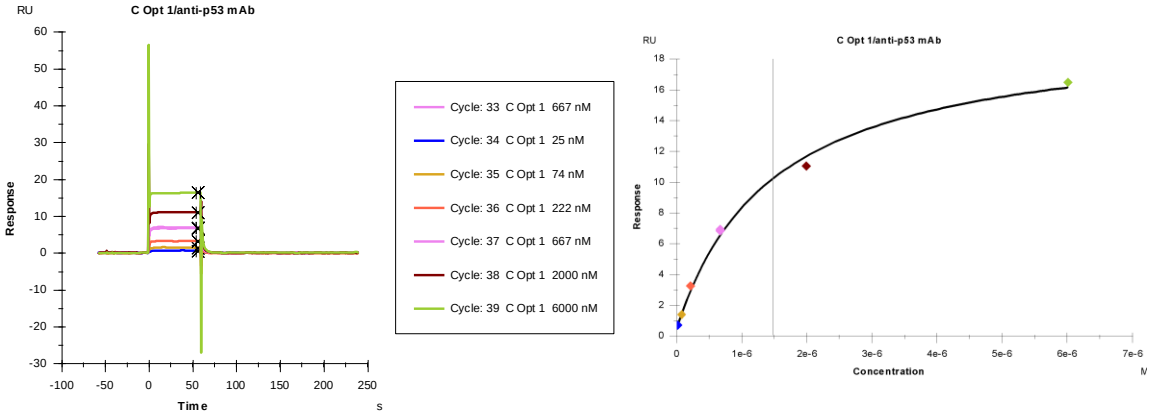
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
2.03E-06	15.8	1.017	0.346

Peptide 3



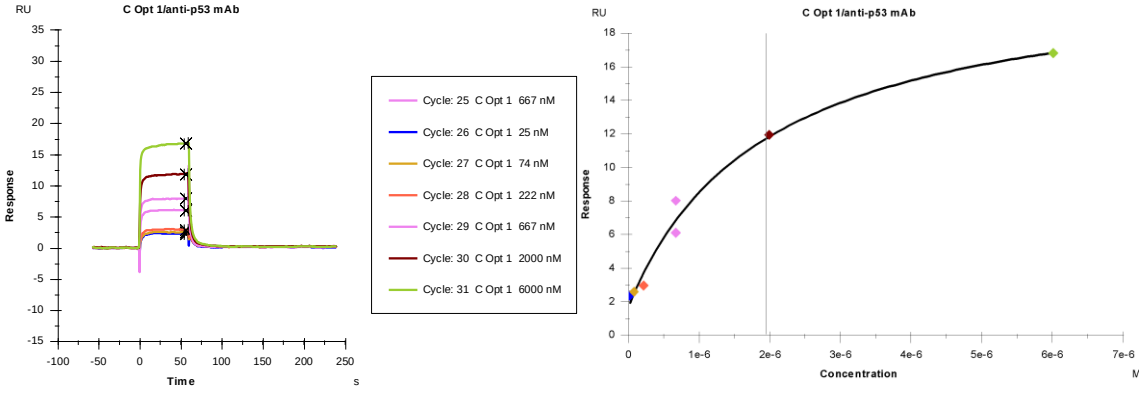
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
3.74E-06	12.11	0.353	0.133

Peptide 4

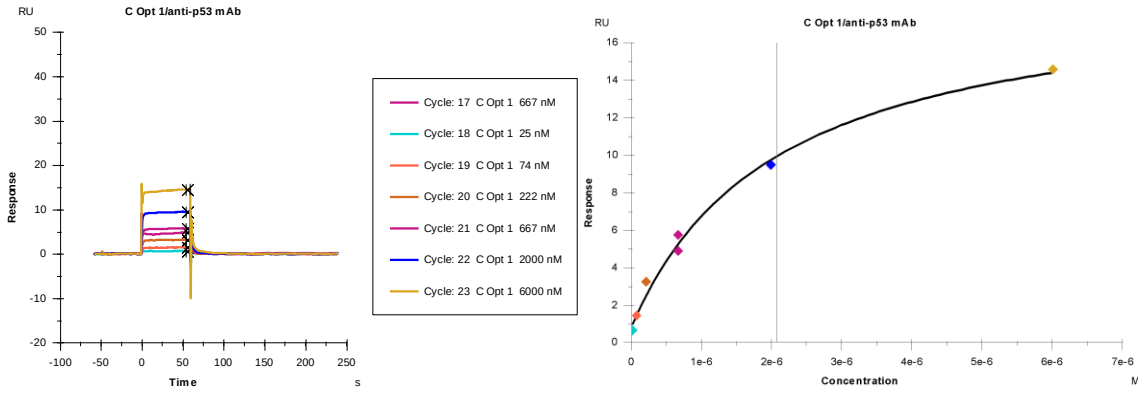


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.48E-06	19.53	0.4988	0.201

Peptide 4

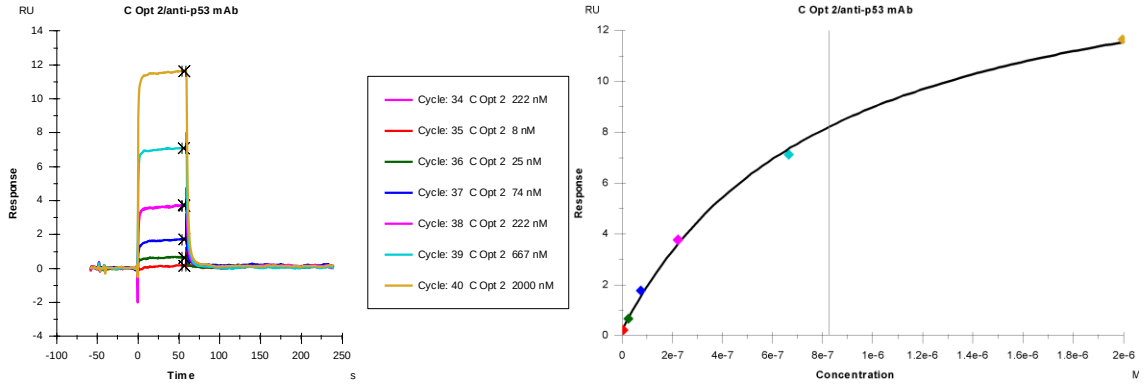


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.94E-06	20.06	1.673	0.674

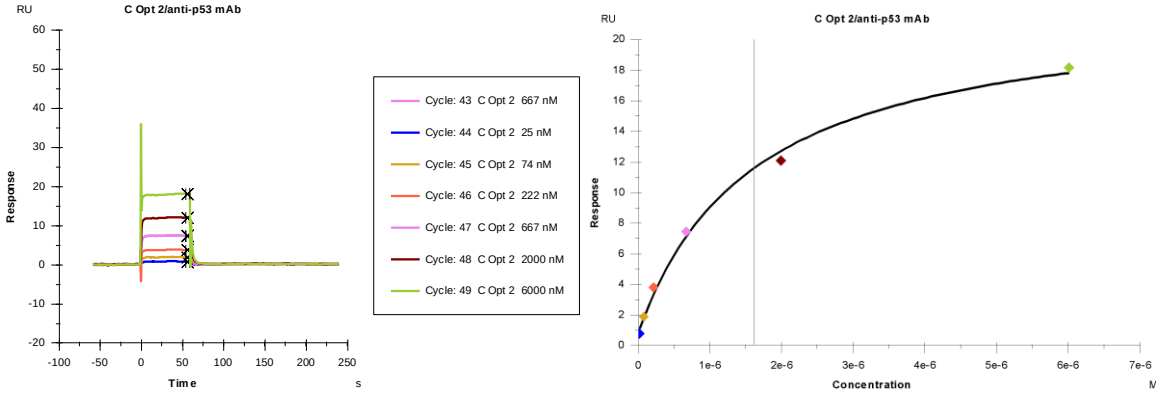


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
2.08E-06	18.32	0.8112	0.266

Peptide 5

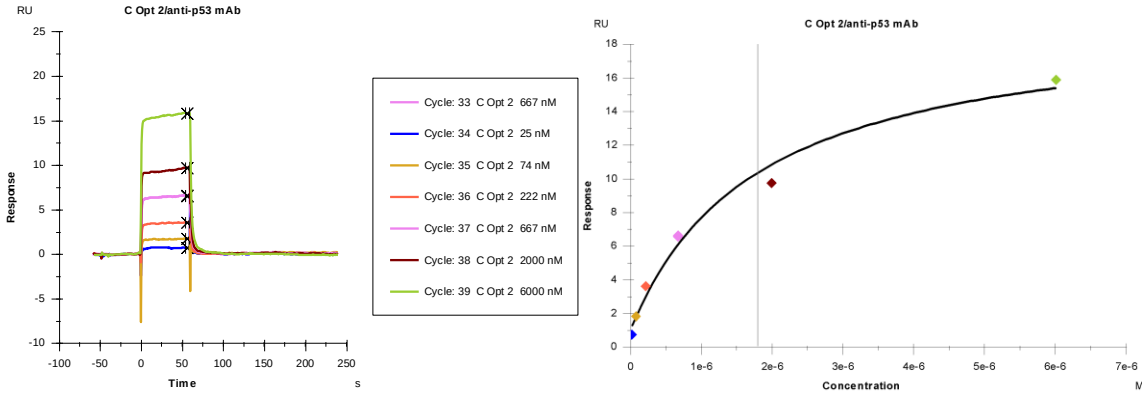


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
8.26E-07	16.03	0.2076	0.0498



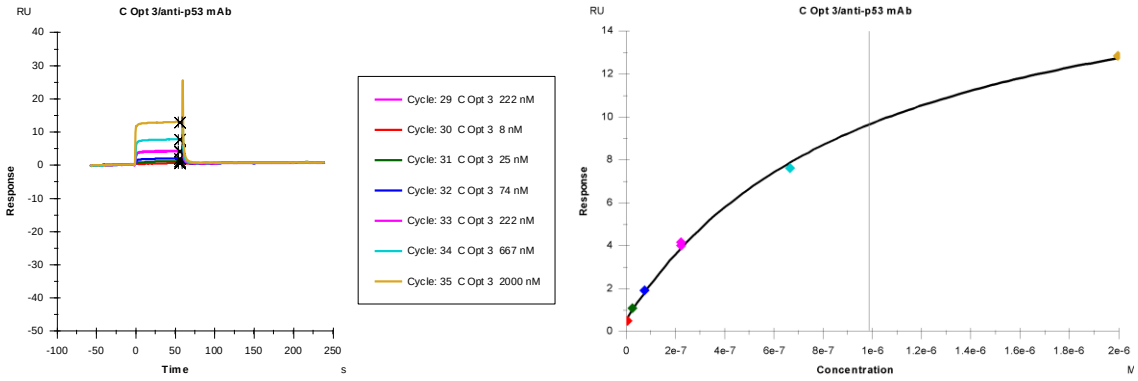
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.62E-06	21.57	0.8275	0.266

Peptide 5



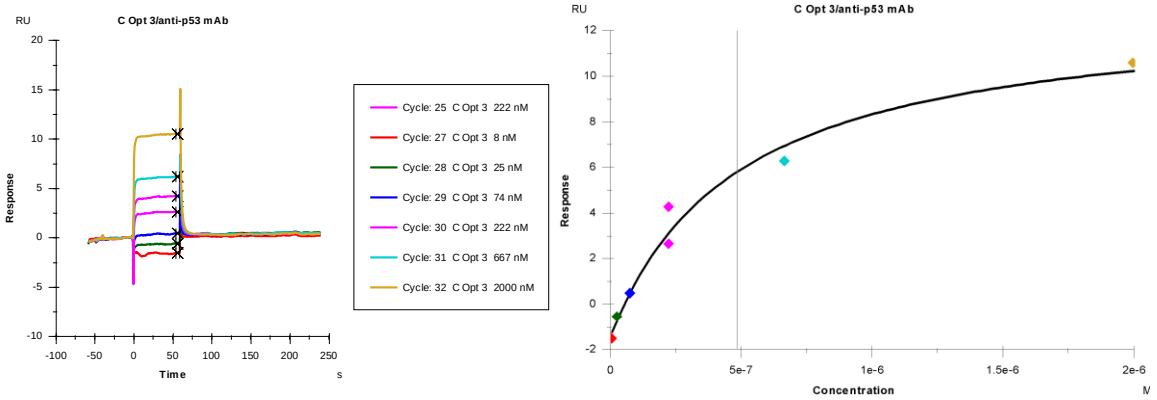
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.80E-06	18.67	1.039	0.644

Peptide 6

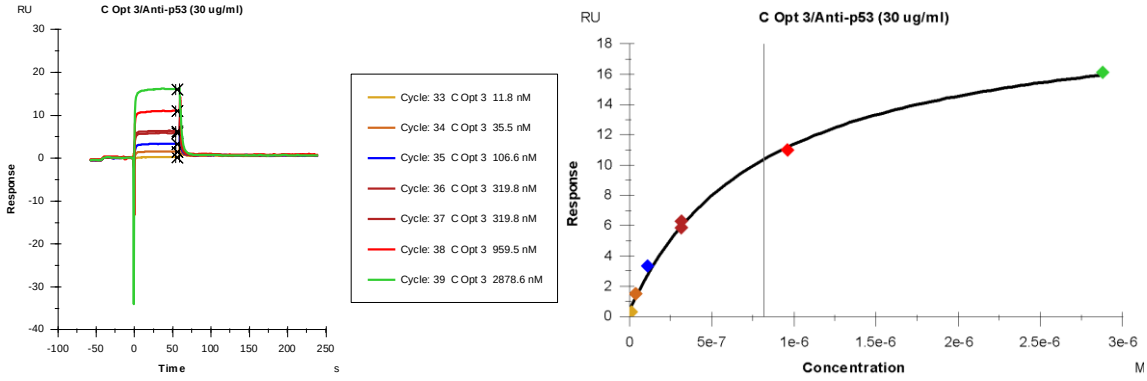


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
9.87E-07	18.24	0.5267	0.0508

Peptide 6

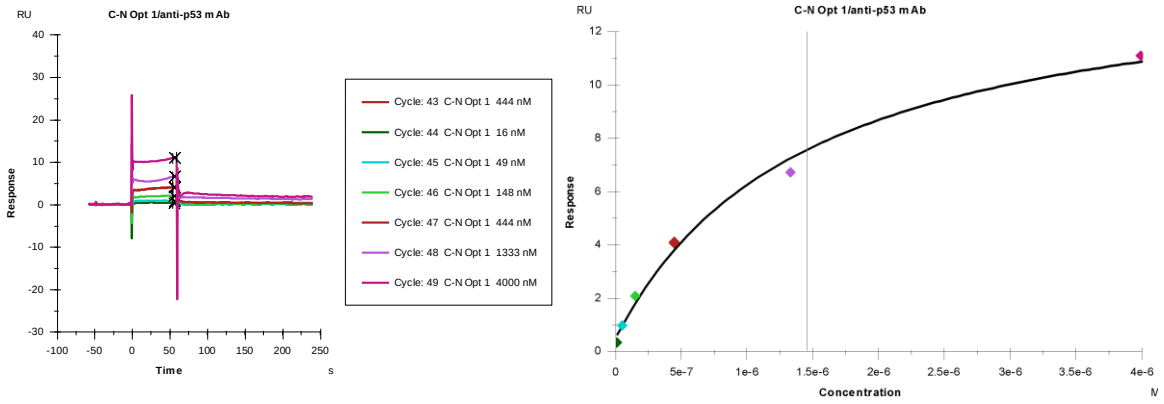


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
4.85E-07	14.53	-1.453	0.564

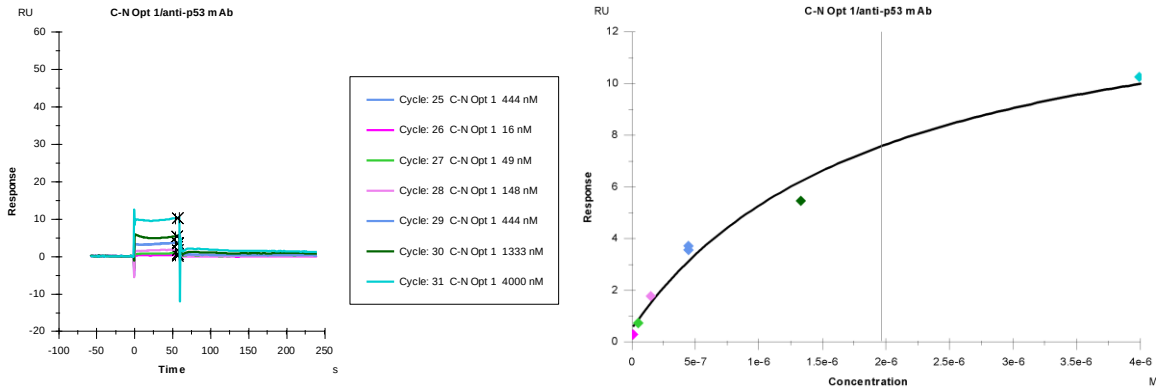


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
8.15E-07	19.9	0.4408	0.185

Peptide 7

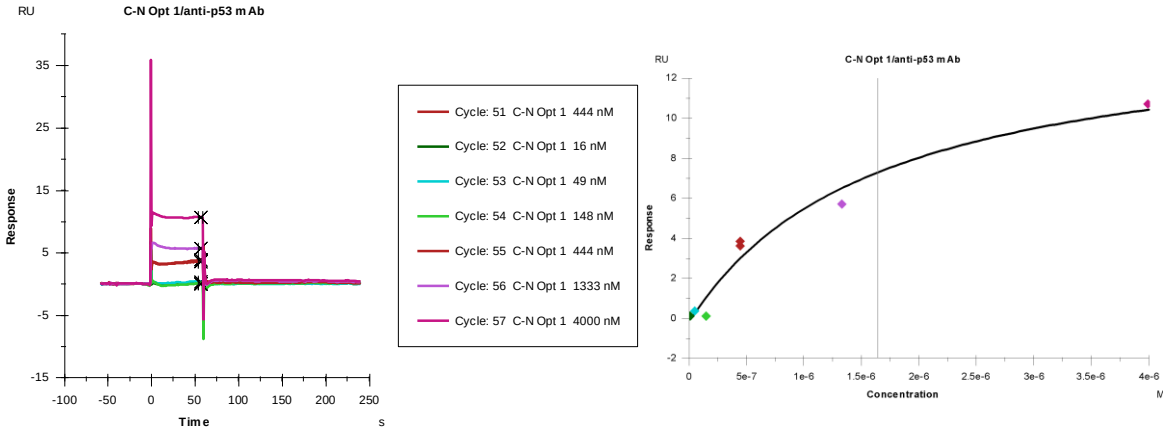


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.46E-06	14.17	0.4974	0.174



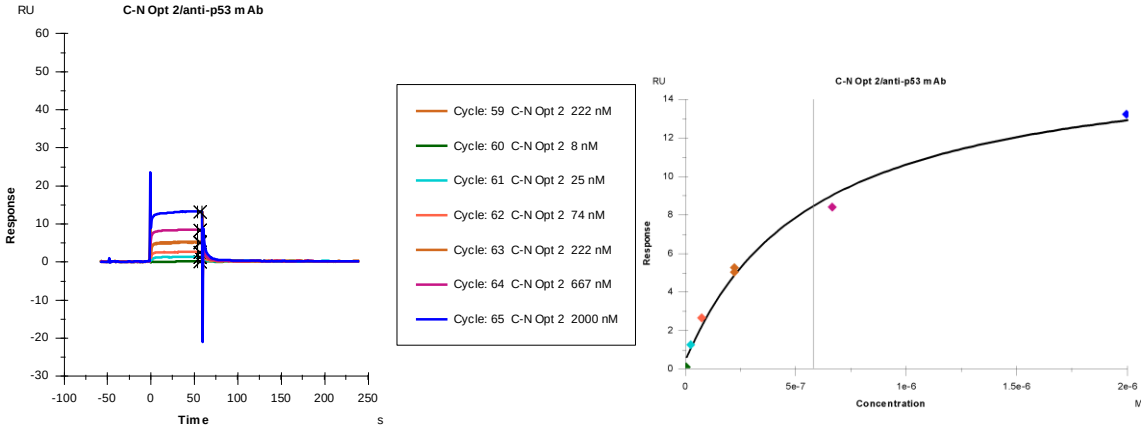
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.96E-06	14.12	0.5223	0.345

Peptide 7



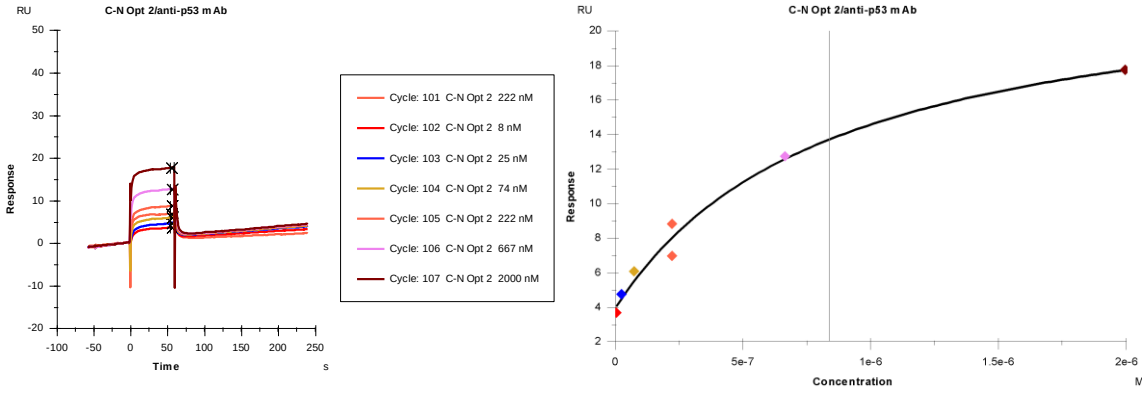
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
1.65E-06	14.98	-0.1786	0.675

Peptide 8

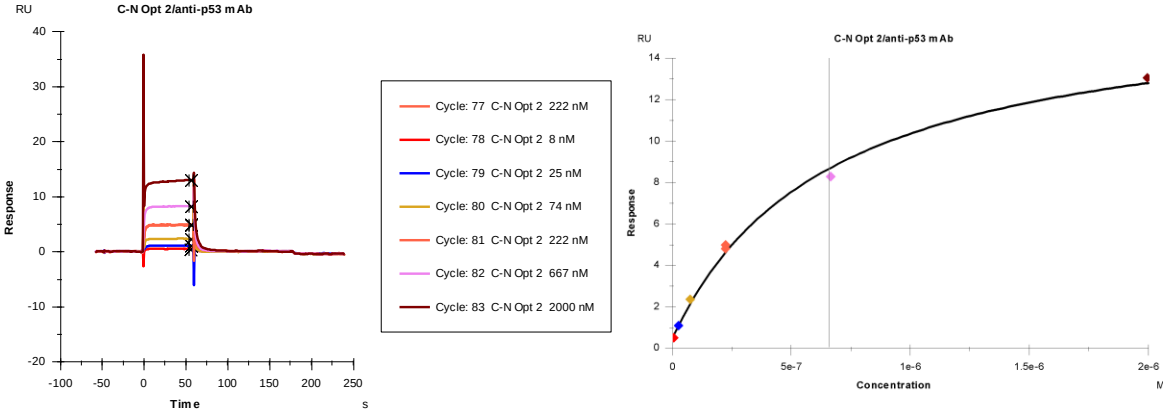


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
5.80E-07	16.14	0.4079	0.281

Peptide 8

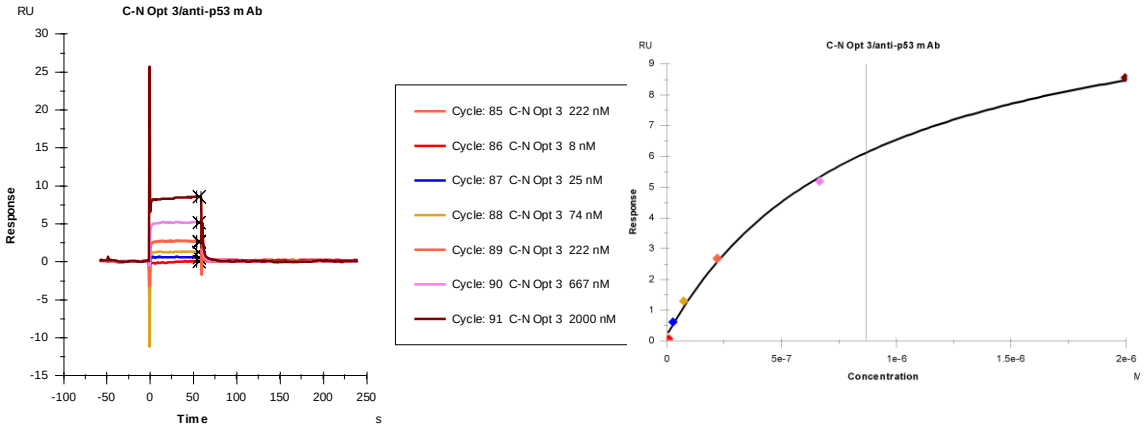


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
8.41E-07	19.63	3.928	0.572

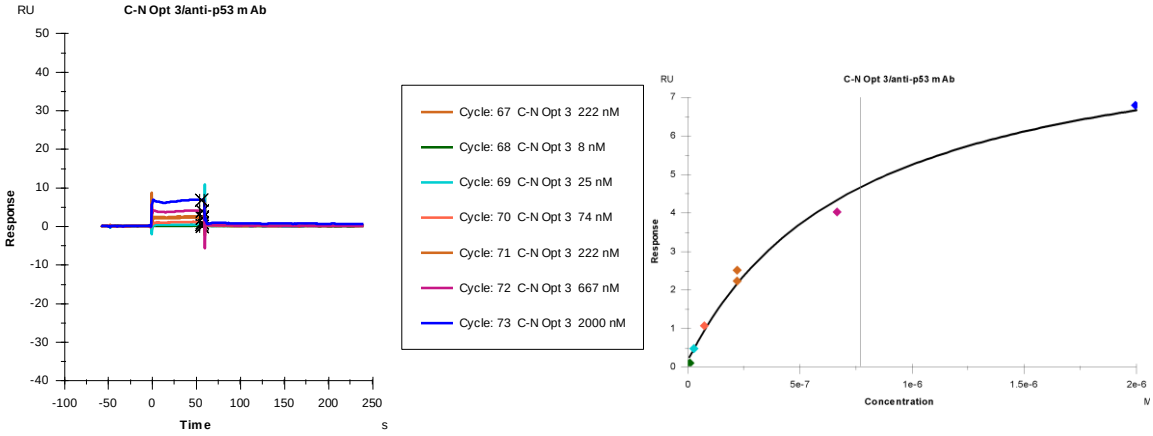


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
6.58E-07	16.37	0.4861	0.122

Peptide 9

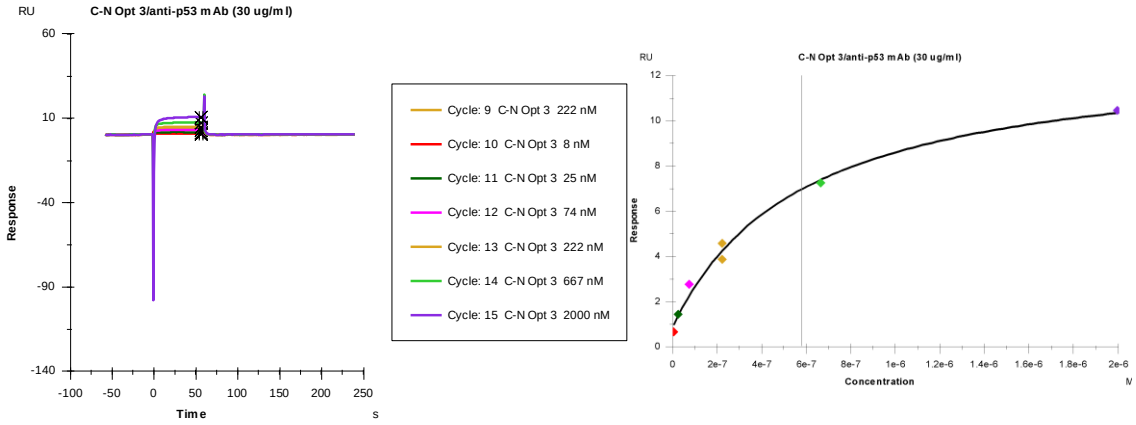


KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
8.70E-07	11.9	0.1769	0.0303



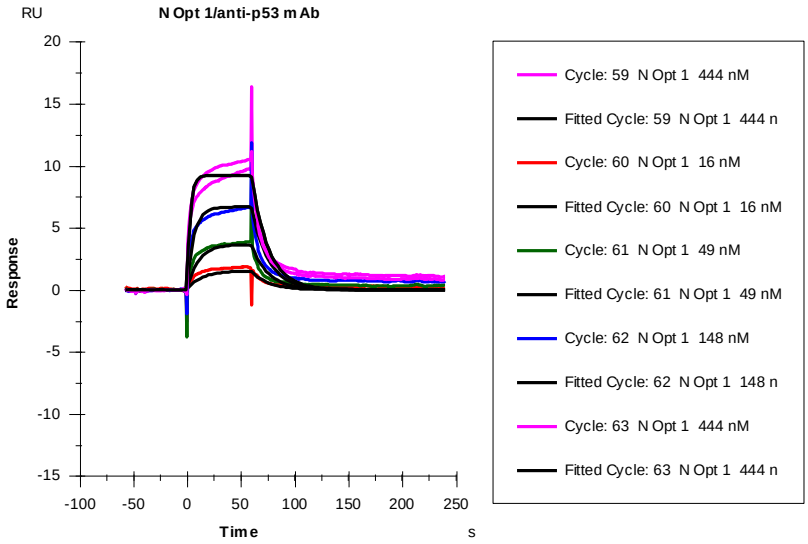
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
7.69E-07	9.003	0.1683	0.0675

Peptide 9



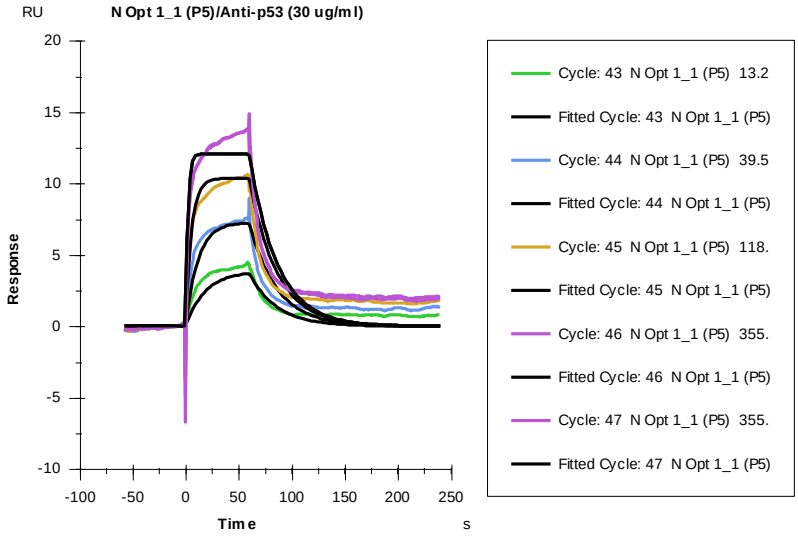
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
5.80E-07	12.26	0.8472	0.177

Peptide 10

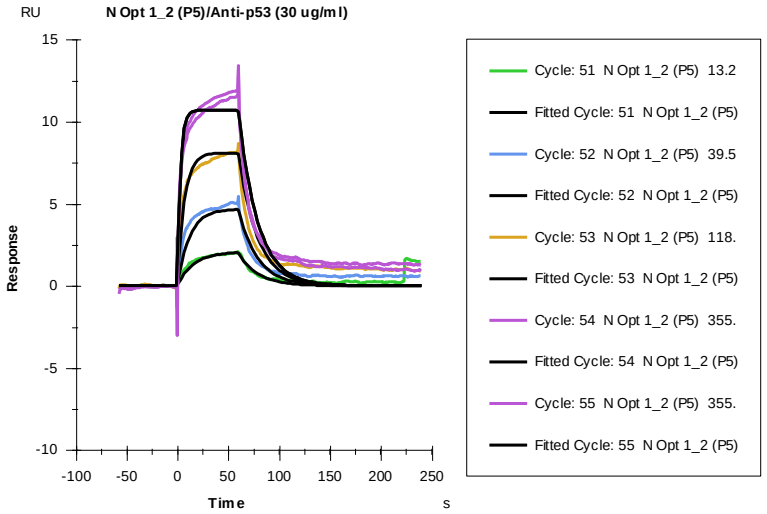


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.24E+05	0.06488	1.04E-07	11.35		8.35E+13				0.406	7
Cycle: 59 444 nM					4.44E-07		30	2.60E+14	0		
Cycle: 60 16 nM					1.60E-08		30	2.60E+14	0		
Cycle: 61 49 nM					4.90E-08		30	2.60E+14	0		
Cycle: 62 148 nM					1.48E-07		30	2.60E+14	0		
Cycle: 63 444 nM					4.44E-07		30	2.60E+14	0		

Peptide 10

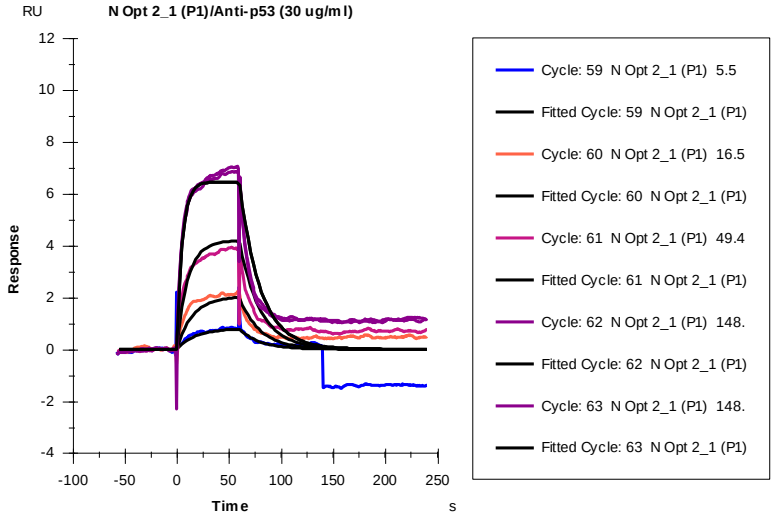


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.21E+06	0.03847	3.19E-08	13.16		9.14E+11				1.24	9
Cycle: 43 13.2 nM					1.32E-08		30	2.84E+12	0		
Cycle: 44 39.5 nM					3.95E-08		30	2.84E+12	0		
Cycle: 45 118.4 nM					1.18E-07		30	2.84E+12	0		
Cycle: 46 355.3 nM					3.55E-07		30	2.84E+12	0		
Cycle: 47 355.3 nM					3.55E-07		30	2.84E+12	0		

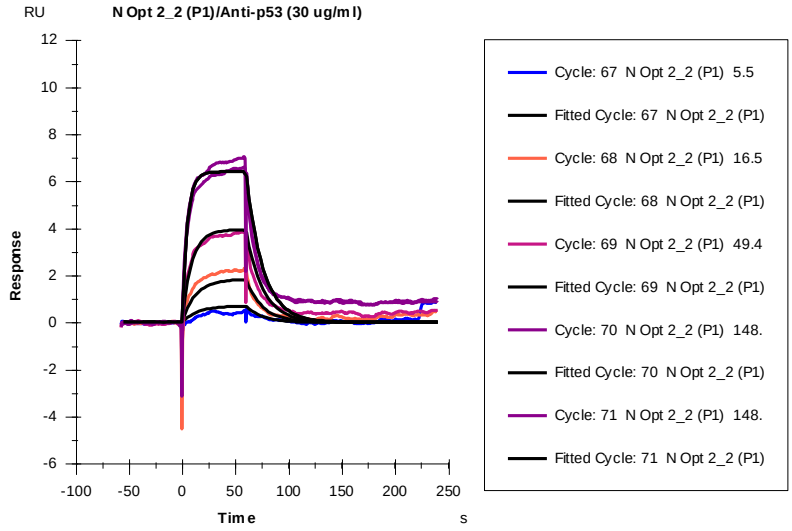


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.81E+05	0.05322	6.81E-08	12.75		1.62E+11				0.53	7
Cycle: 51 13.2 nM					1.32E-08		30	5.02E+11	0		
Cycle: 52 39.5 nM					3.95E-08		30	5.02E+11	0		
Cycle: 53 118.4 nM					1.18E-07		30	5.02E+11	0		
Cycle: 54 355.3 nM					3.55E-07		30	5.02E+11	0		
Cycle: 55 355.3 nM					3.55E-07		30	5.02E+11	0		

Peptide 11

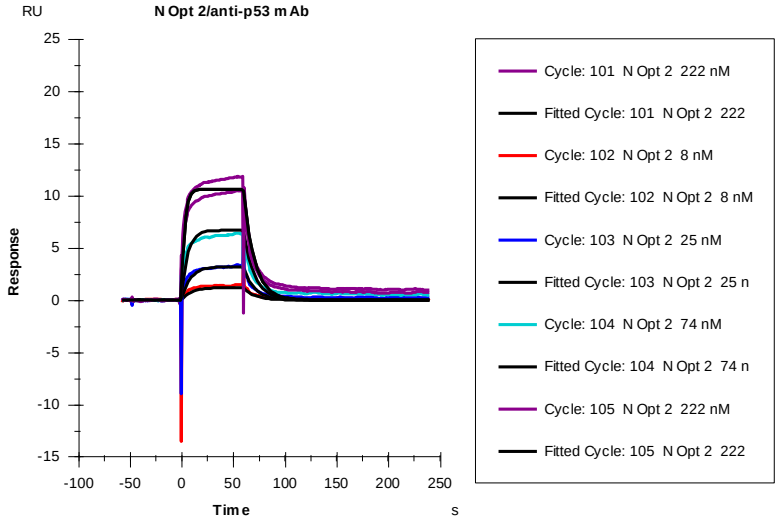


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	9.12E+05	0.05009	5.50E-08	8.855		7.90E+12				0.438	12
Cycle: 59 5.5 nM					5.50E-09		30	2.46E+13	0		
Cycle: 60 16.5 nM					1.65E-08		30	2.46E+13	0		
Cycle: 61 49.4 nM					4.94E-08		30	2.46E+13	0		
Cycle: 62 148.1 nM					1.48E-07		30	2.46E+13	0		
Cycle: 63 148.1 nM					1.48E-07		30	2.46E+13	0		



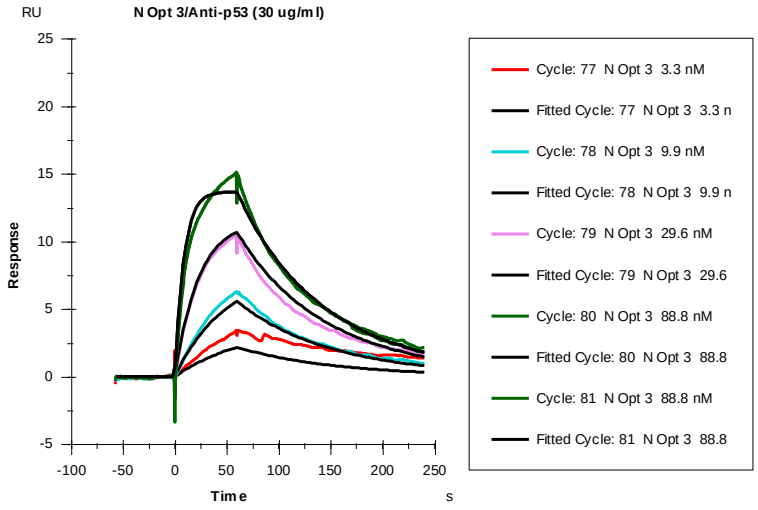
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	9.49E+05	0.06387	6.73E-08	9.312		3.26E+09				0.207	7
Cycle: 67 5.5 nM					5.50E-09		30	1.01E+10	0		
Cycle: 68 16.5 nM					1.65E-08		30	1.01E+10	0		
Cycle: 69 49.4 nM					4.94E-08		30	1.01E+10	0		
Cycle: 70 148.1 nM					1.48E-07		30	1.01E+10	0		
Cycle: 71 148.1 nM					1.48E-07		30	1.01E+10	0		

Peptide 11



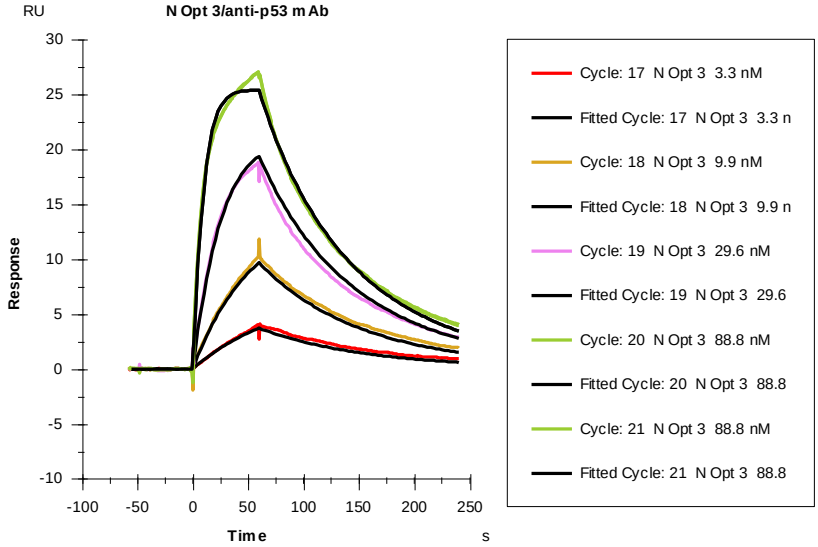
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.02E+06	0.09514	9.36E-08	15.11		2.84E+11				0.396	7
Cycle: 101 222 nM					2.22E-07		30	8.83E+11	0		
Cycle: 102 8 nM					8.00E-09		30	8.83E+11	0		
Cycle: 103 25 nM					2.50E-08		30	8.83E+11	0		
Cycle: 104 74 nM					7.40E-08		30	8.83E+11	0		
Cycle: 105 222 nM					2.22E-07		30	8.83E+11	0		

Peptide 12

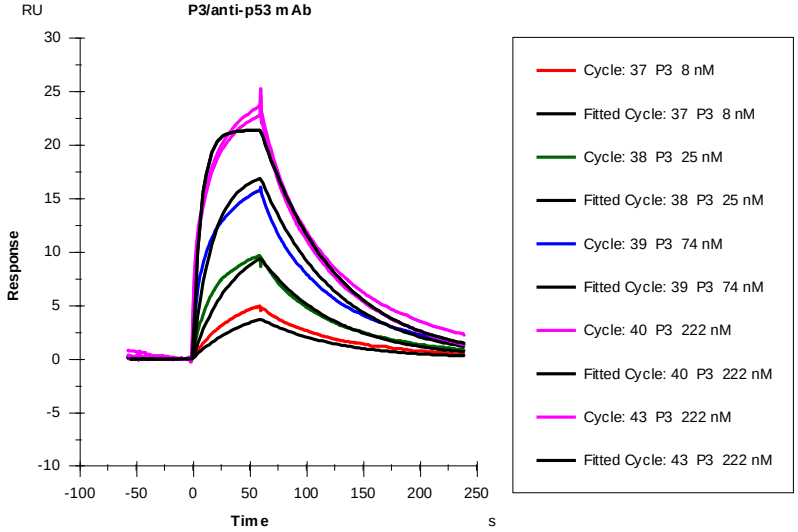


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.31E+06	0.01286	9.84E-09	15.18		2.71E+07				0.349	3
Cycle: 77 3.3 nM					3.30E-09		30	8.43E+07	0		
Cycle: 78 9.9 nM					9.90E-09		30	8.43E+07	0		
Cycle: 79 29.6 nM					2.96E-08		30	8.43E+07	0		
Cycle: 80 88.8 nM					8.88E-08		30	8.43E+07	0		
Cycle: 81 88.8 nM					8.88E-08		30	8.43E+07	0		

Peptide 12

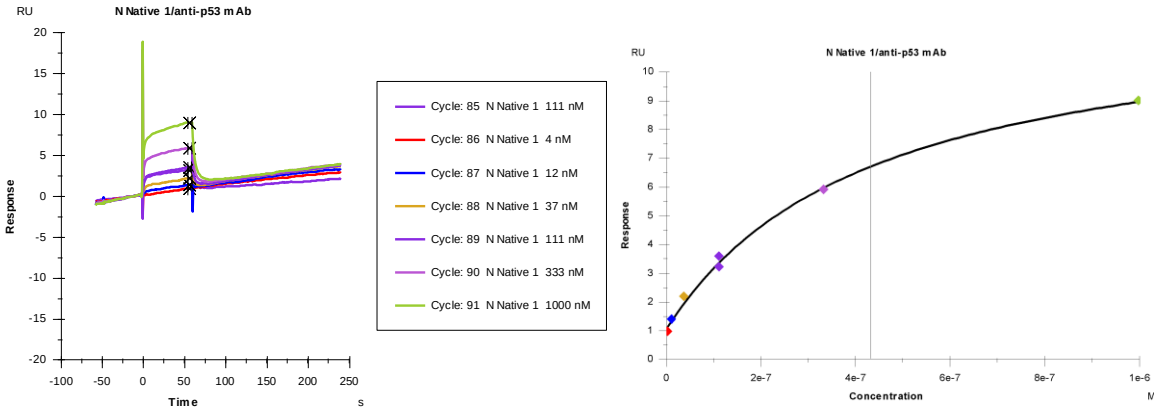


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.29E+06	0.01364	1.06E-08	28.48		3.15E+07				0.226	1
Cycle: 17 3.3 nM					3.30E-09		30	9.79E+07	0		
Cycle: 18 9.9 nM					9.90E-09		30	9.79E+07	0		
Cycle: 19 29.6 nM					2.96E-08		30	9.79E+07	0		
Cycle: 20 88.8 nM					8.88E-08		30	9.79E+07	0		
Cycle: 21 88.8 nM					8.88E-08		30	9.79E+07	0		

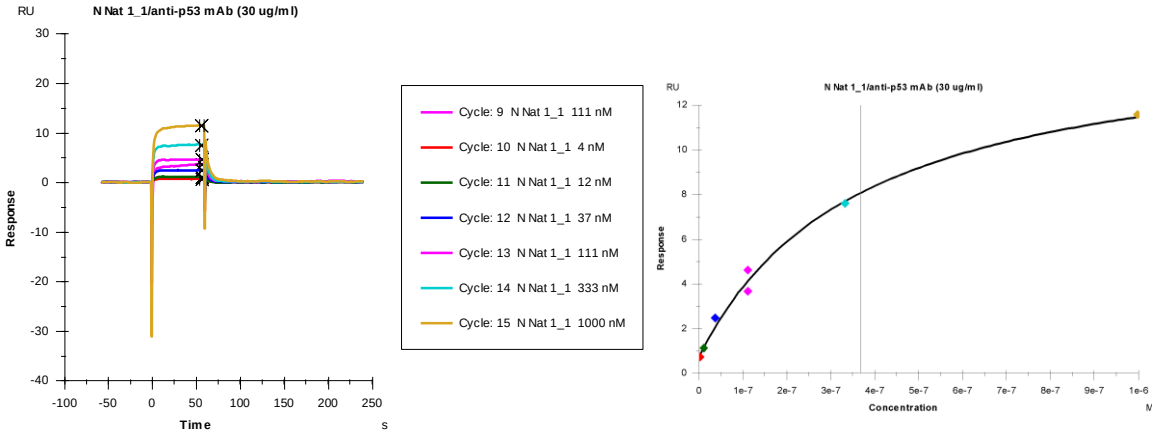


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	5.45E+05	0.01507	2.77E-08	24.01		8.40E+19				0.513	2
Cycle: 37 8 nM					8.00E-09		30	2.61E+20	0		
Cycle: 38 25 nM					2.50E-08		30	2.61E+20	0		
Cycle: 39 74 nM					7.40E-08		30	2.61E+20	0		
Cycle: 40 222 nM					2.22E-07		30	2.61E+20	0		
Cycle: 43 222 nM					2.22E-07		30	2.61E+20	0		

Peptide 13

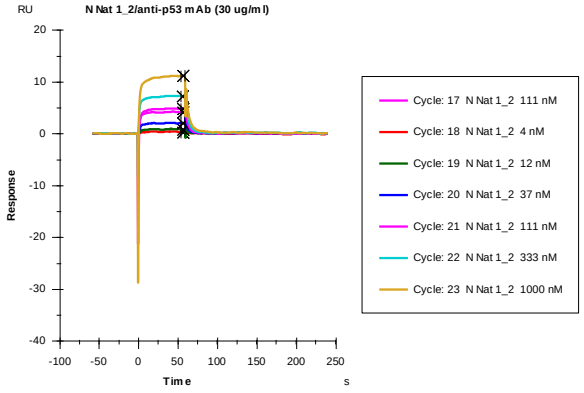


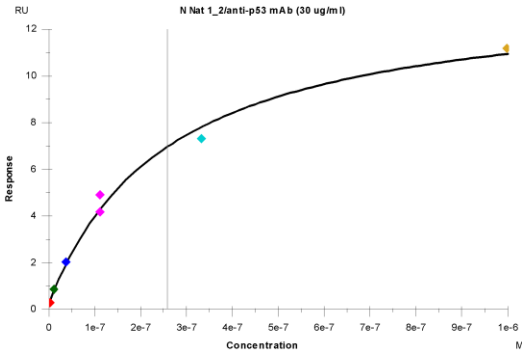
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
4.33E-07	11.35	1.043	0.0449



KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
3.67E-07	14.72	0.7108	0.162

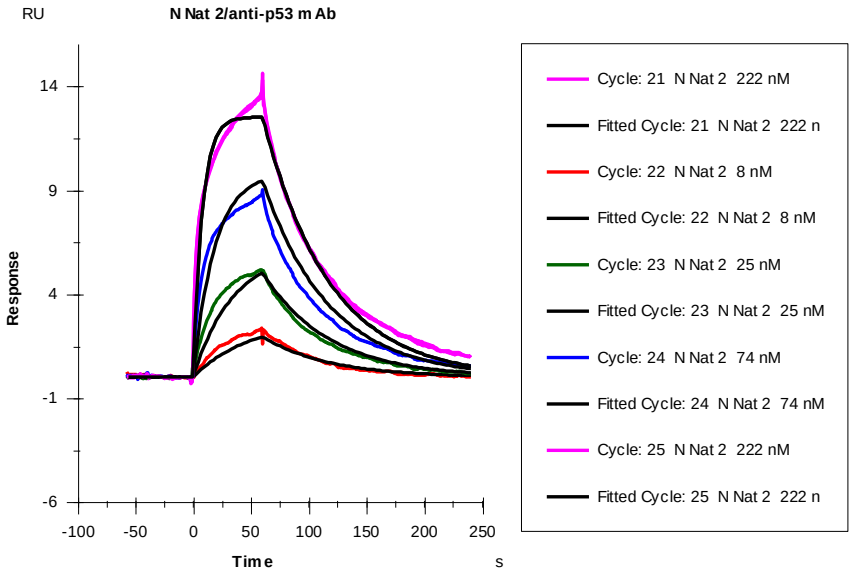
Peptide 13





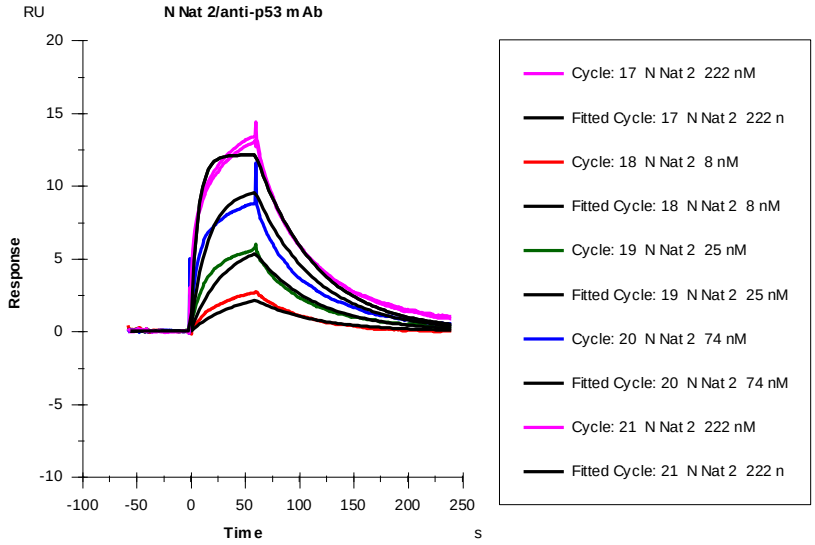
KD (M)	Rmax (RU)	offset (RU)	Chi² (RU²)
2.59E-07	13.5	0.2277	0.185

Peptide 14

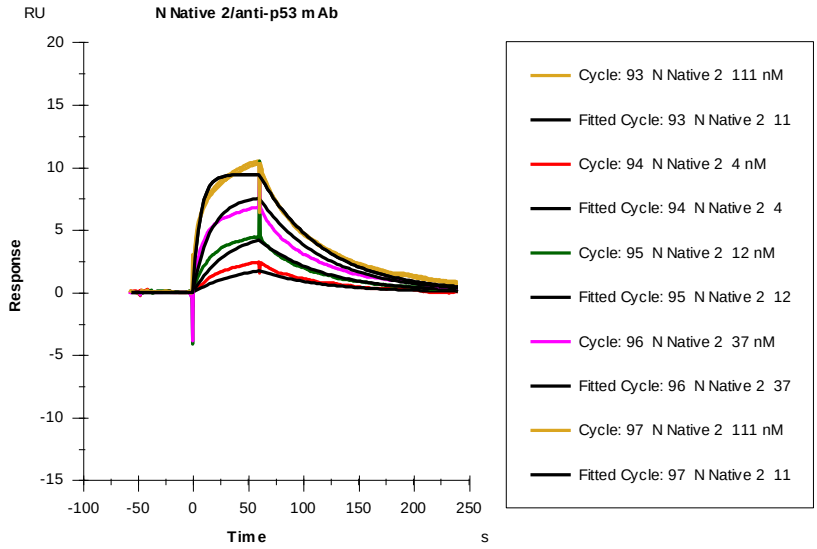


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	4.96E+05	0.0174	3.51E-08	14.47		6.06E+18				0.187	2
Cycle: 21 222 nM					2.22E-07		30	1.88E+19	0		
Cycle: 22 8 nM					8.00E-09		30	1.88E+19	0		
Cycle: 23 25 nM					2.50E-08		30	1.88E+19	0		
Cycle: 24 74 nM					7.40E-08		30	1.88E+19	0		
Cycle: 25 222 nM					2.22E-07		30	1.88E+19	0		

Peptide 14

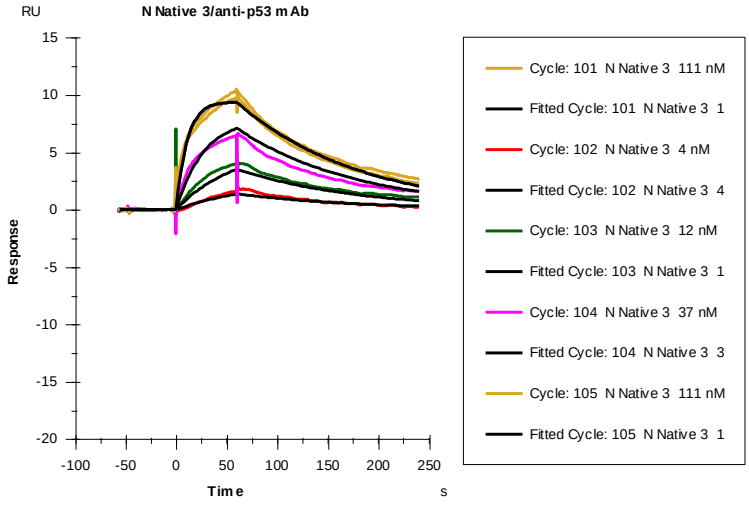


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	5.91E+05	0.01799	3.05E-08	13.78		1.57E+17				0.216	2
Cycle: 17 222 nM					2.22E-07		30	4.88E+17	0		
Cycle: 18 8 nM					8.00E-09		30	4.88E+17	0		
Cycle: 19 25 nM					2.50E-08		30	4.88E+17	0		
Cycle: 20 74 nM					7.40E-08		30	4.88E+17	0		
Cycle: 21 222 nM					2.22E-07		30	4.88E+17	0		

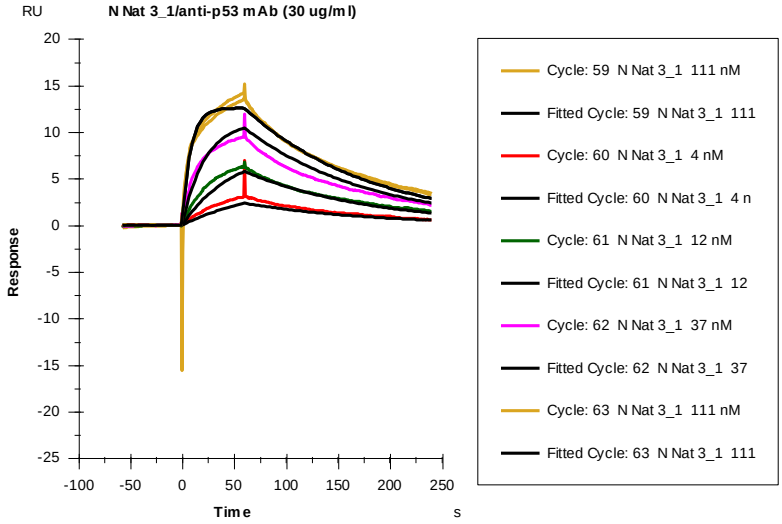


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.22E+06	0.01697	1.39E-08	10.59		2.37E+14				0.179	3
Cycle: 93 111 nM					1.11E-07		30	7.35E+14	0		
Cycle: 94 4 nM					4.00E-09		30	7.35E+14	0		
Cycle: 95 12 nM					1.20E-08		30	7.35E+14	0		
Cycle: 96 37 nM					3.70E-08		30	7.35E+14	0		
Cycle: 97 111 nM					1.11E-07		30	7.35E+14	0		

Peptide 15

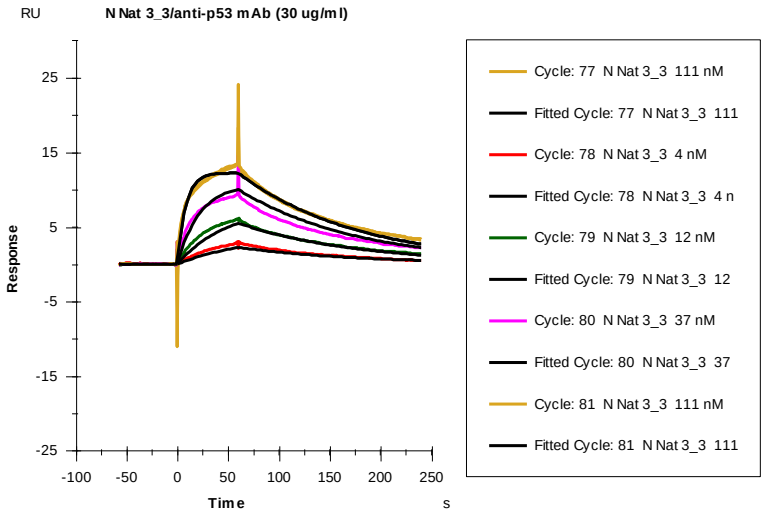


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.66E+05	0.008381	1.09E-08	10.34		2.03E+19				0.179	2
Cycle: 101 111 nM					1.11E-07		30	6.31E+19	0		
Cycle: 102 4 nM					4.00E-09		30	6.31E+19	0		
Cycle: 103 12 nM					1.20E-08		30	6.31E+19	0		
Cycle: 104 37 nM					3.70E-08		30	6.31E+19	0		
Cycle: 105 111 nM					1.11E-07		30	6.31E+19	0		



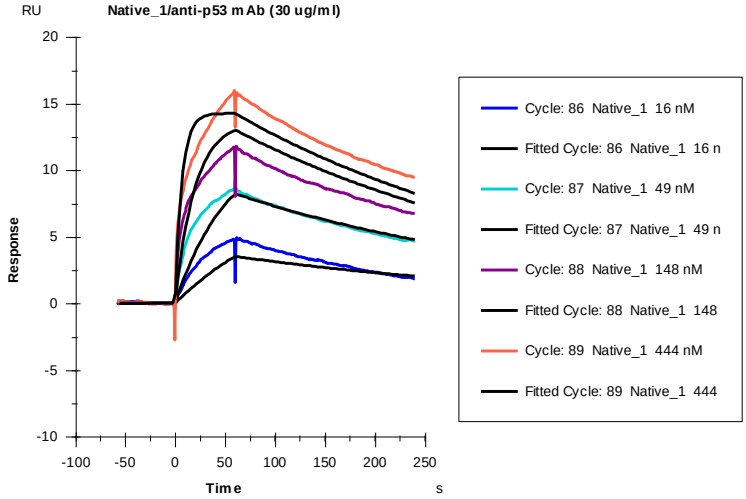
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.04E+06	0.008222	7.92E-09	13.46		3.28E+14				0.345	3
Cycle: 59 111 nM					1.11E-07		30	1.02E+15	0		
Cycle: 60 4 nM					4.00E-09		30	1.02E+15	0		
Cycle: 61 12 nM					1.20E-08		30	1.02E+15	0		
Cycle: 62 37 nM					3.70E-08		30	1.02E+15	0		
Cycle: 63 111 nM					1.11E-07		30	1.02E+15	0		

Peptide 15



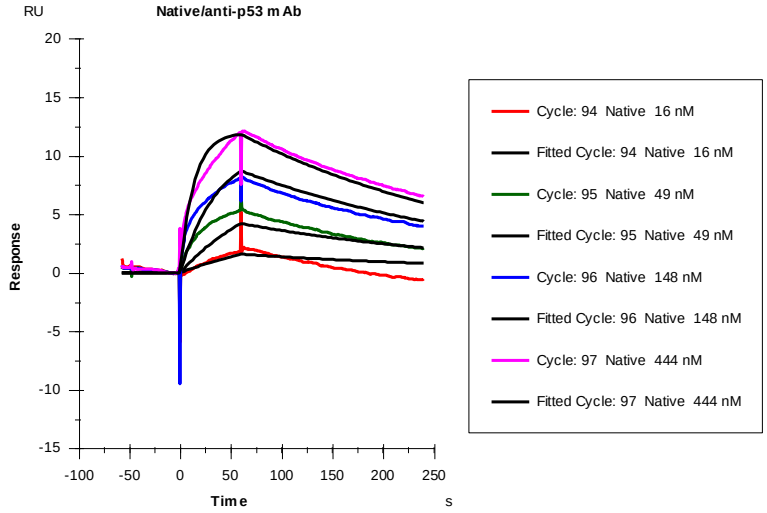
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.00E+06	0.008449	8.41E-09	13.13		3.44E+19				0.309	3
Cycle: 77 111 nM					1.11E-07		30	1.07E+20	0		
Cycle: 78 4 nM					4.00E-09		30	1.07E+20	0		
Cycle: 79 12 nM					1.20E-08		30	1.07E+20	0		
Cycle: 80 37 nM					3.70E-08		30	1.07E+20	0		
Cycle: 81 111 nM					1.11E-07		30	1.07E+20	0		

Peptide 16

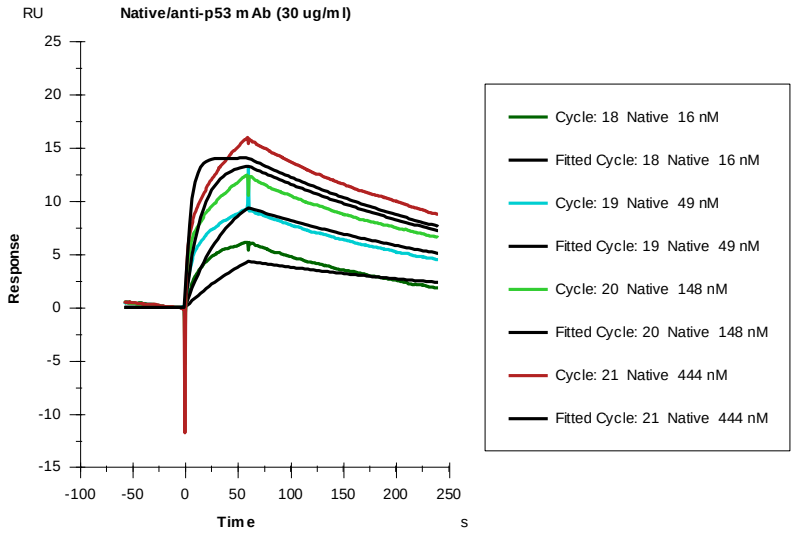


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	3.20E+05	0.003024	9.46E-09	14.57		3.14E+14				0.92	4
Cycle: 86 16 nM					1.60E-08		30	9.75E+14	0		
Cycle: 87 49 nM					4.90E-08		30	9.75E+14	0		
Cycle: 88 148 nM					1.48E-07		30	9.75E+14	0		
Cycle: 89 444 nM					4.44E-07		30	9.75E+14	0		

Peptide 16

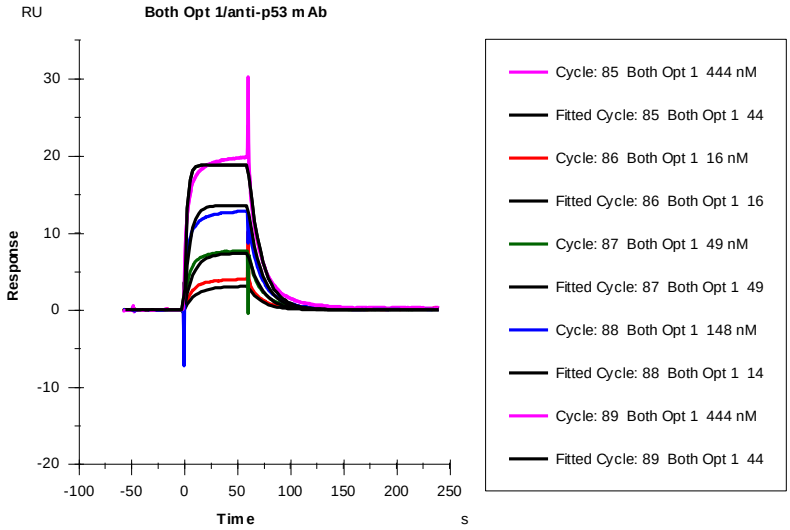


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.57E+05	0.003793	2.42E-08	12.62		1.88E+16				0.503	4
Cycle: 94 16 nM					1.60E-08		30	5.84E+16	0		
Cycle: 95 49 nM					4.90E-08		30	5.84E+16	0		
Cycle: 96 148 nM					1.48E-07		30	5.84E+16	0		
Cycle: 97 444 nM					4.44E-07		30	5.84E+16	0		

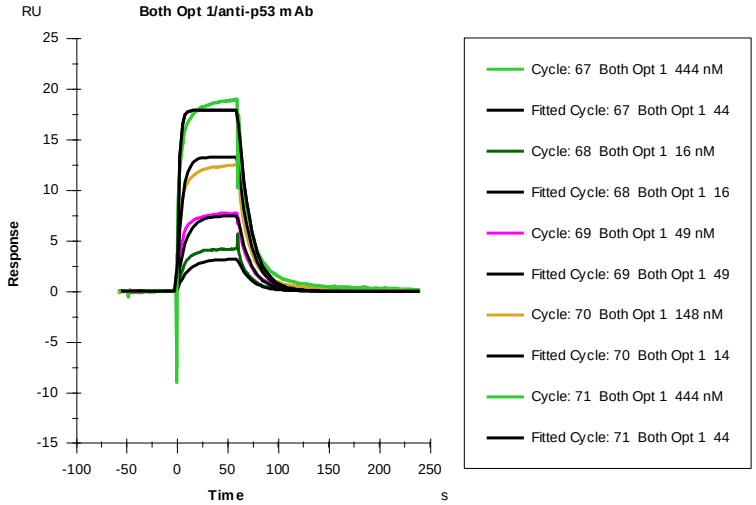


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	4.21E+05	0.0034	8.08E-09	14.28		4.33E+14				1.24	4
Cycle: 18 16 nM					1.60E-08		30	1.34E+15	0		
Cycle: 19 49 nM					4.90E-08		30	1.34E+15	0		
Cycle: 20 148 nM					1.48E-07		30	1.34E+15	0		
Cycle: 21 444 nM					4.44E-07		30	1.34E+15	0		

Peptide 17

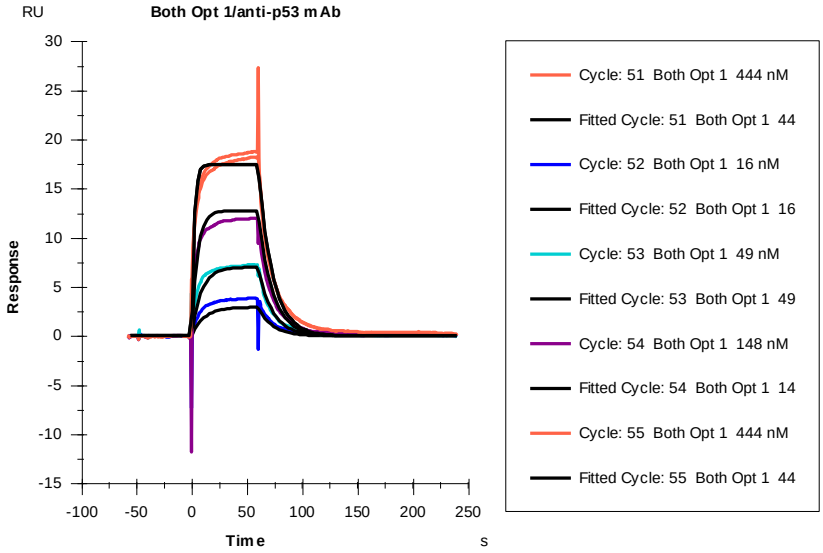


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.31E+05	0.07833	1.07E-07	23.26		1.35E+19				0.302	3
Cycle: 85 444 nM					4.44E-07		30	4.21E+19	0		
Cycle: 86 16 nM					1.60E-08		30	4.21E+19	0		
Cycle: 87 49 nM					4.90E-08		30	4.21E+19	0		
Cycle: 88 148 nM					1.48E-07		30	4.21E+19	0		
Cycle: 89 444 nM					4.44E-07		30	4.21E+19	0		



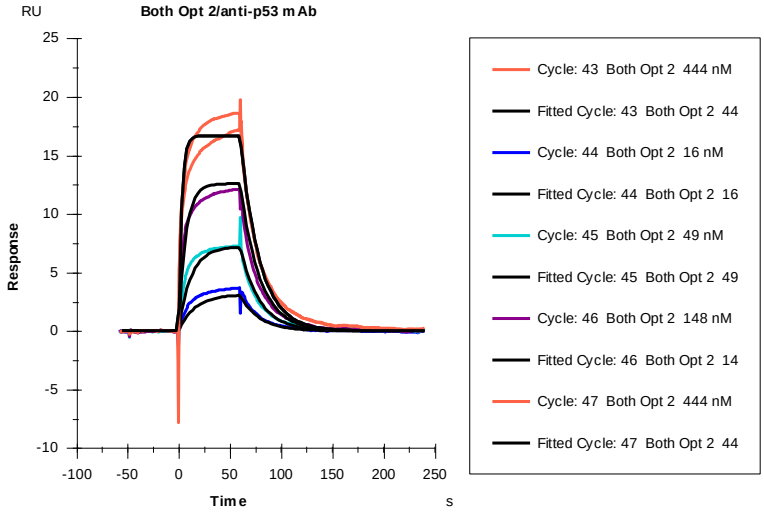
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	8.38E+05	0.07853	9.38E-08	21.64		6.19E+18				0.303	3
Cycle: 67 444 nM					4.44E-07		30	1.92E+19	0		
Cycle: 68 16 nM					1.60E-08		30	1.92E+19	0		
Cycle: 69 49 nM					4.90E-08		30	1.92E+19	0		
Cycle: 70 148 nM					1.48E-07		30	1.92E+19	0		
Cycle: 71 444 nM					4.44E-07		30	1.92E+19	0		

Peptide 17



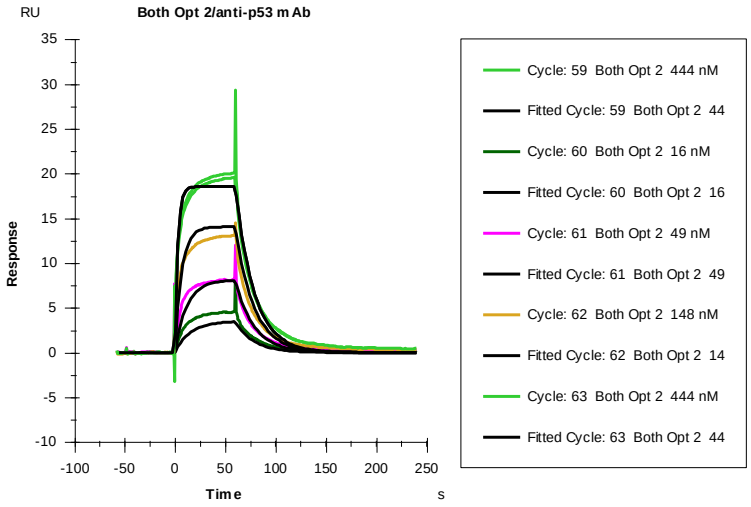
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.69E+05	0.07779	1.01E-07	21.43		1.16E+19				0.288	3
Cycle: 51 444 nM					4.44E-07		30	3.60E+19	0		
Cycle: 52 16 nM					1.60E-08		30	3.60E+19	0		
Cycle: 53 49 nM					4.90E-08		30	3.60E+19	0		
Cycle: 54 148 nM					1.48E-07		30	3.60E+19	0		
Cycle: 55 444 nM					4.44E-07		30	3.60E+19	0		

Peptide 18

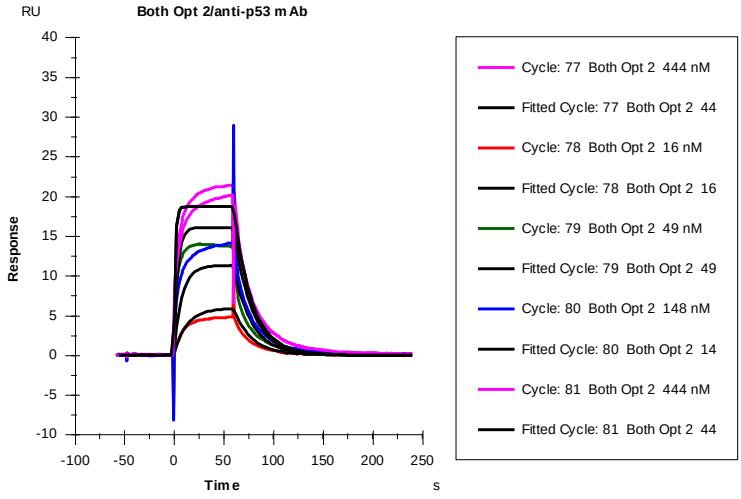


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.16E+05	0.0534	8.67E-08	19.94		4.98E+19				0.282	2
Cycle: 43 444 nM					4.44E-07		30	1.55E+20	0		
Cycle: 44 16 nM					1.60E-08		30	1.55E+20	0		
Cycle: 45 49 nM					4.90E-08		30	1.55E+20	0		
Cycle: 46 148 nM					1.48E-07		30	1.55E+20	0		
Cycle: 47 444 nM					4.44E-07		30	1.55E+20	0		

Peptide 18

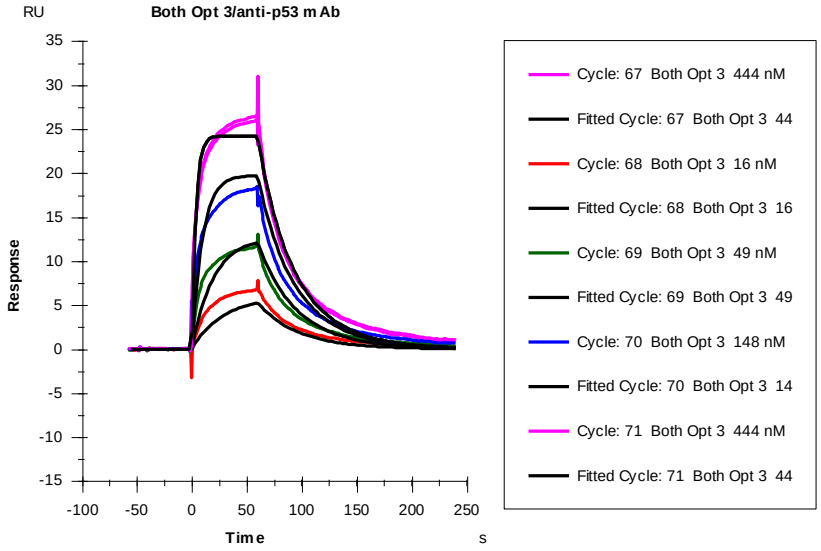


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.38E+05	0.05386	8.44E-08	22.06		6.29E+18				0.394	3
Cycle: 59 444 nM					4.44E-07		30	1.95E+19	0		
Cycle: 60 16 nM					1.60E-08		30	1.95E+19	0		
Cycle: 61 49 nM					4.90E-08		30	1.95E+19	0		
Cycle: 62 148 nM					1.48E-07		30	1.95E+19	0		
Cycle: 63 444 nM					4.44E-07		30	1.95E+19	0		

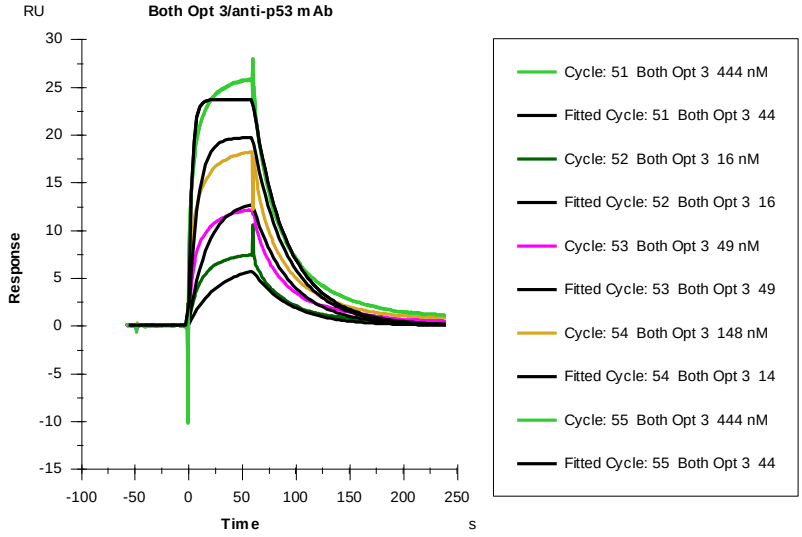


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.40E+06	0.05462	3.91E-08	20.29		2.06E+17				1.74	7
Cycle: 77 444 nM					4.44E-07		30	6.39E+17	0		
Cycle: 78 16 nM					1.60E-08		30	6.39E+17	0		
Cycle: 79 49 nM					4.90E-08		30	6.39E+17	0		
Cycle: 80 148 nM					1.48E-07		30	6.39E+17	0		
Cycle: 81 444 nM					4.44E-07		30	6.39E+17	0		

Peptide 19

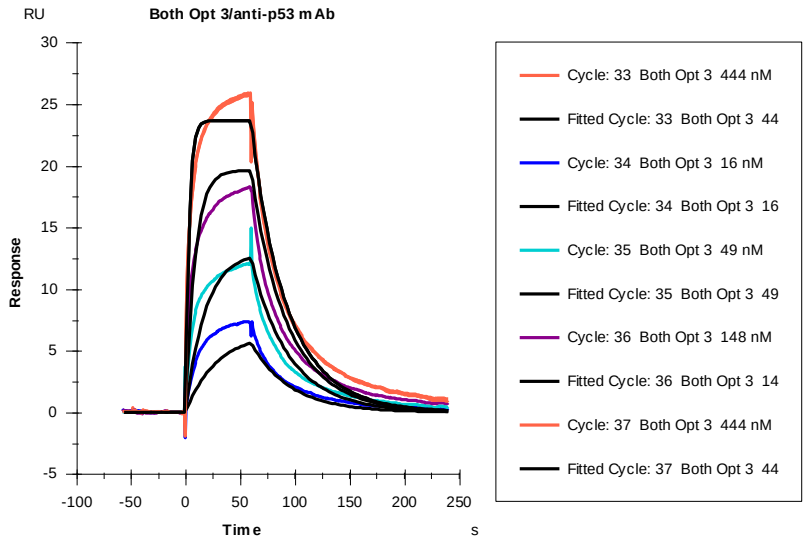


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	5.94E+05	0.0338	5.69E-08	27.27		1.92E+07				0.893	3
Cycle: 67 444 nM					4.44E-07		30	5.97E+07	0		
Cycle: 68 16 nM					1.60E-08		30	5.97E+07	0		
Cycle: 69 49 nM					4.90E-08		30	5.97E+07	0		
Cycle: 70 148 nM					1.48E-07		30	5.97E+07	0		
Cycle: 71 444 nM					4.44E-07		30	5.97E+07	0		



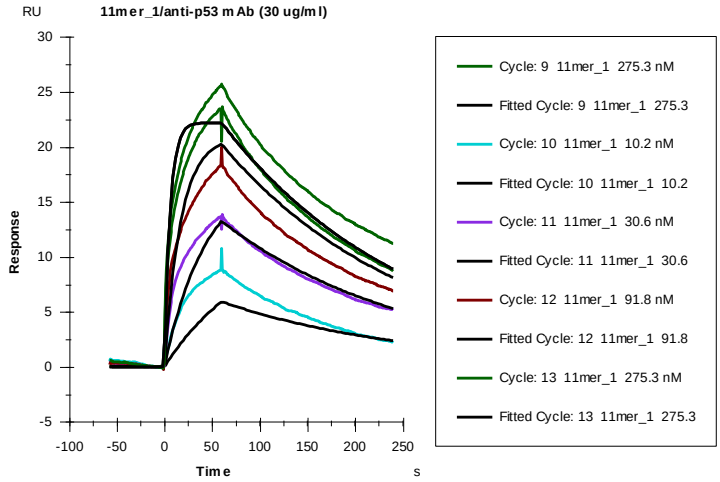
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.01E+05	0.03491	4.98E-08	26.31		2.09E+07				1.05	3
Cycle: 51 444 nM					4.44E-07		30	6.50E+07	0		
Cycle: 52 16 nM					1.60E-08		30	6.50E+07	0		
Cycle: 53 49 nM					4.90E-08		30	6.50E+07	0		
Cycle: 54 148 nM					1.48E-07		30	6.50E+07	0		
Cycle: 55 444 nM					4.44E-07		30	6.50E+07	0		

Peptide 19



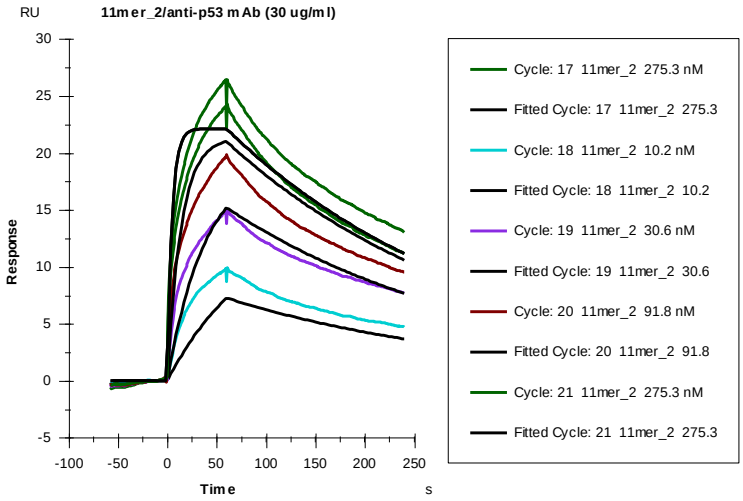
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.94E+05	0.03527	5.08E-08	26.35		1.93E+07				1.08	4
Cycle: 33 444 nM					4.44E-07		30	5.99E+07	0		
Cycle: 34 16 nM					1.60E-08		30	5.99E+07	0		
Cycle: 35 49 nM					4.90E-08		30	5.99E+07	0		
Cycle: 36 148 nM					1.48E-07		30	5.99E+07	0		
Cycle: 37 444 nM					4.44E-07		30	5.99E+07	0		

Peptide 20

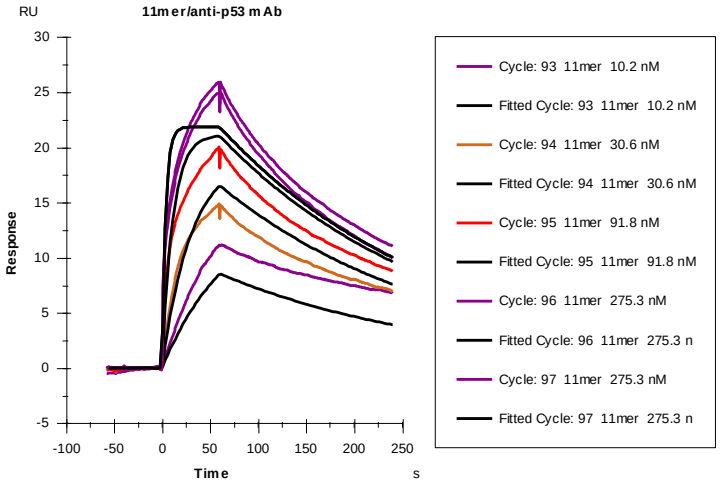


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	5.76E+05	0.005093	8.85E-09	22.9		9.73E+19				2.47	4
Cycle: 9 275.3 nM					2.75E-07		30	3.02E+20	0		
Cycle: 10 10.2 nM					1.02E-08		30	3.02E+20	0		
Cycle: 11 30.6 nM					3.06E-08		30	3.02E+20	0		
Cycle: 12 91.8 nM					9.18E-08		30	3.02E+20	0		
Cycle: 13 275.3 nM					2.75E-07		30	3.02E+20	0		

Peptide 20

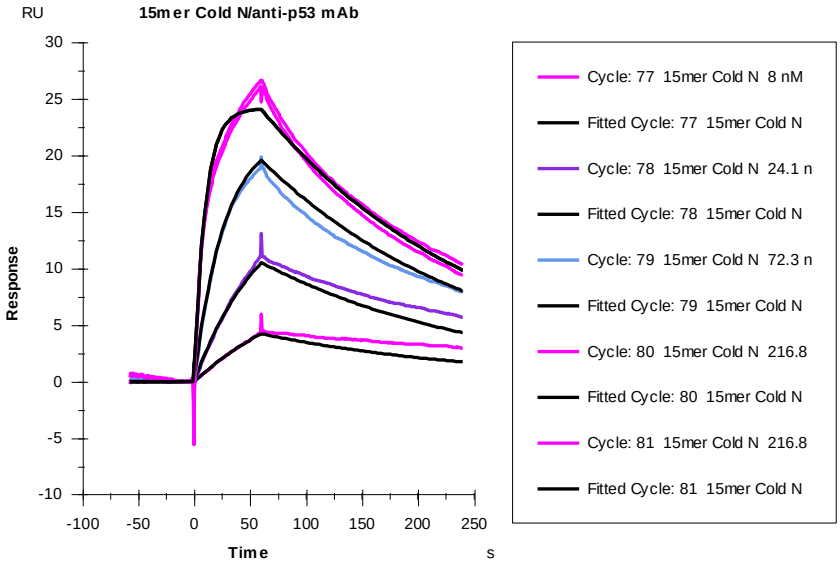


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.23E+05	0.003798	5.25E-09	22.53		4.69E+20				2.59	4
Cycle: 17 275.3 nM					2.75E-07		30	1.46E+21	0		
Cycle: 18 10.2 nM					1.02E-08		30	1.46E+21	0		
Cycle: 19 30.6 nM					3.06E-08		30	1.46E+21	0		
Cycle: 20 91.8 nM					9.18E-08		30	1.46E+21	0		
Cycle: 21 275.3 nM					2.75E-07		30	1.46E+21	0		

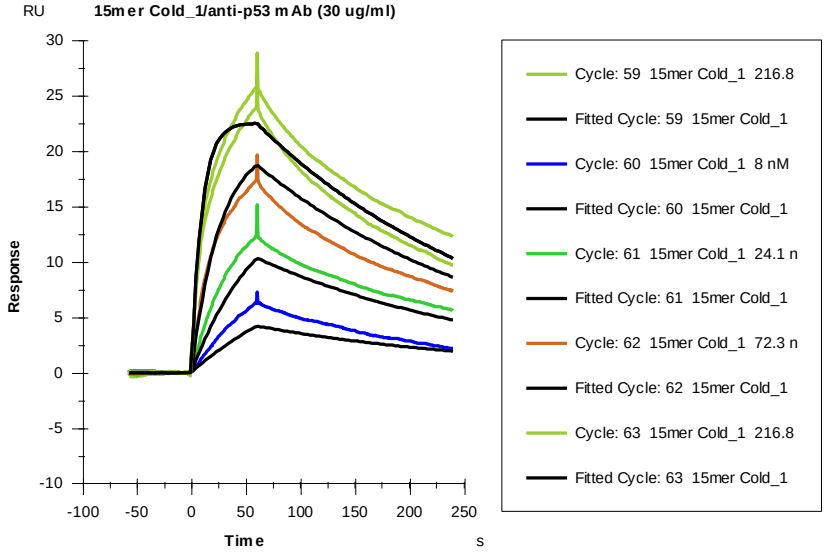


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	9.23E+05	0.004335	4.70E-09	22.2		3.72E+13				2.94	5
Cycle: 93 10.2 nM					1.02E-08		30	1.16E+14	0		
Cycle: 94 30.6 nM					3.06E-08		30	1.16E+14	0		
Cycle: 95 91.8 nM					9.18E-08		30	1.16E+14	0		
Cycle: 96 275.3 nM					2.75E-07		30	1.16E+14	0		
Cycle: 97 275.3 nM					2.75E-07		30	1.16E+14	0		

Peptide 21

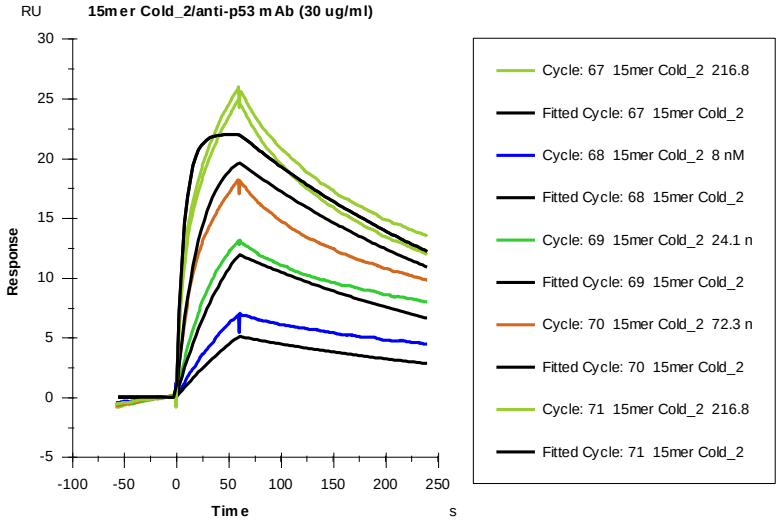


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	4.41E+05	0.004978	1.13E-08	25.4		2.54E+14				0.602	2
Cycle: 77 8 nM					8.00E-09		30	7.90E+14	0		
Cycle: 78 24.1 nM					2.41E-08		30	7.90E+14	0		
Cycle: 79 72.3 nM					7.23E-08		30	7.90E+14	0		
Cycle: 80 216.8 nM					2.17E-07		30	7.90E+14	0		
Cycle: 81 216.8 nM					2.17E-07		30	7.90E+14	0		



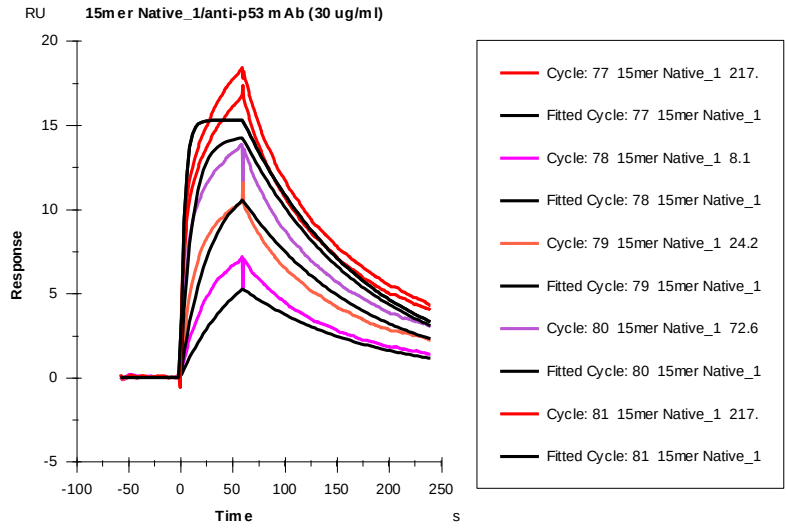
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	4.69E+05	0.004324	9.21E-09	23.5		4.98E+14				1.73	3
Cycle: 59 216.8 nM					2.17E-07		30	1.55E+15	0		
Cycle: 60 8 nM					8.00E-09		30	1.55E+15	0		
Cycle: 61 24.1 nM					2.41E-08		30	1.55E+15	0		
Cycle: 62 72.3 nM					7.23E-08		30	1.55E+15	0		
Cycle: 63 216.8 nM					2.17E-07		30	1.55E+15	0		

Peptide 21



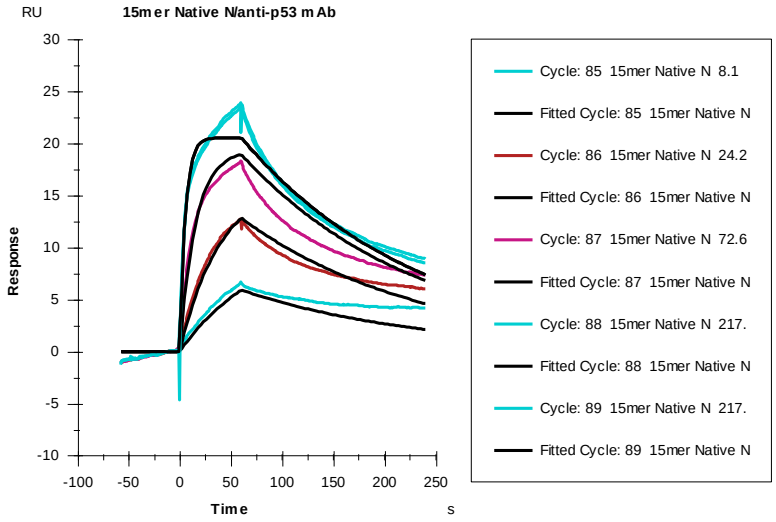
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	5.90E+05	0.003279	5.56E-09	22.58		5.37E+14				1.95	4
Cycle: 67 216.8 nM					2.17E-07		30	1.67E+15	0		
Cycle: 68 8 nM					8.00E-09		30	1.67E+15	0		
Cycle: 69 24.1 nM					2.41E-08		30	1.67E+15	0		
Cycle: 70 72.3 nM					7.23E-08		30	1.67E+15	0		
Cycle: 71 216.8 nM					2.17E-07		30	1.67E+15	0		

Peptide 22

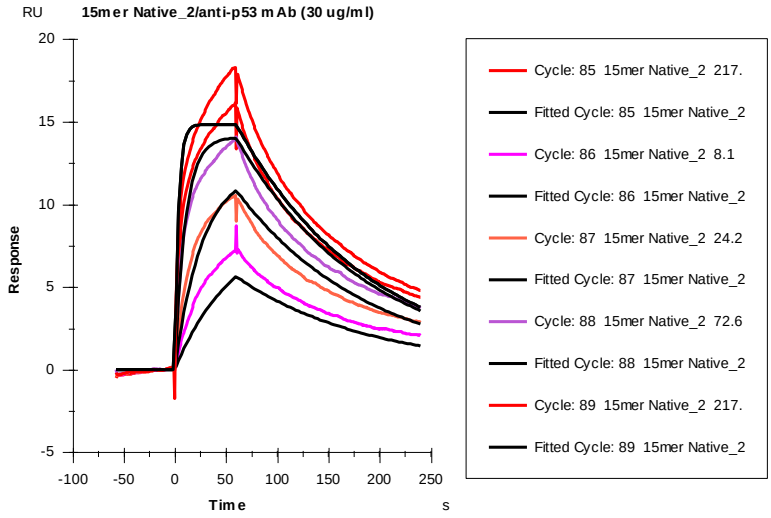


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.10E+06	0.008492	7.70E-09	15.81		2.68E+13				0.804	4
Cycle: 77 217.7 nM					2.18E-07		30	8.33E+13	0		
Cycle: 78 8.1 nM					8.10E-09		30	8.33E+13	0		
Cycle: 79 24.2 nM					2.42E-08		30	8.33E+13	0		
Cycle: 80 72.6 nM					7.26E-08		30	8.33E+13	0		
Cycle: 81 217.7 nM					2.18E-07		30	8.33E+13	0		

Peptide 22

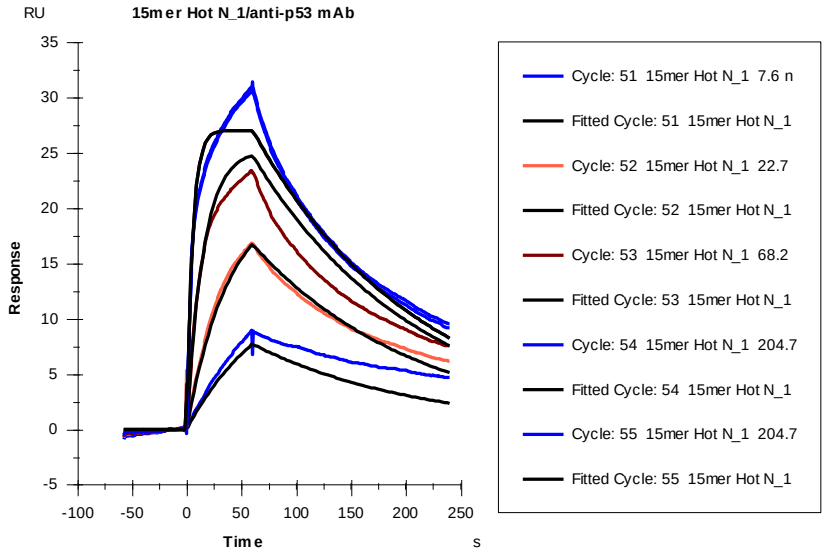


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	8.08E+05	0.005697	7.05E-09	21.23		8.73E+13				1.03	3
Cycle: 85 8.1 nM					8.10E-09		30	2.71E+14	0		
Cycle: 86 24.2 nM					2.42E-08		30	2.71E+14	0		
Cycle: 87 72.6 nM					7.26E-08		30	2.71E+14	0		
Cycle: 88 217.7 nM					2.18E-07		30	2.71E+14	0		
Cycle: 89 217.7 nM					2.18E-07		30	2.71E+14	0		

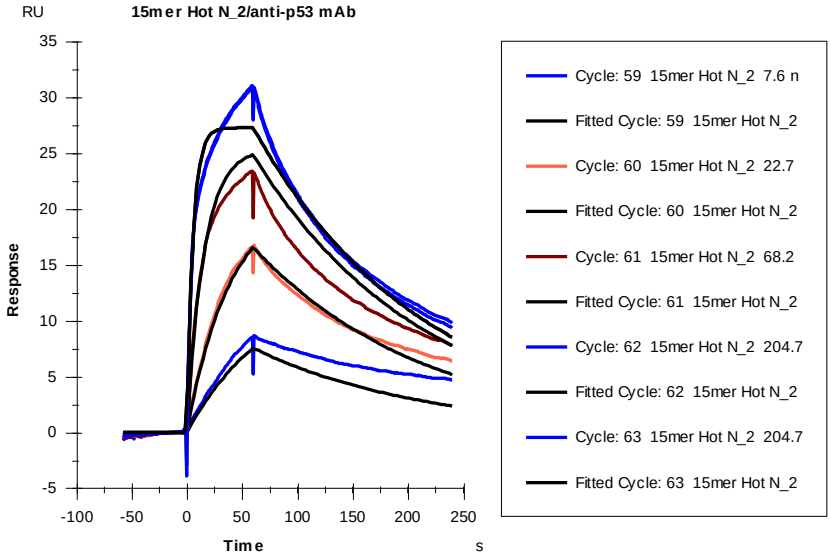


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	1.22E+06	0.007612	6.22E-09	15.25		5.43E+19				0.839	4
Cycle: 85 217.7 nM					2.18E-07		30	1.69E+20	0		
Cycle: 86 8.1 nM					8.10E-09		30	1.69E+20	0		
Cycle: 87 24.2 nM					2.42E-08		30	1.69E+20	0		
Cycle: 88 72.6 nM					7.26E-08		30	1.69E+20	0		
Cycle: 89 217.7 nM					2.18E-07		30	1.69E+20	0		

Peptide 23

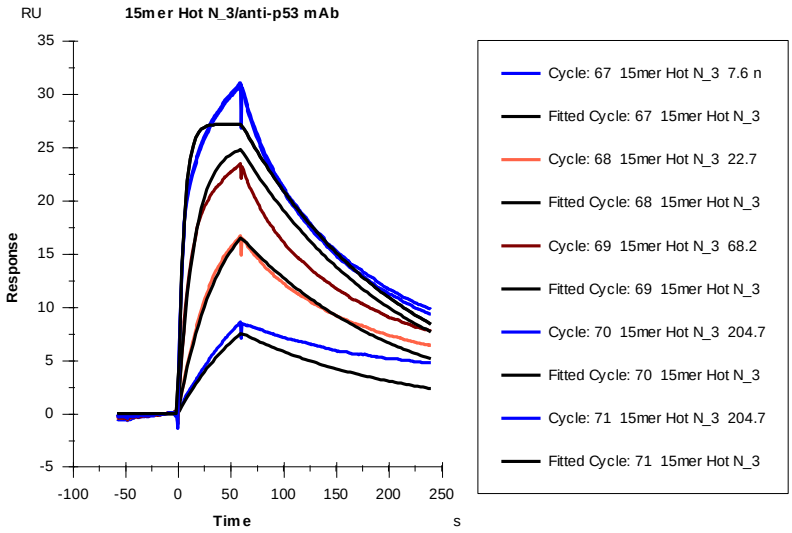


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	8.72E+05	0.006589	7.56E-09	28.04		3.45E+18				1.59	3
Cycle: 51 7.6 nM					7.60E-09		30	1.07E+19	0		
Cycle: 52 22.7 nM					2.27E-08		30	1.07E+19	0		
Cycle: 53 68.2 nM					6.82E-08		30	1.07E+19	0		
Cycle: 54 204.7 nM					2.05E-07		30	1.07E+19	0		
Cycle: 55 204.7 nM					2.05E-07		30	1.07E+19	0		



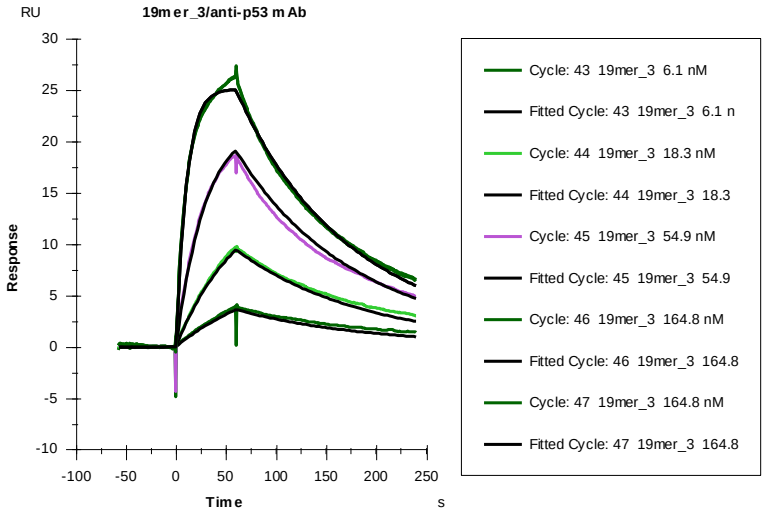
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	8.39E+05	0.006469	7.71E-09	28.3		3.55E+19				1.51	3
Cycle: 59 7.6 nM					7.60E-09		30	1.10E+20	0		
Cycle: 60 22.7 nM					2.27E-08		30	1.10E+20	0		
Cycle: 61 68.2 nM					6.82E-08		30	1.10E+20	0		
Cycle: 62 204.7 nM					2.05E-07		30	1.10E+20	0		
Cycle: 63 204.7 nM					2.05E-07		30	1.10E+20	0		

Peptide 23



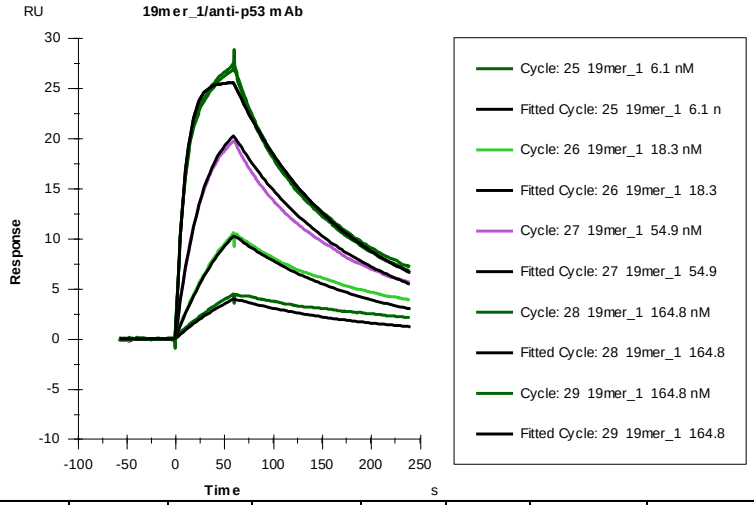
Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	8.42E+05	0.006532	7.76E-09	28.22		6.76E+16				1.55	3
Cycle: 67 7.6 nM					7.60E-09		30	2.10E+17	0		
Cycle: 68 22.7 nM					2.27E-08		30	2.10E+17	0		
Cycle: 69 68.2 nM					6.82E-08		30	2.10E+17	0		
Cycle: 70 204.7 nM					2.05E-07		30	2.10E+17	0		
Cycle: 71 204.7 nM					2.05E-07		30	2.10E+17	0		

Peptide 24

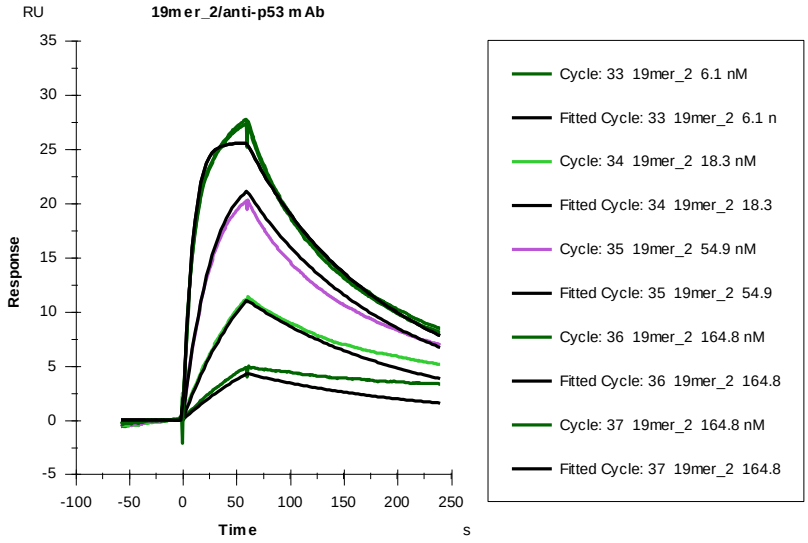


Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.04E+05	0.009157	1.52E-08	27.42		1.94E+07				0.165	1
Cycle: 43 6.1 nM					6.10E-09		30	6.03E+07	0		
Cycle: 44 18.3 nM					1.83E-08		30	6.03E+07	0		
Cycle: 45 54.9 nM					5.49E-08		30	6.03E+07	0		
Cycle: 46 164.8 nM					1.65E-07		30	6.03E+07	0		
Cycle: 47 164.8 nM					1.65E-07		30	6.03E+07	0		

Peptide 24



Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	6.87E+05	0.008795	1.28E-08	27.55		1.76E+07				0.281	1
Cycle: 25 6.1 nM					6.10E-09		30	5.47E+07	0		
Cycle: 26 18.3 nM					1.83E-08		30	5.47E+07	0		
Cycle: 27 54.9 nM					5.49E-08		30	5.47E+07	0		
Cycle: 28 164.8 nM					1.65E-07		30	5.47E+07	0		
Cycle: 29 164.8 nM					1.65E-07		30	5.47E+07	0		



Curve	ka (1/Ms)	kd (1/s)	KD (M)	Rmax (RU)	Conc (M)	tc	Flow (ul/min)	kt (RU/Ms)	RI (RU)	Chi² (RU²)	U-value
	7.94E+05	0.008001	1.01E-08	27.12		1.49E+07				0.58	2
Cycle: 33 6.1 nM					6.10E-09		30	4.62E+07	0		
Cycle: 34 18.3 nM					1.83E-08		30	4.62E+07	0		
Cycle: 35 54.9 nM					5.49E-08		30	4.62E+07	0		
Cycle: 36 164.8 nM					1.65E-07		30	4.62E+07	0		
Cycle: 37 164.8 nM					1.65E-07		30	4.62E+07	0		

References

1. Pedersen, J. W.; Gentry-Maharaj, A.; Fourkala, E. O.; Dawnay, A.; Burnell, M.; Zaikin, A.; Pedersen, A. E.; Jacobs, I.; Menon, U.; Wandall, H. H., Early detection of cancer in the general population: a blinded case-control study of p53 autoantibodies in colorectal cancer. *British journal of cancer* **2013,** *108* (1), 107-14.
2. Macdonald, I. K.; Parsy-Kowalska, C. B.; Chapman, C. J., Autoantibodies: Opportunities for Early Cancer Detection. *Trends in Cancer* **2017,** *3* (3), 198-213.
3. Qin, J.; Zeng, N.; Yang, T.; Wan, C.; Chen, L.; Shen, Y.; Wen, F., Diagnostic Value of Autoantibodies in Lung Cancer: a Systematic Review and Meta-Analysis. *Cellular Physiology and Biochemistry* **2018,** *51* (6), 2631-2646.
4. Williams, L. J.; Schendt, B. J.; Fritz, Z. R.; Attali, Y.; Lavroff, R. H.; Yarmush, M. L., A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme. *Technology (Singap World Sci)* **2019,** *7* (1-2), 12-39.
5. Lee, C. W.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E., Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of CREB binding protein. *Biochemistry* **2010,** *49* (46), 9964-71.
6. Liao, H. X.; Bonsignori, M.; Alam, S. M.; McLellan, J. S.; Tomaras, G. D.; Moody, M. A.; Kozink, D. M.; Hwang, K. K.; Chen, X.; Tsao, C. Y.; Liu, P.; Lu, X.; Parks, R. J.; Montefiori, D. C.; Ferrari, G.; Pollara, J.; Rao, M.; Peachman, K. K.; Santra, S.; Letvin, N. L.; Karasavvas, N.; Yang, Z. Y.; Dai, K.; Pancera, M.; Gorman, J.; Wiehe, K.; Nicely, N. I.; Rerks-Ngarm, S.; Nitayaphan, S.; Kaewkungwal, J.; Pitisuttithum, P.; Tartaglia, J.; Sinangil, F.; Kim, J. H.; Michael, N. L.; Kepler, T. B.; Kwong, P. D.; Mascola, J. R.; Nabel, G. J.; Pinter, A.; Zolla-Pazner, S.; Haynes, B. F., Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity* **2013,** *38* (1), 176-86.
7. van Den Elsen, J. M.; Kuntz, D. A.; Hoedemaeker, F. J.; Rose, D. R., Antibody C219 recognizes an alpha-helical epitope on P-glycoprotein. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (24), 13679-84.
8. Potocnakova, L.; Bhide, M.; Pulzova, L. B., An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction. *Journal of immunology research* **2016,** *2016*, 6760830.
9. Galanis, K. A.; Nastou, K. C.; Papandreou, N. C.; Petichakis, G. N.; Iconomidou, V. A., Linear B-cell epitope prediction: a performance review of currently available methods. *bioRxiv* **2019**, 833418.
10. Malonis, R. J.; Lai, J. R.; Vergnolle, O., Peptide-Based Vaccines: Current Progress and Future Challenges. *Chemical reviews* **2020,** *120* (6), 3210-3229.
11. Hos, B. J.; Tondini, E.; van Kasteren, S. I.; Ossendorp, F., Approaches to Improve Chemically Defined Synthetic Peptide Vaccines. *Frontiers in Immunology* **2018,** *9* (884).
12. Marqus, S.; Pirogova, E.; Piva, T. J., Evaluation of the use of therapeutic peptides for cancer treatment. *Journal of Biomedical Science* **2017,** *24* (1), 21.
13. Lau, J. L.; Dunn, M. K., Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry* **2018,** *26* (10), 2700-2707.
14. Chow, D.; Nunalee, M. L.; Lim, D. W.; Simnick, A. J.; Chilkoti, A., Peptide-based Biopolymers in Biomedicine and Biotechnology. *Mater Sci Eng R Rep* **2008,** *62* (4), 125-155.

15. Altunbas, A.; Pochan, D. J., Peptide-Based and Polypeptide-Based Hydrogels for Drug Delivery and Tissue Engineering. In *Peptide-Based Materials*, Deming, T., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 135-167.
16. Eskandari, S.; Guerin, T.; Toth, I.; Stephenson, R. J., Recent advances in self-assembled peptides: Implications for targeted drug delivery and vaccine engineering. *Advanced Drug Delivery Reviews* **2017,** *110-111*, 169-187.
17. Sanchez, A. B.; Nguyen, T.; Dema-Ala, R.; Kummel, A. C.; Kipps, T. J.; Messmer, B. T., A general process for the development of peptide-based immunoassays for monoclonal antibodies. *Cancer Chemother Pharmacol* **2010,** *66* (5), 919-925.
18. Guidotti, G.; Brambilla, L.; Rossi, D., Cell-Penetrating Peptides: From Basic Research to Clinics. *Trends in Pharmacological Sciences* **2017,** *38* (4), 406-424.
19. Vanhee, P.; van der Sloot, A. M.; Verschueren, E.; Serrano, L.; Rousseau, F.; Schymkowitz, J., Computational design of peptide ligands. *Trends in Biotechnology* **2011,** *29* (5), 231-239.
20. Wu, C.-H.; Liu, I. J.; Lu, R.-M.; Wu, H.-C., Advancement and applications of peptide phage display technology in biomedical science. *Journal of Biomedical Science* **2016,** *23* (1), 8.
21. Pirkhezranian, Z.; Tahmoorespur, M.; Monhemi, H.; Sekhavati, M. H., Computational Peptide Engineering Approach for Selection the Best Engendered Camel Lactoferrin-Derive Peptide with Potency to Interact with DNA. *International Journal of Peptide Research and Therapeutics* **2020**.
22. Zanuy, D.; Sayago, F. J.; Revilla-López, G.; Ballano, G.; Agemy, L.; Kotamraju, V. R.; Jiménez, A. I.; Cativiela, C.; Nussinov, R.; Sawvel, A. M.; Stucky, G.; Ruoslahti, E.; Alemán, C., Engineering strategy to improve peptide analogs: from structure-based computational design to tumor homing. *Journal of Computer-Aided Molecular Design* **2013,** *27* (1), 31-43.
23. Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L., Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem* **2015,** *8*, 37-47.
24. Kmiecik, S.; Kouza, M.; Badaczewska-Dawid, A. E.; Kloczkowski, A.; Kolinski, A., Modeling of Protein Structural Flexibility and Large-Scale Dynamics: Coarse-Grained Simulations and Elastic Network Models. *Int J Mol Sci* **2018,** *19* (11), 3496.
25. Abe, K.; Kobayashi, N.; Sode, K.; Ikebukuro, K., Peptide ligand screening of α-synuclein aggregation modulators by in silico panning. *BMC Bioinformatics* **2007,** *8* (1), 451.
26. Gee, M. H.; Sibener, L. V.; Birnbaum, M. E.; Jude, K. M.; Yang, X.; Fernandes, R. A.; Mendoza, J. L.; Glassman, C. R.; Garcia, K. C., Stress-testing the relationship between T cell receptor/peptide-MHC affinity and cross-reactivity using peptide velcro. *Proceedings of the National Academy of Sciences of the United States of America* **2018,** *115* (31), E7369-E7378.
27. Gorelik, M.; Davidson, A. R., Distinct peptide binding specificities of Src homology 3 (SH3) protein domains can be determined by modulation of local energetics across the binding interface. *The Journal of biological chemistry* **2012,** *287* (12), 9168-9177.
28. Soussi, T., p53 Antibodies in the Sera of Patients with Various Types of Cancer: A Review. *Cancer Research* **2000,** *60* (7), 1777.
29. Suppiah, A.; Greenman, J., Clinical utility of anti-p53 auto-antibody: systematic review and focus on colorectal cancer. *World J Gastroenterol* **2013,** *19* (29), 4651-4670.
30. King, C. A.; Bradley, P., Structure-based prediction of protein-peptide specificity in Rosetta. *Proteins* **2010,** *78* (16), 3437-49.
31. Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M., Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding. *Journal of the American Chemical Society* **2013,** *135* (41), 15579-15584.

32. Singh, H.; Singh, S.; Singh Raghava, G. P., Peptide Secondary Structure Prediction using Evolutionary Information. *bioRxiv* **2019**, 558791.
33. Muñoz, V.; Serrano, L., Elucidating the folding problem of helical peptides using empirical parameters. *Nature structural biology* **1994,** *1* (6), 399-409.
34. Muñoz, V.; Serrano, L., Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *Journal of molecular biology* **1995,** *245* (3), 275-296.
35. Muñoz, V.; Serrano, L., Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J Mol Biol* **1995,** *245* (3), 297-308.
36. Shen, Y.; Maupetit, J.; Derreumaux, P.; Tufféry, P., Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. *Journal of Chemical Theory and Computation* **2014,** *10* (10), 4745-4758.
37. Thévenet, P.; Shen, Y.; Maupetit, J.; Guyon, F.; Derreumaux, P.; Tufféry, P., PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res* **2012,** *40* (Web Server issue), W288-93.
38. Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P., PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research* **2016,** *44* (W1), W449-W454.
39. Yan, R.; Xu, D.; Yang, J.; Walker, S.; Zhang, Y., A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports* **2013,** *3* (1), 2619.

Chapter 4: Optimizing the Hidden Symmetry Model Using Solvent Interaction Terms and a Simple Protein Statistical Mechanics Toy Model

Zachary Fritz[1], Lawrence Williams[2], Martin Yarmush[1],

[1]Department of Biomedical Engineering, Rutgers, The State University of New Jersey

[2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey

Contribution

As the lead author of this chapter, I (Zachary Fritz) developed the toy model, performed all calculations, analyzed the data, made all the figures, and wrote the manuscript.

Introduction

A simple, fast, and user-friendly protein thermodynamic model that can accurately provide a per-residue decomposition of interaction free energy would fill a valuable niche in the protein engineering and computational drug design fields. We previously introduced the Hidden Symmetry Model (HSyM, see Chapter 2), which can perform this function in a matter of seconds using only sequence and structure information from a protein's native conformation, substantially reducing computational demand compared to molecular dynamics simulations and Monte Carlo sampling methods. HSyM was first applied to predicting the thermostability changes of buried point mutations in T4 lysozyme and was able to do so with high accuracy (R^2 = 0.71) when compared with experimental data [1].

While the model was successful with these interior and partially buried mutations, which we classified as Mutation Class I (MC-I), an interesting trend was observed for those solvent exposed mutation sites, which we termed Mutation Class II (MC-II). HSyM's thermostability predictions for the MC-II mutants gave a negative correlation (R^2 =-0.22) with experimental data. We attributed this to the fact that the model does not currently take into account solvation effects for different types of residues, such as if the residue is polar, charged, or nonpolar. This issue manifests itself in the problem that the model will predict a *stabilizing* effect for any mutation to a higher gamma value residue (typically more hydrophobic), regardless of context. This assumption of course does not hold true for all mutations. For example, many mutations to arginine of solvent exposed hydrophobic residues in

acetylcholinesterase were found to be stabilizing [2], and conversely mutating the solvent exposed R96 site in T4 lysozyme to more hydrophobic/nonpolar residues tends to be highly destabilizing [3].

To address this problem, we proposed applying a solvent interaction efficiency parameter (τ) to help differentiate the favorability of residues' interactions with water molecules. This parameter ranges from 0 to 1 and weighs the efficiency of a residue's interaction with water based on the residue's identity (polar or nonpolar) and the magnitude of its model-calculated μ value, which is directly proportional to its interaction free energy. This parameter was included in the model when it was first introduced, but was set to a default value of 1.0 for all interactions (residue-residue and residue-water). To aid us in determining the set of residue values for τ, we developed a simple 2D protein toy model that was inspired by similar models by Dill [4, 5]. This toy model allowed us to simulate solvent effects, protein folding and unfolding, and other complex behavior within a system of a limited, known number of possible protein conformations. We conducted a statistical mechanical analysis of this toy model to predict the effects of certain types of mutations on folded state stability and how changing the solvent interaction efficiency parameter values influences these stability changes. Going forward, we believe this toy model will serve as a "proving ground" for optimizing other model parameters in a simplified manner, rather than running hundreds of trial-and-error calculations on actual protein structures.

Materials and Methods

The Toy Model

Our toy model represents a merging of elements from Ken Dill's 2D heteropolymer models [4, 5] and HSyM itself. Protein conformations and water molecules are represented on a 2D square

lattice, with an amino acid residue occupying one square and unoccupied squares representing solvent. The polypeptide chain is represented by a numbered six residue hexomino configuration, with the squares constituting residues *i* and *i+1* always sharing a side to represent a peptide bond. Residues are allowed to move into adjacent unoccupied squares and make 90° and 180° turns as long as they are still connected to their neighboring residues. All possible chain conformations are shown in **Figure 4.1 (A),** with each conformer's associated symmetry and residue numbering allowing for multiple nonredundant stereoisomers and enantiomers (**Figure 4.1 (B)**), bringing the total number of conformations up to 71.

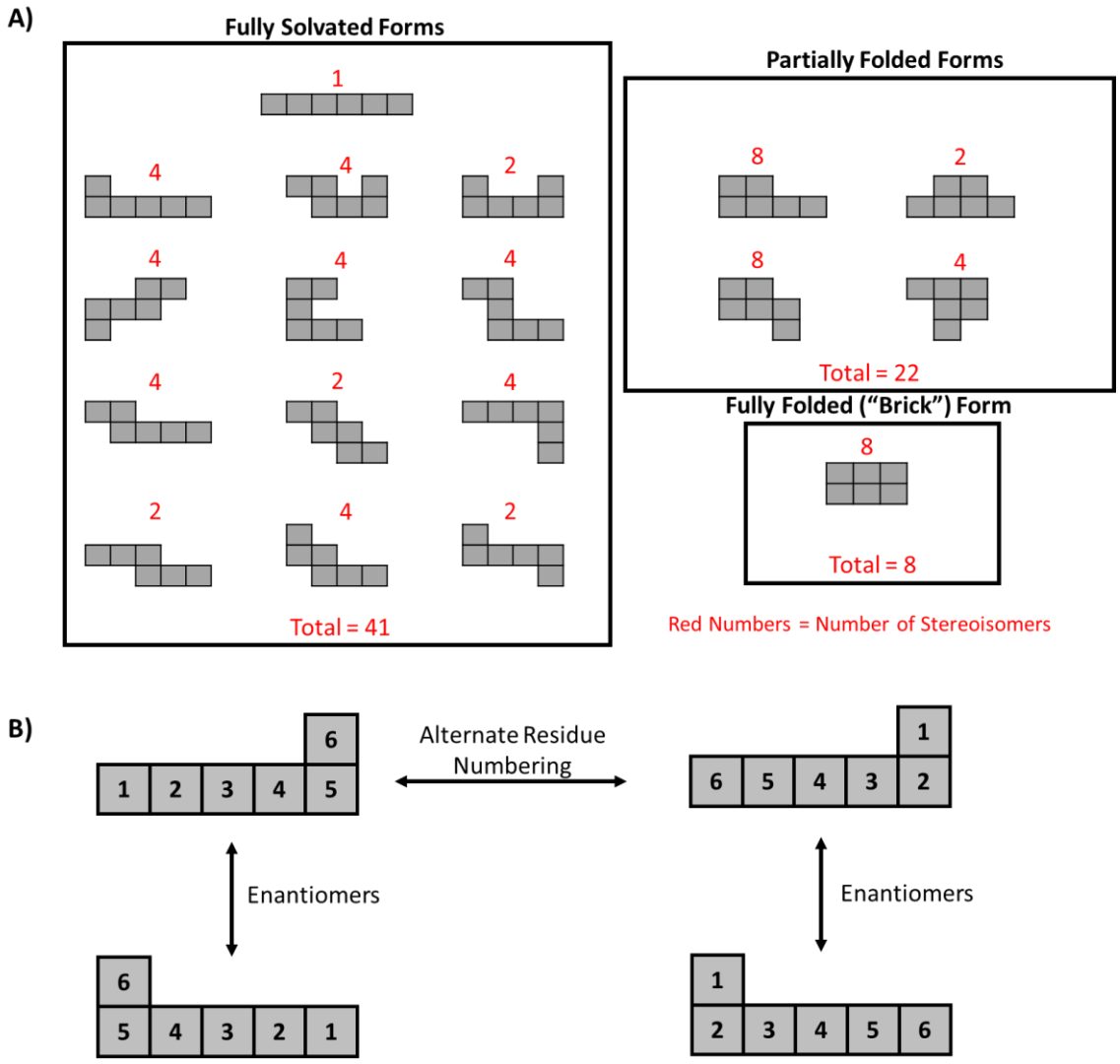


Figure 4.1: Toy Model Hexomino Conformations. (A) All possible hexomino conformations of the six-residue polypeptide chain used in the toy model. Hexominos with more than two terminal ends are not permitted. The hexominos are categorized by the number of residue-residue contacts they have: Solvated conformations have none, Partially Folded have one, and Fully Folded have two. The red numbers above each hexomino indicate the number stereoisomers it has, for a total of 71 distinct conformations. (B) An example of how a single hexomino shape can constitute several nonredundant conformations.

Residues are assigned either solvent or residue contacts based on the identity of the non-peptide bonded squares they share a side with. In this way the terminal residues (1 and 6) are allowed three contacts and the interior residues (2-5) are permitted two. Unoccupied solvent squares are permitted to have multiple residue contacts. We classified the chain conformations based on the number of residue-residue contacts they have, with Fully Folded conformations having two, Partially Folded having one, and Solvated conformations having none. To represent solvent displacement upon folding and to conserve the total number of contacts (14) in the system, for every residue-residue contact formed an implicit solvent-solvent contact is also added.

Every residue is assigned a binary identity, either Polar (P) or Nonpolar (N). Like the HSyM, each residue and solvent square is also given a unitless interaction factor (μ) value, representing its interaction free energy contribution. Unlike HSyM these values were assigned manually and not calculated based on the residue's identity, the identity of its neighbors, or the chain's secondary structure. Residue μ values were also assigned in a binary manner, either being "Hot" (μ = 170) or "Cold" (μ = 120); the cutoff value between hot and cold is an additional parameter to be optimized. The unoccupied solvent squares were set to a default μ value of 110, the same value used in Chapter 2.

To determine the total free energy of each conformation, we calculated the sum of all the contacting interaction factor products in a manner similar to the energy calculations done in the T4 Lysozyme thermostability study (Chapter 2):

$$E_{Conf} = -\lambda \sum_{i,j} \tau_{i,j}\, \mu_i \mu_j$$

where E_{Conf} is the total energy of a conformation, λ is a scaling factor with units of kcal/mol, $\tau_{i,j}$ is the interaction efficiency parameter for the contact, and μ_i and μ_j are interaction factor values for a contacting residue-residue, residue-solvent, or solvent-solvent pair. For the purposes of this study the τ values for all residue-residue and solvent-solvent interactions were set to a constant value of 1, as we are primarily interested in optimizing the interaction efficiencies for residue-solvent contacts.

Statistical Mechanics Calculations

Since the toy model has a defined number of protein conformations (71) with known energy values, we were able to use a fundamental equation of statistical mechanics [6] to determine the probability of the system existing in any given conformation:

$$p(E_{Conf}) = \frac{\exp(-\beta E_{Conf})}{\sum_N \exp(-\beta E_N)} = \frac{\exp(-\beta E_{Conf})}{Z}$$

Where $p(E_{Conf})$ is the probability of the system existing in the given conformation, β is the inverse product of the temperature T (in Kelvin) and Boltzmann's constant (k), N is the total number of conformations/microstates, and the denominator Z is known as the partition function. From the individual conformation probabilities, the group probabilities for the Fully Folded, Partially Folded, and Solvated conformation classifications can be determined for a given temperature, and thus the Percent Unfolded character of the system at a given temperature can be calculated:

$$\%\, Unfolded = 100 \times (p_{Solvated\ Group} + 0.5 \times p_{Partially\ Folded\ Group})$$

Tau Parameter Optimization Workflow

Figure 4.2 summarizes the steps taken to optimize the solvent interaction efficiency values. First, mutations at two different types of sites—Solvent Exposed and Partially Buried-- were

classified into four groups—N Hot to P Cold, N Hot to P Hot, N Hot to N Cold, and P Hot to P Cold—and ranked according to their predicted level of stabilization/destabilization of the Fully Folded state. Next, single conformation analyses were performed using a single toy model Fully Folded or Solvated conformation. In these analyses each of the four types of point mutations was performed at either of the two types of mutation sites with dozens of τ value combinations. Energy calculations were done using a VBA-based Excel macro. All non-mutated residues were assigned an N Hot identity. For water-residue interactions, the τ parameter was allowed to have four different values, depending on if the residue was N Hot, N Cold, P Cold, or P Hot. τ combinations that gave the expected mutant stability ranking were selected for further testing. These combinations were used in a statistical mechanics analysis in which mutations (at specific numbered residues) were applied to all 71 possible conformations of the toy model system at a range of temperatures to see how much the mutation stabilized or destabilized the folded ensemble. Finally, the best performing τ combination was applied to HSyM to predict point mutation thermostability changes in T4 lysozyme.

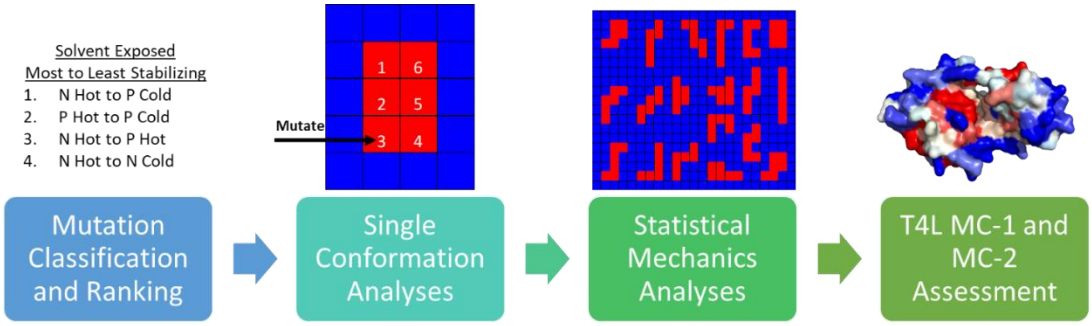


Figure 4.2: Toy Model Solvent Interaction Efficiency Parameter Optimization Process

T4 Lysozyme MC-1 and MC-2 Mutant Data

For parameter optimization assessment, HSyM was used to predict point mutation thermostability shifts in T4 Lysozyme (PDB: 3FA0) with the updated solvent interaction efficiency terms in the same manner as was presented in Chapter 2, but this time the MC-2

(solvent exposed) mutations were included in addition to the original MC-1 (buried) mutations. The MC-2 class consists of 43 mutations found at sites 96, 105, 131, and 157, while the MC-1 class consists of 28 mutations found at sites 3, 11, 115, 117, 119, 132, and 133. For residue-solvent interaction efficiency assignments, the Polar (P) residues were considered to be D, E, K, N, Q, R, and S; all other residues were considered Nonpolar (N). Experimental thermostability data was obtained from a compiled review [3] and was fitted against the model's predictions.

Results

Single Conformation Mutation Studies

For the single conformation analysis, we used the "U"-shaped Fully Folded conformer shown in **Figure 4.3 (A)**. Starting from the Solvent Exposed (no residue-residue contacts) residue 3 position, we ranked the four types of mutations with respect to their predicted magnitudes of stabilization; two slightly different rankings were chosen due to uncertainty over whether the N Hot to N Cold or N Hot to P Hot mutations would be more stabilizing at this site. The wildtype (WT) state was established as assigning all residues an N Hot (μ =170) designation, with an exception for the P Hot to P Cold mutation, in which the WT state was all P Hot (μ =170) residues. **Figure 4.3 (B) and (C)** shows the two τ combinations that yielded the desired stability rankings, along with the degree of stabilization for each mutation type (more negative values are more stabilizing). τ values greater than 0.9 or less than 0.2 were avoided, as they were judged to be too extreme to be realistic.

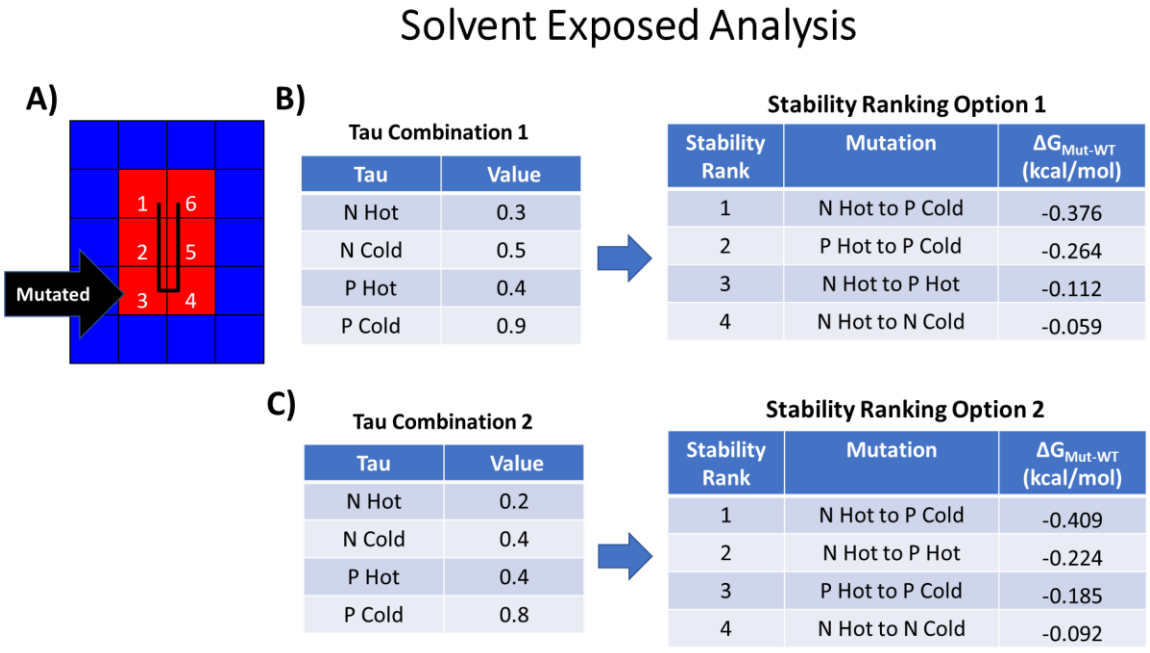


Figure 4.3: Single Conformation Solvent Exposed Site Analysis **(A)** The Fully Folded conformation used in this analysis, with the mutated site 3 indicated. The U-shaped black lines represent the peptide backbone. **(B) & (C)** The residue-solvent τ values for each type of residue (in the tables on the left) that gave the desired mutation stability rankings (in the tables on the right).

A similar analysis was done for the residue 2 position, which was classified as a Partially Buried (1 residue-residue contact, 1 residue-solvent contact). **Figure 4.4** shows the τ values from the Solvent Exposed analysis applied to the mutations at this position. In contrast to the Solvent Exposed site, most of the mutations were destabilizing, though this was expected as mutating to colder and more polar residues should weaken the residue-residue contact (with N Hot residue 5) that this site possesses. Thus, the τ values selected from the Solvent Exposed site are still feasible.

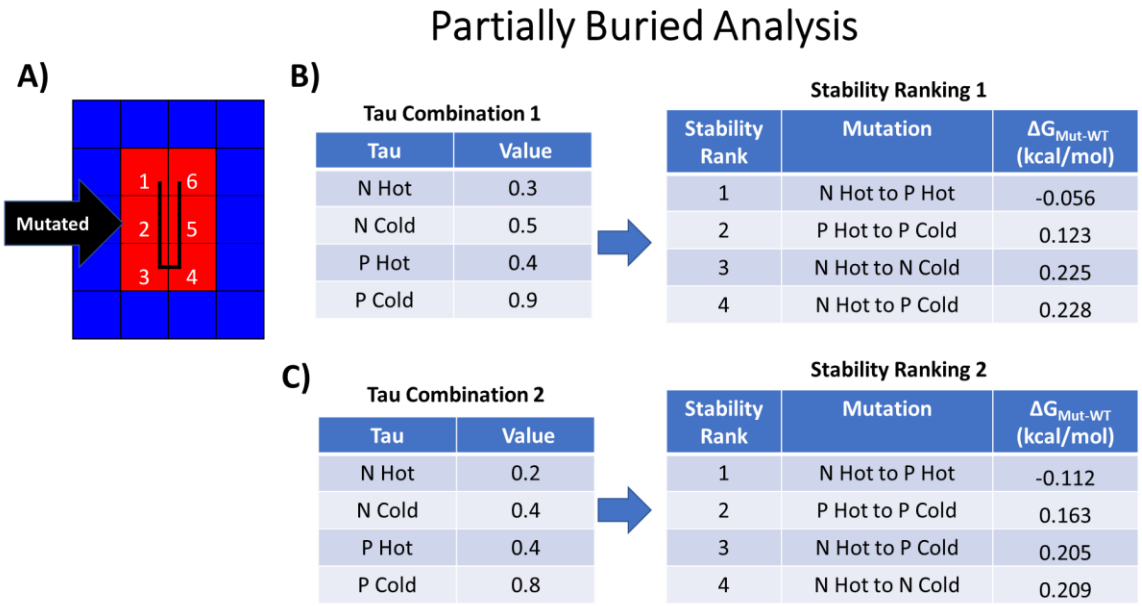


Figure 4.4: Single Conformation Partially Buried Site Analysis **(A)** The Fully Folded conformation used in this analysis, with the mutated Site 2 indicated. **(B) & (C)** The same τ combinations used in **Figure 4.3 (B) & (C)** were applied to the Partially Buried site mutations, with associated stability rankings shown in the tables to the right.

A comparison analysis was also done on a Solvated conformation (**Figure 4.5 (A)**) using the selected τ value combinations. Unsurprisingly, the decreases in stability for the mutations at residue 3 in the Solvated conformation exactly matched those of the Fully Folded conformation, as in both conformations the only contacts residue 3 has are two residue-solvent contacts. Conversely, the mutations at residue 2 tended to stabilize the Solvated conformation (**Figure 4.5 (B) & (C)**). This also made intuitive sense: more polar polypeptide chains are more likely to be soluble in aqueous solutions, and trends from our work with HSyM suggest that colder (lower μ value) residues seem to favor being solvent exposed. However, it is interesting to note that the stability of the Solvated state mutations does not necessarily inversely correlate with the destabilizing effect of the mutations at this site in the Fully Folded conformation.

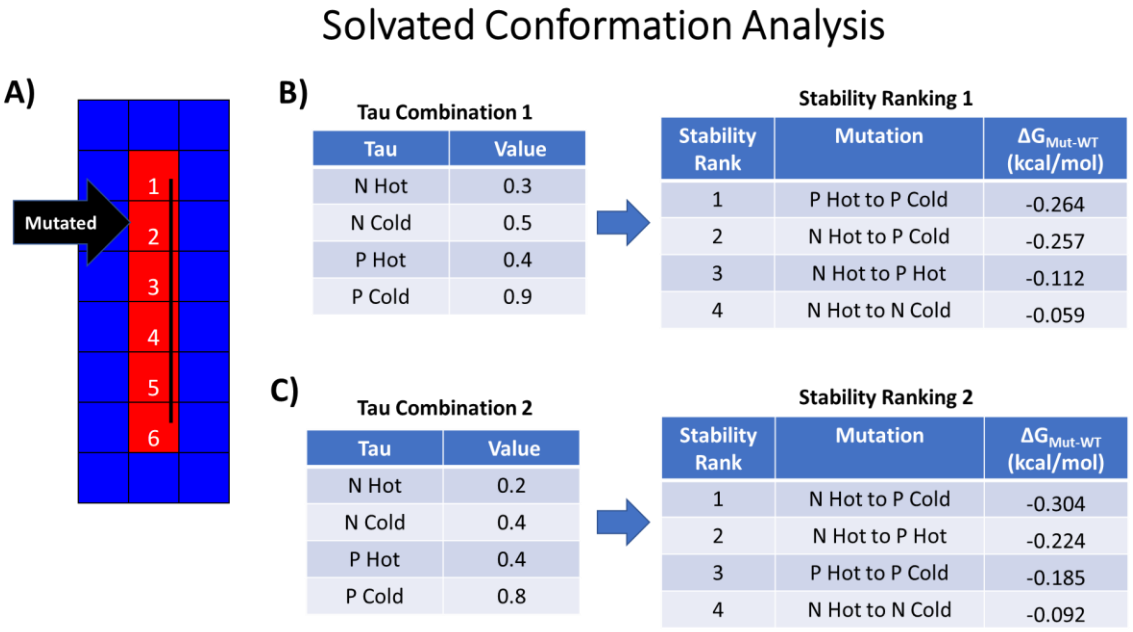


Figure 4.5: Solvated Single Conformation Analysis **(A)** The Solvated conformation and mutation site used in this analysis. Mutations at site 3 gave the same stability values as those at site 3 in the Fully Folded conformation (**Figure 4.3**), and so are not shown here. **(B) & (C)** The τ values from the Fully Folded analyses were applied to mutating site 2 in the Solvated conformation, with the resulting stability changes shown in the tables to the right.

Statistical Mechanics Analysis

The two sets of solvent interaction efficiency (τ) values chosen in the single conformation analysis were applied to all 71 possible toy model conformations in a statistical mechanics analysis. The same four types of mutations used in the single conformation analysis were applied to residues 2 and 3 in all conformations, with the wildtype (WT) state again being all residues set to a N Hot designation (except for the P Hot to P Cold mutation). **Figures 4.6** and **4.7** show the melting curves for the toy model system for both mutant sites and τ combinations. Increasing the N Hot or P Hot (the default residue types for the WT) τ values by an increment of

0.1 shifted both WT and mutant melting curves to the left by about 50K, indicating an increased preference for the Extended conformations and a destabilization of the Folded states. It should also be noted that changing the scaling factor, λ (set to 3E-5 kcal/mol for this study), will also shift the melting curves, as shown in **Supplementary Figure S4.1.** Decreasing λ will increase the slope of the transition region of the curve, as it will take a smaller amount of thermal energy (temperature increase) to overcome the smaller energy differences between the Solvated and Folded conformations, causing the protein to denature at lower temperatures.

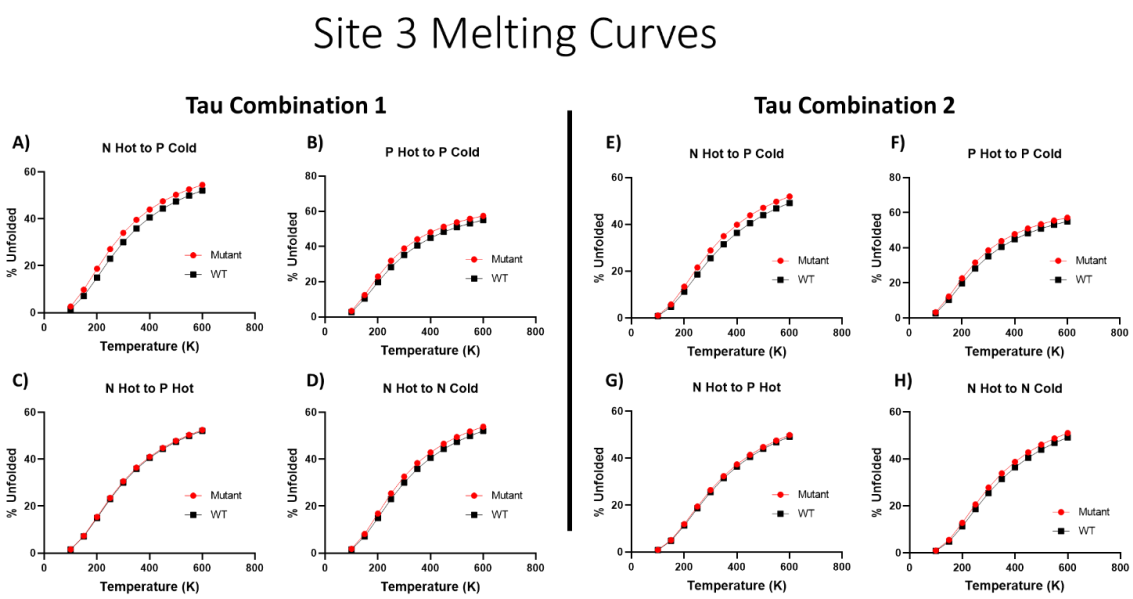


Figure 4.6: Melting Curves for Wildtype (WT) and Site 3 Mutants. **(A)-(D)** mutations used the τ values from Combination 1 (**Figure 4.3 (B)**), with average % Unfolded differences between mutant and WT of 3.54%, 3.32%, 0.47%, and 2.27% respectively for the region between 200 and 500K. **(E)-(H)** mutations used the τ values from Combination 2 (**Figure 4.3 (C)**), with average % Unfolded differences between mutant and WT of 3.10%, 3.02%, 0.83%, and 2.13% respectively for the region between 200 and 500K.

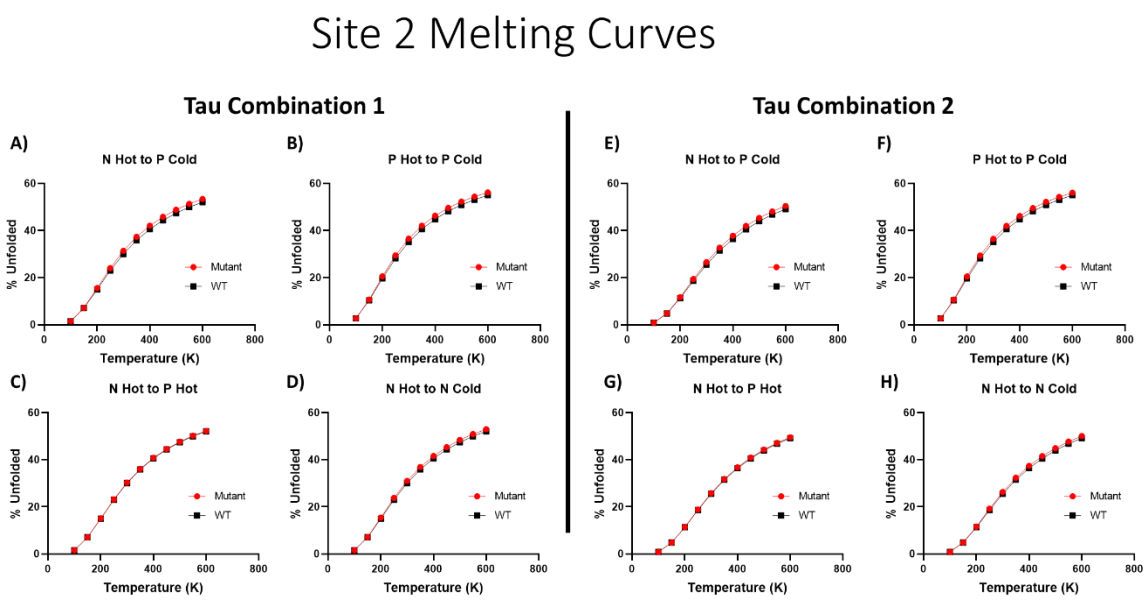


Figure 4.7: Melting Curves for Wildtype (WT) and Site 2 Mutants. **(A)-(D)** mutations used the τ values from Combination 1 (**Figure 4.3 (B)**), with average % Unfolded differences between mutant and WT of 1.27%, 1.33%, 0.19%, and 0.88% respectively for the region between 200 and 500K. **(E)-(H)** mutations used the τ values from Combination 2 (**Figure 4.3 (C)**), with average % Unfolded differences between mutant and WT of 1.10%, 1.22%, 0.31%, and 0.77% respectively for the region between 200 and 500K.

Most of the mutations at Site 3 tended to slightly destabilize the system compared to WT, ranging from a 2-4% increase in the unfolded percentage for the portion of the melting curve from 200 to 500K. The exception to this was the N Hot to P Hot mutation, which was a very neutral mutation with the unfolded percentage difference between mutant and WT not exceeding 1% for the same temperature range. By contrast, all of the mutations at Site 2 were very neutral, with the unfolded differences between mutant and WT not exceeding 1.5%.

The differences in destabilization between the mutants at sites 2 and 3 comes from the how often each site is fully solvent exposed (2 residue-solvent contacts and no solvent-solvent

contacts) in the ensemble of conformations. Site 3 is only fully solvent exposed in half of the Fully Folded conformations and 64% of the Partially Folded ones, whereas Site 2 is fully solvent exposed in 75% of the Fully Folded conformations and 73% of the Partially Folded ones. Since our τ values favor contacts between polar and/or cold residues with solvent, the mutations at Site 2 were better able to compensate for any destabilizing effects due to disruption of residue-residue contacts. The effect of solvent accessibility on mutation destabilization was also demonstrated with two double mutations of sites 2 and 3 from N Hot to P Cold for both sets of τ values (**Supplementary Figure S4.2).** Since both sites are only completely solvent exposed in a quarter of the Fully Folded conformations, the mutation to a colder polar residue will prove unfavorable for the residue-residue contacts these sites have and greater destabilization will occur. A similar effect is seen when mutating the terminal residue at site 1 (**Supplementary Figure S4.3)**, which is never completely solvent exposed in the Fully Folded conformations.

T4 Lysozyme MC-1 and MC-2 Assessment

The two selected τ combinations were implemented into the existing HSyM code and algorithm and used to predict thermostability changes in T4 lysozyme for both the MC-1 (mostly buried) and MC-2 (mostly solvent exposed) mutation classes. The μ cutoff value for what residues were considered "hot" was set to 135 and above, and only D, E, K, N, Q, R, and S residues were considered "Polar". **Figure 4.8** shows the MC-1 and MC-2 HSyM calculated thermostability changes fitted against experimental data for both τ value combinations as well as the original default setting of all τ values set to 1.0. It is apparent that neither selected τ combination improved the accuracy of the model with regards to the MC-2 mutations. Increasing or decreasing the hot μ cutoff value by increments of 5 between a range of 120-150 did not significantly improve the correlations (data not shown), though removing HSyM's entropy and

temperature terms from the calculations did cause the MC-2 correlations to have a more negative correlation (from R^2 = -0.09 to -0.46 for τ combination 2).

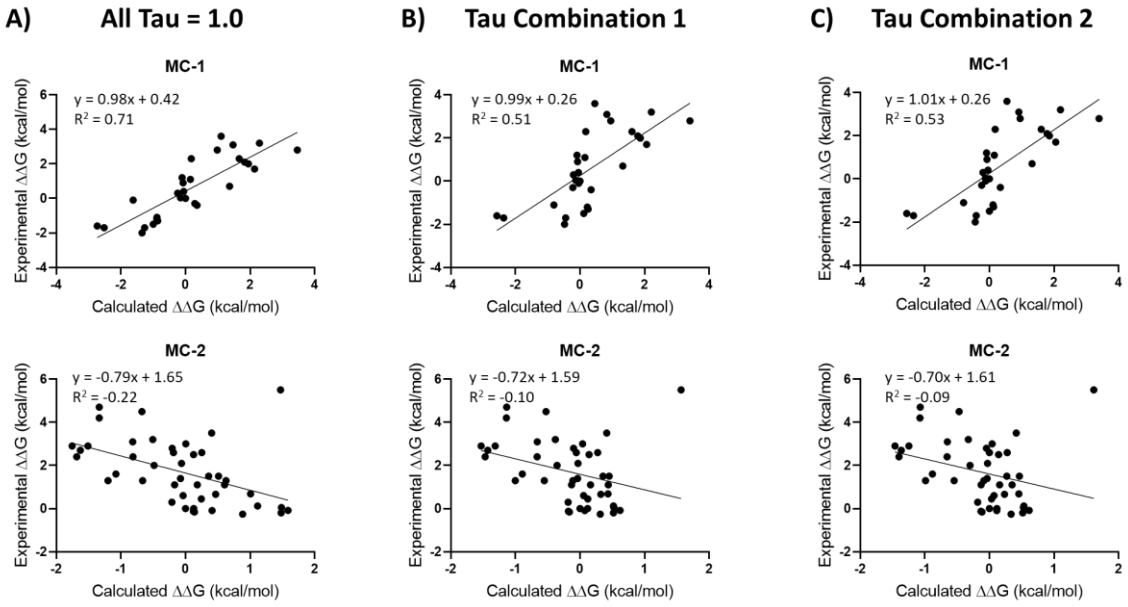


Figure 4.8: Comparison of HSyM Calculated and Experimental Thermostability Data for MC-1 and MC-2 Mutations of T4 Lysozyme. **(A)** The HSyM calculations were done in the same way as Chapter 2, with τ = 1.0 for all residue-solvent and residue-residue interactions. (The line of best fit and R^2 values differ slightly from the original published data due to modifications made to HSyM's contact and entropy assignment coding). **(B)-(C)** HSyM calculations were done using the 4-fold residue-solvent τ values from τ Combinations 1 and 2. All residue-residue τ values remained 1.0.

Conclusions and Discussion

While the solvent interaction efficiency parameters chosen in this analysis did not yield improved predictive fits for the solvent exposed MC-2 mutations, the toy model developed herein did provide several valuable insights. For example, while residue 96 in T4 lysozyme is classified as an MC-2 site and is indeed on the surface of the protein, it also makes contacts with

several other (mostly warm or hot) residues, similar to the "partially buried" sites in the toy model that have both solvent and residue contacts. Determining the proper balance in solvent interaction efficiency terms for these sorts of "hybrid" residues will be a central challenge in improving HSyM. Several possible alternatives to a 4-value residue-solvent τ parameter based on residue identity (N or P) and μ value (Hot or Cold) should be tried, including a binary designation based on residue identity (N or P), a linear function/continuum of τ values rather than 4 discrete values, and a method that takes into account the number of residue contacts an amino acid has before assigning a τ value. Additionally, more mutations from other thermodynamically well-characterized proteins (barnase, staph nuclease) should be assessed, with a special emphasis on categorizing mutants based on if they are buried, solvent exposed, or a blend of the two.

Of course, the problem of the MC-2 mutations might lie primarily in another facet of the model. It is highly likely, for example, that the interaction efficiency parameters for certain types of residue-residue contacts should not be equal to the default value of 1.0 (100% efficient); it makes more sense for hydrophobic residues to suffer a penalty when interacting with polar and charged side chains. Other potential sources of error in HSyM include limiting residues to 6 contacts, not taking into account side chain length changes (and potential gain or loss of contacts) when mutating residues, or a more fundamental flaw in one of the model's other parameters (γ, χ, or σ).

In contrast with the single conformation analyses, in the statistical mechanics analysis we could not isolate a single residue position that was always either fully solvent exposed or partially buried. While this made stabilizing the Folded conformations by mutating a residue from N to P or Hot to Cold difficult, it is more representative of reality than a single conformation analysis. Proteins have many stable and semi-stable conformations, each with its own ensemble of

microstates with an associated free energy; indeed, the "native state" of a protein actually consists of an ensemble of fluctuating microconformations [7]. For this reason, it is certain that a given residue will not always be solvent exposed, buried, or partially buried for all protein conformations. The fact that such a simple toy model can simulate protein behavior so well makes it an attractive tool for us to test and optimize HSyM. In the near future we hope to use it to optimize not only the interaction efficiency parameter for both solvent-residue and residue-residue contacts, but also the water/solvent μ value and its dependence on temperature in order to simulate complex behavior like cold denaturation [8].

Supplementary Figures

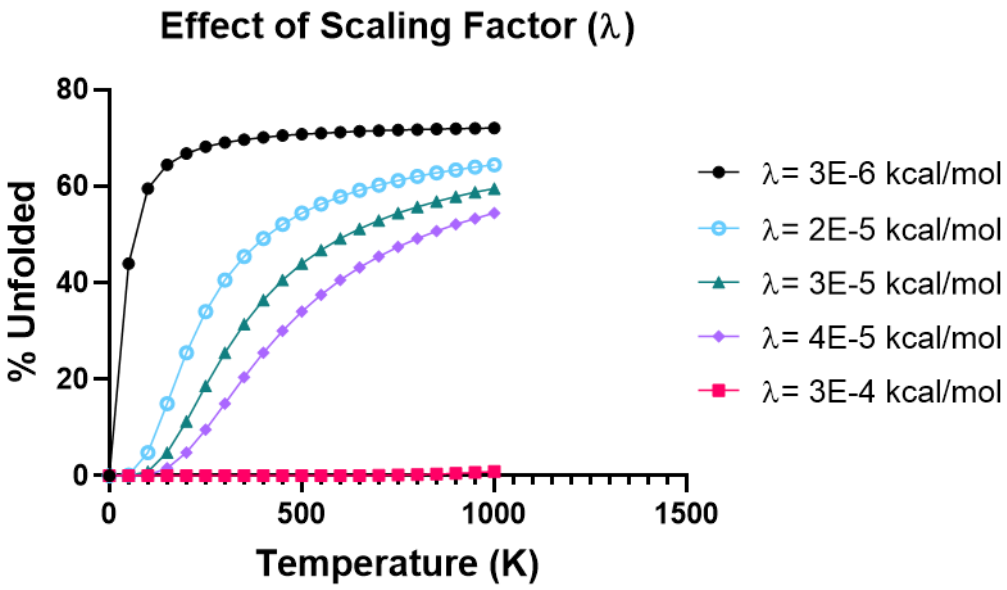


Figure S4.1: The Effect of the Scaling Factor (λ) on Toy Model Melting Curves. The WT (all N Hot) state with Tau Combination 2 was used to generate these curves. The default λ value used in this study was 3E-5 kcal/mol.

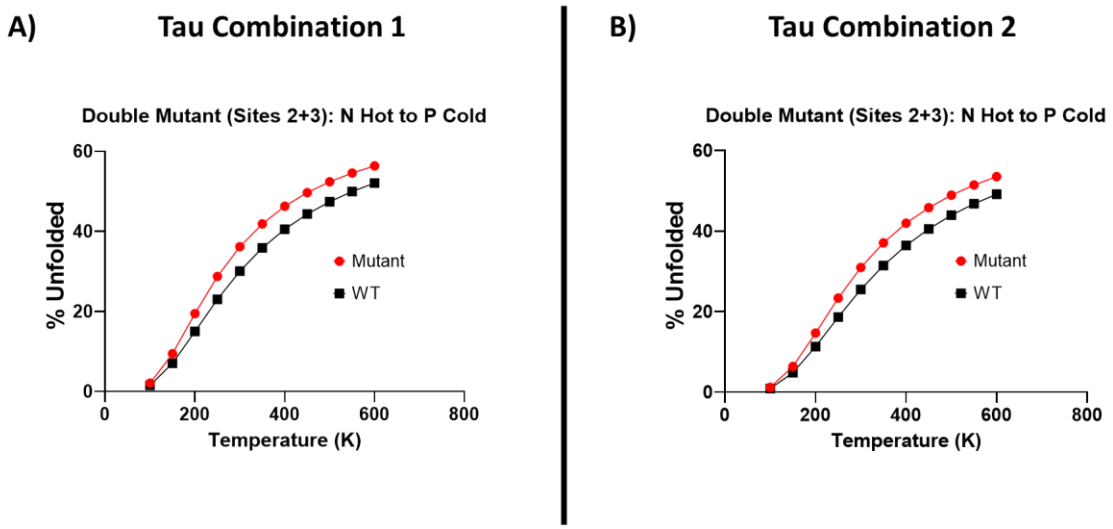


Figure S4.2: Melting curves for a Double N Hot to P Cold Mutation at Sites 2 and 3. **(A)** The mutation using τ Combination 1, with an average % Unfolded difference between mutant and WT of 5.48% for the region between 200 and 500 K. **(B)** The mutation using τ Combination 2, with an average % Unfolded difference between mutant and WT of 4.99% for the region between 200 and 500 K.

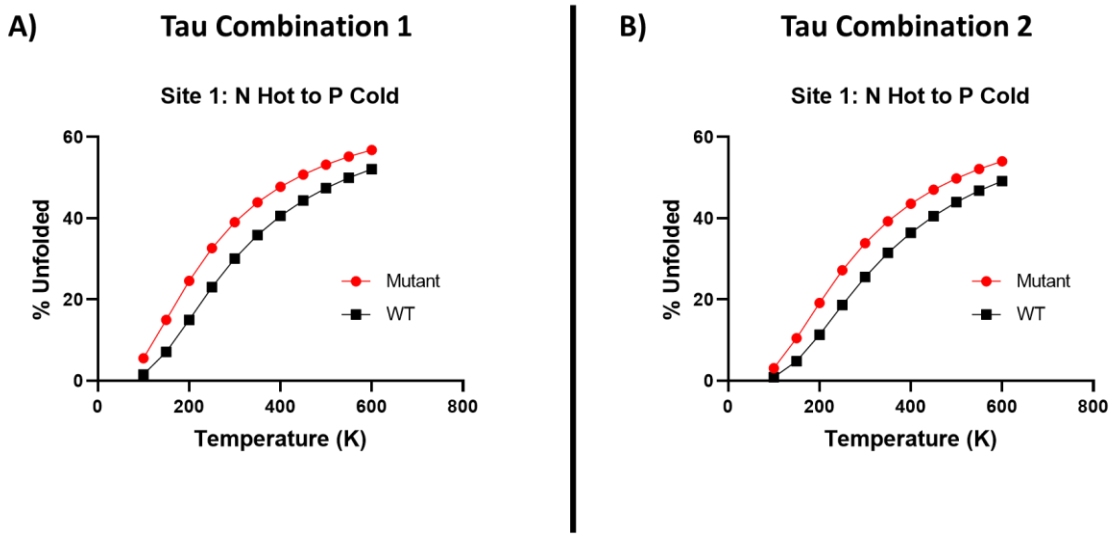


Figure S4.3: Melting Curves for the N Hot to P Cold Mutation at Site 1. **(A)** The mutation using τ Combination 1, with an average % Unfolded difference between mutant and WT of 7.94% for the region between 200 and 500 K. **(B)** The mutation using τ Combination 2, with an average % Unfolded difference between mutant and WT of 7.44% for the region between 200 and 500 K.

References

1. Williams, L. J.; Schendt, B. J.; Fritz, Z. R.; Attali, Y.; Lavroff, R. H.; Yarmush, M. L., A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme. *Technology (Singap World Sci)* **2019,** *7* (1-2), 12-39.
2. Strub, C.; Alies, C.; Lougarre, A.; Ladurantie, C.; Czaplicki, J.; Fournier, D., Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. *BMC Biochemistry* **2004,** *5* (1), 9.
3. Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W., Lessons from the lysozyme of phage T4. *Protein Science* **2010,** *19* (4), 631-641.
4. Miller, R.; Danko, C. A.; Fasolka, M. J.; Balazs, A. C.; Sun Chan, H.; Dill, K. A., Folding kinetics of proteins and copolymers. *The Journal of Chemical Physics* **1992,** *96* (1), 768-780.
5. Chan, H. S.; Dill, K. A., Transition states and folding dynamics of proteins and heteropolymers. *The Journal of Chemical Physics* **1994,** *100* (12), 9238-9257.
6. Wereszczynski, J.; McCammon, J. A., Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly reviews of biophysics* **2012,** *45* (1), 1-25.
7. James, L. C.; Tawfik, D. S., Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences* **2003,** *28* (7), 361-368.
8. Sanfelice, D.; Temussi, P. A., Cold denaturation as a tool to measure protein stability. *Biophys Chem* **2016,** *208*, 4-8.

Chapter 5: Preliminary Work on Developing a Fully Integrated Microfluidic Device for Point-of-Care Blood Diagnostics

Zachary Fritz[1], Anil Shrirao[1], Martin Yarmush[1]

Contribution

As the lead author of this chapter, I planned and conducted the majority of the experiments, analyzed the data, made the figures, and wrote the manuscript.

Introduction

Microfluidic devices have ushered in the promise of point-of-care diagnostics that can be utilized immediately in the clinic or in low-resource locations. This technology, often referred to as biomedical microelectromechanical systems (BioMEMS) or "lab-on-a-chip", aims to not only reduce the physical laboratory and cost footprints for common procedures, but also may allow these procedures to be performed faster and easier due to the smaller sample and reagent volumes required [1, 2]. Prototype BioMEMS devices have been developed by research teams to perform a variety of tasks, including blood counts [3], flow cytometry [4], and various biochemical assays [5, 6]. A subset of BioMEMS, known as centrifugal microfluidics or "lab-on-a-disk", further reduces the footprint of these devices by eliminating the need for fluid pumps, instead using centrifugal force to actuate fluid flow through the microfluidic chip [7].

Our objective is to develop a fully integrated, centrifugal microfluidic device that will contain all the components necessary to perform a diagnostic blood or sera immunoassay. An example of one such immunoassay would use Luminex microbeads functionalized with peptide antigens, designed using HSyM as described in Chapter 3, to detect cancer associated autoantibodies in patient blood. The use of microbeads allows us greater biomarker capture surface area, the possibility of multiplexed assays, and the ability to try unique sample mixing strategies [8, 9]. This work will be building off of the success of a device previously developed by our lab, a centrifugal microfluidic platform able to perform white blood cell counts [10, 11]. Our aim is to adapt this device by optimizing the microfluidic chip, sample mixing method, and optical

detection system towards a microbead-based immunoassay function. Other elements of the device which contributed to its success, such as its centrifugal fluid actuation and automated sample handling, will be retained.

This chapter presents two microfluidic chip designs based on two different sample mixing strategies: magnetic microbead mixing and mixing using surface acoustic waves (SAWs). Sample and reagent mixing is a critical concern in microfluidic immunoassays, as the low Reynolds number creeping flow profile associated with these devices limits mixing to slow diffusion unless special strategies are employed [12]. The disadvantages and advantages of each method are outlined, as well as the results of preliminary mixing and fluid actuation experiments. A section on how the optical detection system will have to be redesigned to accommodate a microbead-based immunoassay is also presented. Finally, the Conclusions and Discussion section summarizes the work done to date and discusses the ongoing and future objectives that must be accomplished.

Magnetic Microbead Mixing

Various groups have utilized the magnetic character of microbeads to overcome the obstacle of low Reynolds number flow in microfluidic channels and achieve efficient sample mixing [13, 14]. The largest inspiration we drew from was the batch-mode centrifugal magnetic mixing developed by Grumann et al., in which the researchers placed permanent magnets in alternating positions above the orbit of the spinning microfluidic disk [15]. The most efficient mixing was achieved when these magnets were combined with periodically switching the direction of the microfluidic disk's spin, with optimum mixing occurring with a maximum spin speed of 8 Hz and a spin switching acceleration of 32 Hz s^{-1}. This method of mixing was deemed most applicable to our device, as the immunoassay's sample and reagent mixing steps will involve batch-mode

mixing within a central mixing chamber. A major advantage of this method of mixing is its ease of implementation, as no on-chip electric or mechanical features are required.

Magnetic Mixing Chip Design and Manufacturing

Prototype microfluidic chips for this method of sample mixing were made out of alternating layers of 1/16” polyacrylic and pressure sensitive adhesive using an Epilog Zing laser cutter. SolidWorks software was used to design the microchannels, microchambers, and other features. To maintain a microfluidic flow profile, the inlet and mixing chambers and the microchannels connecting them were patterned in the thin first adhesive layer, while a waste channel and chamber was added to the second adhesive and acrylic layers beneath this. A micropore membrane filter (Sterlitech)—either a transparent polyester (PETE) filter with 0.4 µm pore size or an opaque polycarbonate (PCTE) filter with 1.0 µm pore size—was placed underneath the central mixing chamber between the polyacrylic and adhesive layers to allow waste liquid to flow downwards while trapping the 6 µm Luminex microbeads within the chamber. An example device design is shown in **Figure 5.1**.

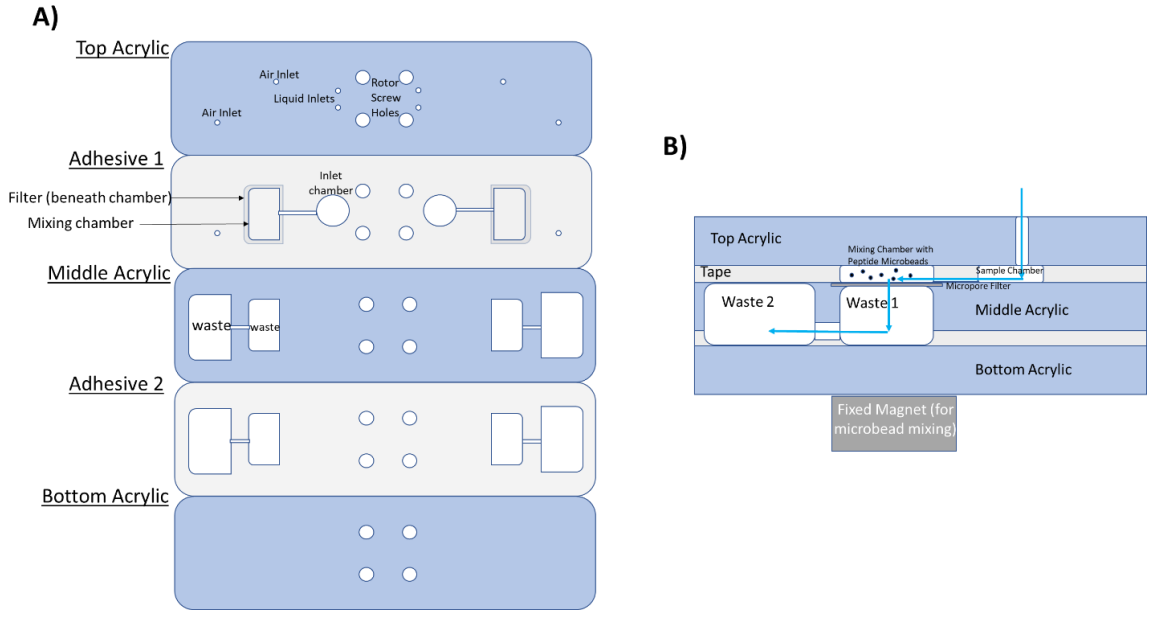
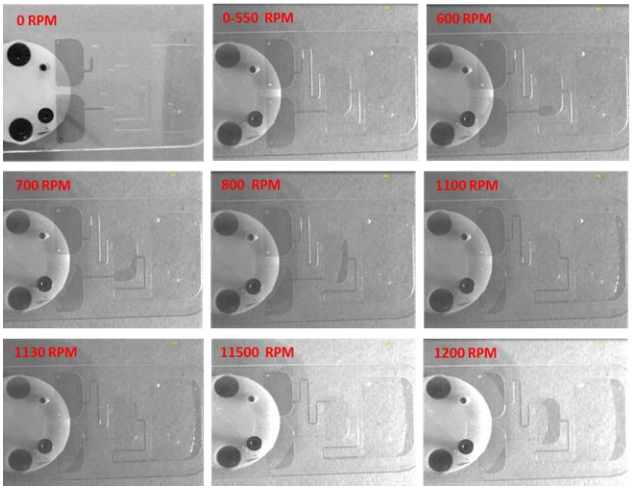


Figure 5.1 (previous page)**:** Prototype Microfluidic Chip Design for Magnetic Mixing Method. **(A)** Typical features for each layer of the chip, with the final design having additional microchannel geometries (straight, serpentine, and siphon) for sequential flow control. **(B)** Direction of liquid flow through the chip.

Flow Actuation Testing

Immunoassays require the precise addition and removal of reagents and buffers in sequential steps. While many microfluidic devices make use of pneumatic, magnetic, and other types of valves that require active actuation [16], centrifugal systems offer the advantage of using passive valve designs that depend on microchannel geometry, capillary forces, and changes in the spin speed and centrifugal force to control fluid flow [7, 17]. Examples of these types of valves include capillary burst valves and siphon valves.



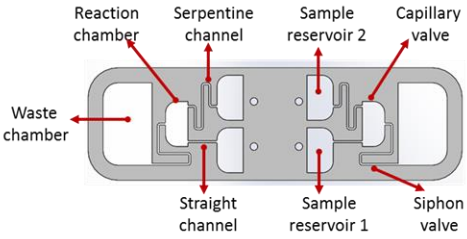


Figure 5.2: Sequential Flow Control by Varying Microchannel Geometries. Each unit of this prototype chip consists of two reservoir chambers connected to a reaction chamber by two different types of capillary valves (straight channel or serpentine channel), with the reaction chamber venting to a waste chamber via a siphon valve. In the photo series demonstration, accelerating the rotational spin to 550 RPM allows fluid from sample reservoir 1 (red arrows) to

prime the straight channel, then overcome the capillary forces to enter the reaction chamber at 600 RPM. Further acceleration (700-800 RPM) allows the fluid to begin exiting the reaction chamber and priming the siphon valve. Acceleration to 1100 RPM causes fluid to fully exit the siphon valve, while fluid from sample reservoir 2 (yellow arrows) primes the serpentine capillary valve. Finally, at 1200 RPM this fluid bursts through the capillary valve to enter the reaction chamber. These images were obtained using a strobe light synchronized with the spin speed of the chip.

To demonstrate the feasibility of using passive valves for differential flow control in our device, we designed a prototype chip with three different types of valve geometries: a straight channel capillary valve, a serpentine capillary valve, and a siphon valve. A 3D-printed custom stage allowed us to secure the chip with screws to a spinning rotor which was controlled with a custom LabView program. **Figure 5.2** shows how sequential fluid flow into and out of a central chamber was achieved by altering the spin speed of the disk.

A simple experiment was also performed to compare the two types of membrane filters' permeability to fluid flow. The same volume of dyed water was added to chambers on the same chip with either the 0.4 µm pore size PETE filter or the 1.0 µm pore size PCTE filter affixed to the bottom. The chip was spun at 1000 RPM for 30 seconds, and as **Figure 5.3** shows almost all of the liquid passed through the 1 µm filter while only a fraction passed through the 0.4 µm filter. However, a transparent material like PETE might be better suited for optical detection of the microbead, so additional filter testing using other pore sizes and materials is needed.

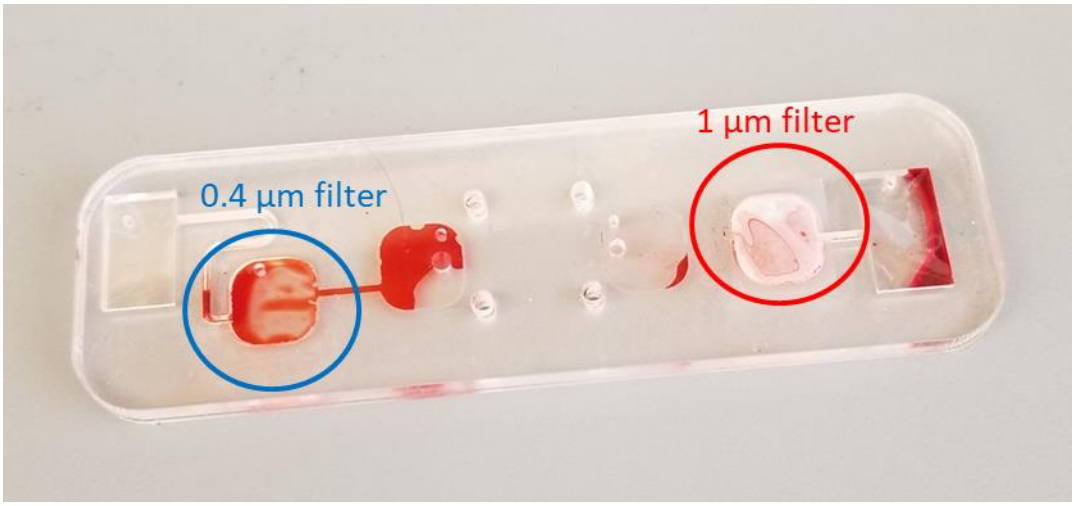


Figure 5.3: Comparison of Two Micropore Membrane Filter for Liquid Flow Permittance

Magnetic Mixing Preliminary Testing

We employed a similar mixing strategy as Grumann et al, in which small square (6.35 x 6.35 mm) neodymium magnets were placed in alternating positions (bordering the inside and outside walls of the microfluidic chip's mixing chamber) underneath the spinning microfluidic chip. The chip's spin direction was also periodically switched, usually either every 30 or 60 seconds. When using high concentrations of microbeads at low to moderate (20-300 RPM) spin speeds it was apparent that this mixing method biased the beads towards either upper or lower walls of the mixing chamber, depending on the spin direction (**Figure 5.4 (A)**). By contrast, if magnets weren't used the centrifugal force tended to send the beads to the outside wall.

To assess the efficacy of this mixing method, the mixing chambers of two chips were loaded with the same amount of p53 peptide antigen-functionalized microbeads (see Chapter 3) and a solution of an anti-p53 monoclonal antibody at 100 ng/mL. In these experiments one chip was always kept stationary as a control, while the other subjected to varying magnetic mixing strategies (magnet placement, spin speed, spin switching interval times). At the end of this primary mixing/incubation step, the microbeads were removed from each chip and placed into a

96-well filter plate where the remaining immunoassay steps (washing, detection antibody incubation, and fluorophore labeling) were carried out using a standard BioRad protocol; the bead fluorescent signals were then read using a BioPlex 200 plate reader. In every experiment, there was no significant difference between the mixed and unmixed/stationary microbeads (**Figure 5.4 (B)-(C)**). Potential reasons for this and ways to possibly rectify this problem are explored in the Conclusions and Discussion section of this chapter.

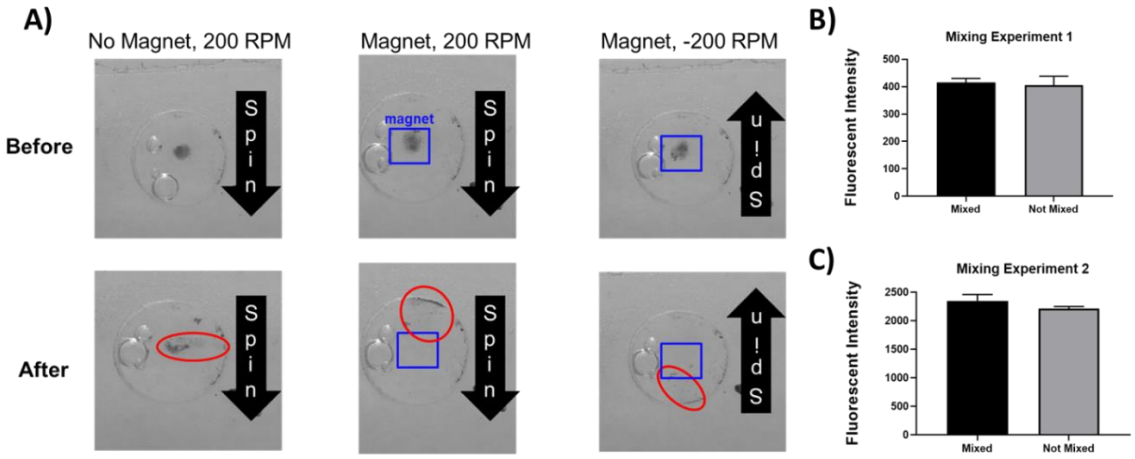


Figure 5.4: Magnetic Microbead Mixing Experiments. **(A)** Not using any magnets and spinning the chip tended to bias the microbeads towards the outer (right) wall, while employing magnets at low to moderate spin speeds would bias the beads in the direction opposite of the spin direction (upper and lower walls). **(B)** In this experiment a 3-layer (acrylic, adhesive microchamber layer, acrylic) chip was spun with 4 uL of total liquid (2 uL microbead solution, 2 uL antibody standard) at speeds of 100 and -100 RPM for 30 minutes, with spin switching every 1 minute. **(C)** In this experiment a 5 layered chip (acrylic, adhesive microchamber layer, acrylic microchamber layer, adhesive microchamber layer, acrylic) with a 70 µL total mixing volume (10 µL microbeads solution + 60 µL antibody standard) was spun at 300 and -300 RPM for 30 minutes with spin switching every 1 minute.

Surface Acoustic Wave (SAW) Mixing

Surface acoustic waves (SAWs) are high frequency (10–1,000 MHz) mechanical waves propagated along the surface of a material rather than through the bulk of it [18]. In BioMEMS SAWs are typically generated using a type of electrode, interdigitated transducers (IDTs), patterned on a piezoelectric substrate like lithium niobate ($LiNbO_3$) [19]. SAWs have been used in a variety of BioMEMS devices to achieve efficient mixing in microchannels, droplets, and chambers [20-22].

SAW Electrode and Prototype Device Manufacturing

Due to the need to manufacture microscale IDTs and bond them to a $LiNbO_3$ substrate, a combination of photolithography, soft lithography, and a molten metal injection method was used to create a prototype microchannel-based device. A summary of these manufacturing steps is shown in **Figure 5.5**. First, photolithography using negative photoresist (SU-8) was used to pattern the microchannels and IDT features on a silicon wafer at a height of 50 μm (**Figure 5.5 (A)**). IDT features were placed near microchannel corners or junctions for maximum SAW propagation down the length of the channel. This wafer was used as a mold for soft lithography to form the device features in polydimethylsiloxane (PDMS). Molten indium was then pulled into the PDMS IDT channels with a vacuum and allowed to cool, forming the electrodes. This injection method was able to form IDT "fingers" with a minimum width of 35 μm. The PDMS containing the microchannels and IDTs was then bonded to a $LiNbO_3$ substrate.

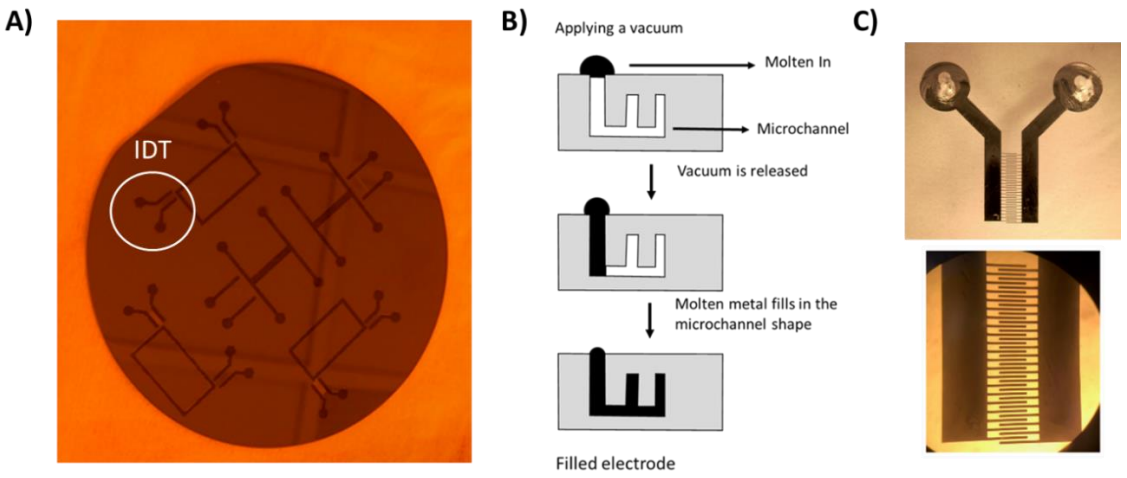


Figure 5.5: SAW IDT and Microchannel Manufacturing. **(A)** The photolithographically patterned mold used to cast the IDT and microchannel features into PDMS via soft lithography. **(B)** The molten metal (indium) injection process for forming the IDT electrodes. **(C)** A completed IDT

SAW Mixing Preliminary Testing

To assess the turbulent mixing ability of the SAWs generated by our IDTs, two streams of dyed water dyed red or green were loaded into the device with a syringe pump. Prior to SAW generation, the two streams exhibited a typical laminar flow profile (**Figure 5.6 (A) & (D)**), with whichever stream that had the higher flow rate assuming a wide profile in the center of the microchannel and the stream with lower flow rate flanking it along the channel walls. SAWs with a frequency of 16 MHz and amplitude of 5.0 dBm were generated by connecting the IDTs to a high frequency generator and amplifier. **Figure 5.6 (B) & (E)** shows the perturbation in the laminar flow after the electrodes were excited. Interestingly, this perturbation was maintained even after the signal generator was shut off (**Figure 5.6 (C) &(F)**), though the time duration of this perturbation was not recorded.

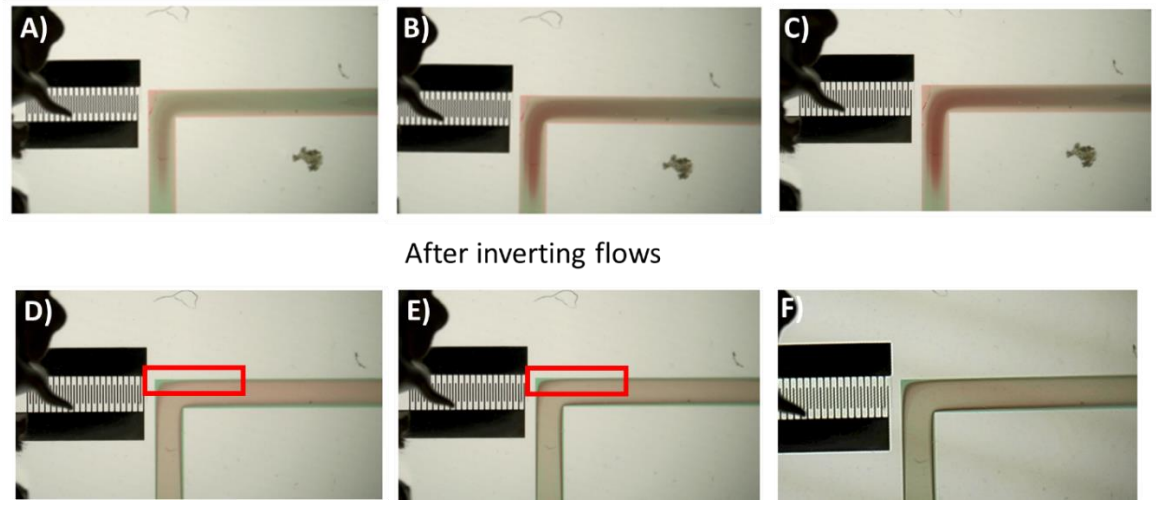


Figure 5.6: SAW Mixing Experiment. The flow profiles of the dyed red and green streams are shown **(A) & (D)** prior to electrode excitation, **(B) & (E)** immediately after electrode excitation, and **(C) & (F)** after the electrode was turned off. The lower images were obtained from inverting the flow rates of the upper images so that the red dyed liquid had the wider central flow profile.

Optical Detection Setup and Preliminary Testing

The previous iteration of our lab's blood analysis device had an integrated optical detection system, consisting of a built-in microscope (10x objective lens, LED light source, interchangeable red and blue filters) and a CMOS camera [10]. While this system sufficed for detecting and imaging white blood cells that are stained with acridine orange, this setup would not be sensitive enough to detect the significantly smaller Luminex microbeads (6.5 µm vs the typical WBC size of 12-15 µm) that are externally labeled with minute amounts of phycoerythrin or a similar fluorophore.

We assessed the detection performance of a silicon photodiode (SPD, Thorlabs, PDA36A, 13 mm^2 area) and an avalanche photodiode (APD, Thorlabs, APD130A, 1 mm diameter active area). The detectors were connected to an oscilloscope (Agilent, DSO1052B) for signal readout in the setup shown in **Figure 5.7 (A)**. P53 peptide antigen-functionalized microbeads (see Chapter 3)

were incubated with anti-p53 monoclonal antibody standards at 0, 25, or 100 ng/mL using a standard BioRad immunoassay kit and protocol, except 100 ug/mL streptavidin-AlexaFluor 488 was used as the labeling fluorophore instead of streptavidin-phycoerythrin. These beads were loaded onto polyacrylic microfluidic chip and localized into central clumps with a neodymium magnet. A 50x objective lens and camera was used to image the beads and position the stage and microfluidic chip, and a blue LED was used as an excitation light source for the AF-488. While the silicon photodiode could not discriminate between the three types of microbeads, the avalanche photodiode did detect a signal difference of 0.9-1.4 mV between the blank and 100 ng/mL microbeads (**Figure 5.7 (C)**).

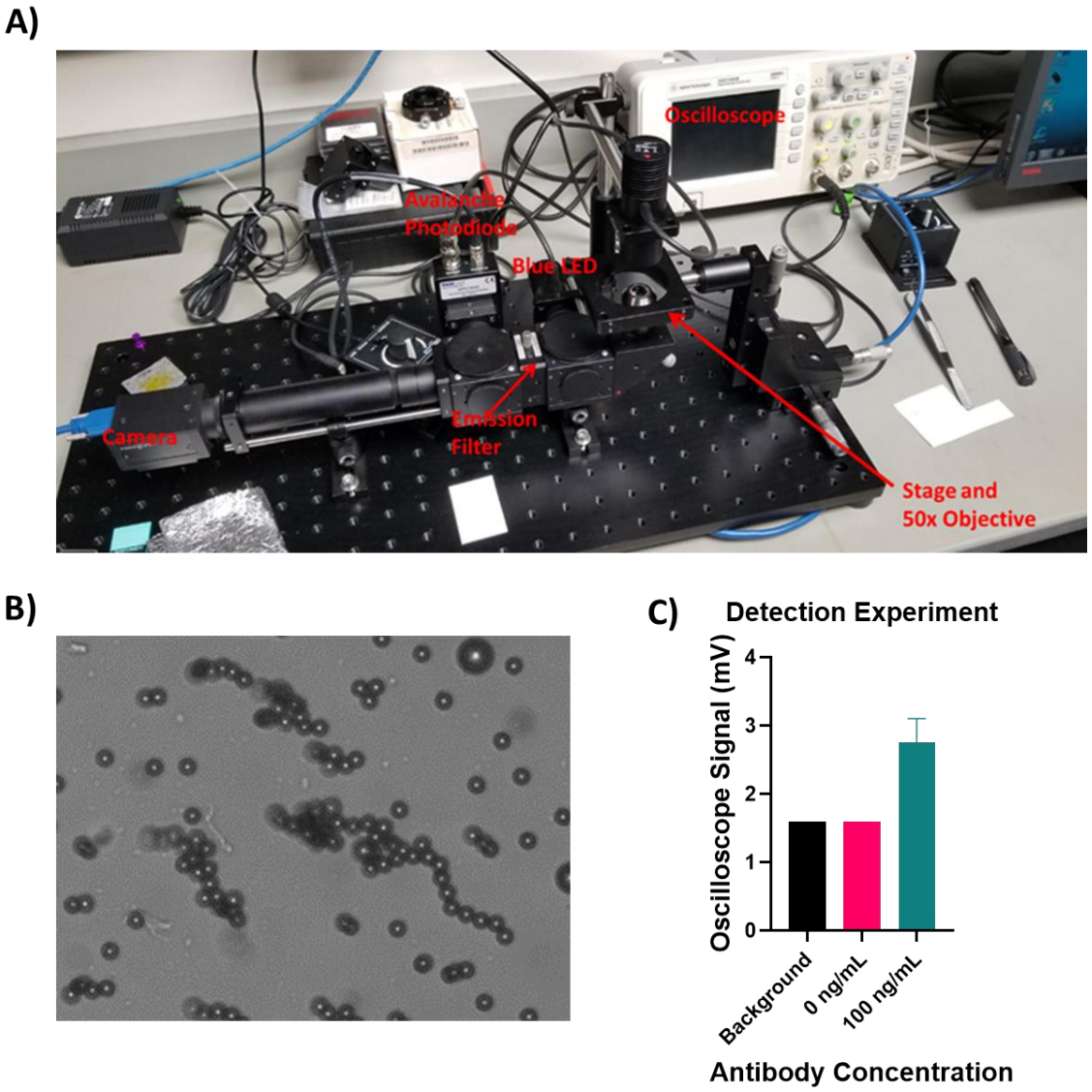


Figure 5.7: Optical Detection System Testing **(A)** The setup used for optical detector testing. A silicon photodiode (not shown) was later substituted with an avalanche photodiode for greater sensitivity. **(B)** An example of a microbead clump imaged and analyzed for fluorescent signal **(C)** Results of a simple detection experiment using the avalanche photodiode detector. "Background" refers to the background signal with no beads present. No detectable difference in fluorescent signal was detected between the background signal and clumps of beads exposed to the negative antibody control, but after controlling for bead number a significant difference

was observed between the negative and positive (100 ng/ml) antibody controls. The oscilloscope settings used were: Channel 1 VPD: 2.00 mV; Time Per Division: 100 ns

Conclusions and Discussion

While the work presented herein is still in the preliminary stages, it presents a valuable jumping-off point for future development of the microfluidic blood diagnostic device. We have presented two options for sample-microbead mixing, demonstrated the feasibility of using microchannel geometry and spin direction switching for sequential flow control, and began to narrow down equipment options for optical detection.

The magnetic mixing method chip design presented here is attractive due to its simplicity and rapid manufacturing time, which considerably saves on time, labor, and material costs. The negative results obtained from the preliminary magnetic mixing experiments do not necessarily invalidate it as mixing option. There are many considerations that should be explored, including using stronger magnets, placing the magnets closer to the chip, and positioning the magnets above (as in the Grumann paper) as well as below the chip. It should be noted that the study by Grumann et al used substantially larger microbeads than us (68 µm vs 6.5 µm), which likely meant their microbeads required higher centrifugal force to pull them to the outside and thus allowed them to use higher spin speeds []. We were also limited in our ability to rapidly switch the spin direction, relying on a manually controlled LabView program to do so. In the future a program should be developed to allow for automated, high frequency spin switching, similar to the frequencies used in the Grumman study.

In contrast to the magnetic mixing method, the SAW-based mixing strategy would require more labor and time intensive chip manufacturing in order to pattern the IDT electrodes. However, the molten metal injection method used here substantially saves time compared to other metal

deposition techniques, though these techniques may prove necessary for manufacturing IDTs with digit widths smaller than 35 µm. Our preliminary mixing experiment and much BioMEMS literature showed support for SAW mixing in a microchannel flow setting, but further experiments with mixing in a batch (microchamber) setting are needed. A central mixing microchamber would likely require at least two IDTs for efficient mixing, so the configuration of these IDTs (parallel vs orthogonal) should also be examined.

The preliminary optical detection experiments showed that while a silicon photodiode was not sensitive enough to detect antibody concentration changes on the functionalized microbeads, the avalanche photodiode could discriminate between the negative and positive control beads. We plan on testing a photomultiplier tube (PMT) detector in the future, which we expect to provide a higher sensitivity than the APD. Other considerations to improve sensitivity and reduce background signal include placing the microfluidic chip stage in a dark housing to reduce ambient light, using a photomask with micro apertures to only image a small area at a time, trying different bead localization methods (magnetization vs centrifugation), and determining an optimal bead concentration.

We envision that once completed, this fully integrated microfluidic device will be able to perform rapid blood diagnostics assays, such as for early detection of cancer, in a point-of-care setting. This device also will bridge a gap between an HSyM application—binding affinity prediction and peptide engineering (Chapter 3)—and a practical clinical use via the incorporation of optimized peptide-functionalized microbeads.

References

1. Foudeh, A. M.; Fatanat Didar, T.; Veres, T.; Tabrizian, M., Microfluidic designs and techniques using lab-on-a-chip devices for pathogen detection for point-of-care diagnostics. *Lab on a Chip* **2012,** *12* (18), 3249-3266.

2. Vashist, S. K., Point-of-Care Diagnostics: Recent Advances and Trends. *Biosensors* **2017,** *7* (4), 62.
3. Hassan, U.; Reddy, B., Jr.; Damhorst, G.; Sonoiki, O.; Ghonge, T.; Yang, C.; Bashir, R., A microfluidic biochip for complete blood cell counts at the point-of-care. *Technology (Singap World Sci)* **2015,** *3* (4), 201-213.
4. Shrirao, A. B.; Fritz, Z.; Novik, E. M.; Yarmush, G. M.; Schloss, R. S.; Zahn, J. D.; Yarmush, M. L., Microfluidic flow cytometry: The role of microfabrication methodologies, performance and functional specification. *Technology (Singap World Sci)* **2018,** *6* (1), 1-23.
5. Lin, Q.; Wen, D.; Wu, J.; Liu, L.; Wu, W.; Fang, X.; Kong, J., Microfluidic Immunoassays for Sensitive and Simultaneous Detection of IgG/IgM/Antigen of SARS-CoV-2 within 15 min. *Analytical Chemistry* **2020,** *92* (14), 9454-9458.
6. Ghodbane, M.; Stucky, E. C.; Maguire, T. J.; Schloss, R. S.; Shreiber, D. I.; Zahn, J. D.; Yarmush, M. L., Development and validation of a microfluidic immunoassay capable of multiplexing parallel samples in microliter volumes. *Lab on a chip* **2015,** *15* (15), 3211-3221.
7. Gorkin, R.; Park, J.; Siegrist, J.; Amasia, M.; Lee, B. S.; Park, J.-M.; Kim, J.; Kim, H.; Madou, M.; Cho, Y.-K., Centrifugal microfluidics for biomedical applications. *Lab on a Chip* **2010,** *10* (14), 1758-1773.
8. Wang, M.; Lomeli, S. H.; Franklin, W. A.; Lee, S.; Pantuck, A. J.; Zeng, G., Optimizing peptide epitope-based autoantibody detection in cancer patients. *Am J Clin Exp Immunol* **2017,** *6* (5), 84-91.
9. Elshal, M. F.; McCoy, J. P., Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA. *Methods* **2006,** *38* (4), 317-323.
10. Balter, M. L.; Chen, A. I.; Colinco, C. A.; Gorshkov, A.; Bixon, B.; Martin, V.; Fromholtz, A.; Maguire, T. J.; Yarmush, M. L., Differential leukocyte counting via fluorescent detection and image processing on a centrifugal microfluidic platform. *Analytical Methods* **2016,** *8* (47), 8272-8279.
11. Balter, M. L.; Leipheimer, J. M.; Chen, A. I.; Shrirao, A.; Maguire, T. J.; Yarmush, M. L., Automated end-to-end blood testing at the point-of-care: Integration of robotic phlebotomy with downstream sample processing. *Technology (Singap World Sci)* **2018,** *6* (2), 59-66.
12. Ward, K.; Fan, Z. H., Mixing in microfluidic devices and enhancement methods. *Journal of micromechanics and microengineering : structures, devices, and systems* **2015,** *25* (9).
13. Owen, D.; Ballard, M.; Alexeev, A.; Hesketh, P. J., Rapid microfluidic mixing via rotating magnetic microbeads. *Sensors and Actuators A: Physical* **2016,** *251*, 84-91.
14. Rida, A.; Lehnert, T.; Gijs, M. *Microfluidic mixer using magnetic beads*; 2003.
15. Grumann, M.; Geipel, A.; Riegger, L.; Zengerle, R.; Ducrée, J., Batch-mode mixing on centrifugal microfluidic platforms. *Lab on a Chip* **2005,** *5* (5), 560-565.
16. Oh, K. W.; Ahn, C. H., A review of microvalves. *Journal of Micromechanics and Microengineering* **2006,** *16* (5), R13-R39.
17. Chen, J. M.; Huang, P.-C.; Lin, M.-G., Analysis and experiment of capillary valves for microfluidics on a rotating disk. *Microfluidics and Nanofluidics* **2008,** *4* (5), 427-437.
18. Yeo, L. Y.; Friend, J. R., Surface Acoustic Wave Microfluidics. *Annual Review of Fluid Mechanics* **2014,** *46* (1), 379-406.
19. Ding, X.; Li, P.; Lin, S.-C. S.; Stratton, Z. S.; Nama, N.; Guo, F.; Slotcavage, D.; Mao, X.; Shi, J.; Costanzo, F.; Huang, T. J., Surface acoustic wave microfluidics. *Lab on a Chip* **2013,** *13* (18), 3626-3649.
20. Ahmed, H.; Park, J.; Destgeer, G.; Afzal, M.; Sung, H. J., Surface acoustic wave-based micromixing enhancement using a single interdigital transducer. *Applied Physics Letters* **2019,** *114* (4), 043702.

21. Shilton, R.; Tan, M. K.; Yeo, L. Y.; Friend, J. R., Particle concentration and mixing in microdrops driven by focused surface acoustic waves. *Journal of Applied Physics* **2008,** *104* (1), 014910.
22. Zhang, Y.; Devendran, C.; Lupton, C.; de Marco, A.; Neild, A., Versatile platform for performing protocols on a chip utilizing surface acoustic wave (SAW) driven mixing. *Lab on a Chip* **2019,** *19* (2), 262-271.

Chapter 6: Dissertation Conclusions

Summary

In this work we have introduced a coarse-grained protein energetics model that is able to quickly and easily convert sequence and structure data from a single protein ground state into a per-residue interaction free energy breakdown. We benchmarked the model by successfully applying it to two very different tasks: predicting point mutation thermostability shifts in T4 Lysozyme (Chapter 2), and using it to design and predict the affinities of engineered peptide antigens for an antibody target for which we had no sequence or structural data (Chapter 3). We also show future promise for the model in optimizing its residue-solvent interaction efficiencies (and potentially other parameters) using a statistical mechanical toy model (Chapter 4), and for bridging the gap between its affinity prediction abilities and a real-world clinical application with a fully integrated microfluidic diagnostic blood analysis platform (Chapter 5).

Model Limitations

As Chapters 2 and 4 demonstrate, the model's relationship with residue-solvent interactions must still be optimized. We believe this will be critical for accurately predicting the effects of mutations at solvent exposed and partially exposed sites. There is also some evidence (not presented in this work) that other parameters of the model—including the amino acid descriptors (γ), propagation length (χ), and solvent/water interaction factor (μ_{water})—may also need to be reassessed.

Our work in Chapter 3 showed that it is likely that certain epitopes and protein-protein/protein-peptide interfaces can reach an "affinity ceiling", in which model-guided modifications will have little to no impact on affinity. It is likely that complete knowledge of an epitope sequence and the target antibody would help to overcome this. Indeed, while the presumed requirement that HSyM needs an experimental protein structure in order to perform optimally is certainly a

limitation, the good results obtained without such a structure in Chapter 3 somewhat provide a counterargument to this.

HSyM as a coarse-grained, scale invariant model will also likely miss out on certain subtle atomic interactions or slight differences in energy between conformations. For this reason, we propose using it as many other coarse-grained models are applied, as a sort of in silico screening tool whose results can be further refined with all-atom methods (MD or MC) or tested with in vitro experimentation.

Future Directions

Presented here are lists of future tasks and experiments related to each topic discussed in this dissertation.

Model Optimization

- Use the toy model to optimize solvent interaction efficiency values by assessing a variety of options, including stepwise and linear functions and/or binary designations based only on residue identity.
- Use the toy model to optimize the solvent/water interaction factor term, and determine if a temperature dependence of this term is applicable
- Conduct sensitivity analyses on other model parameters (χ, γ, λ) to determine if they require further optimization

Thermostability Prediction

- Assess model's performance on a variety of mutation types (buried, solvent exposed, and partially buried) in other thermodynamically well-characterized proteins (barnase, staph nuclease, etc.)

Affinity Prediction & Peptide Engineering

- Select another known linear epitope and accompanying monoclonal antibody to determine if the model can make modifications outside of the native antigenic sequence that will improve affinity
- Select and optimize other known linear epitopes from p53 and other tumor antigens to create a diagnostic cancer autoantibody panel and test these against clinical sera samples
- Assess the model's ability to predict epitopes: start by focusing on warm/hot solvent exposed regions of antigenic proteins, isolate these linear or conformational regions, and test against polyclonal Abs.

Microfluidic Device

- Complete magnetic and SAW mixing experiments and choose a mixing method and accompanying microfluidic chip design.
- Test the detection performance of a PMT and develop an optical detection system protocol with considerations for bead localization, optimal bead concentration, size of the detection area, and positive signal threshold.
- Develop a LabView program to automate steps of the assay protocol, including sample/reagent loading and spin speed control for fluid addition, removal, and mixing steps.