

Multistate Computational Protein Design: Theories, Methods, and Applications

James A. Davey

**A thesis submitted to the Faculty of Graduate and Postdoctoral Studies in
partial fulfillment of the requirements of the PhD degree in Chemistry**

Department of Chemistry

Faculty of Science

University of Ottawa

© James A. Davey, Ottawa, Canada 2016

TABLE OF CONTENTS

List of Tables.....	vi
List of Figures	viii
List of Abbreviations.....	xiii
Abstract	xv
Acknowledgements.....	xvi
Chapter 1. Single-state Computational Protein Design Theory	1
1.1. The Chemical Structure of Proteins	2
1.2. The Complexity of Structure and Sequence Space	10
1.3. Template Preparation Methodology	13
1.4. Rotamer Libraries	14
1.5. Energy Functions	16
1.6. Sequence Optimization Algorithms	19
1.7. Protein Engineering Examples Aided by Single-state Design.....	21
Chapter 2. Multistate Computational Protein Design Theory.....	23
2.1. Single-state Design Challenges.....	25
2.2. Multistate Approaches to Computation Protein Design	28
2.3. Positive and Negative Multistate Design	31
2.4. The Practical Implementation of Multistate Design	33
2.5. Multistate Analysis and Multistate Design	35
2.6. Protein Engineering Examples with Multiple Conformational States	37
2.7. Protein Engineering Examples with Multiple Chemical States	44
2.8. Multistate Computational Protein Design Challenges.....	53

2.9. Thesis Hypothesis and Rationale.....	56
Chapter 3. Ensemble Generation Methodologies.....	58
3.1. Issues with Single-state Design	60
3.2. The Perturbation Followed by Energy Minimization Algorithm	63
3.3. Behaviour of the PertMin Algorithm	67
3.4. Design Hypothesis and Training Set Sequences.....	70
3.5. Creation and Characterization of G β 1 Ensembles	74
3.6. Protein Stability Design Tasks	80
3.7 Enrichment of Sequence Predictions.	85
3.8. Binning of Sequence Predictions.....	86
3.9. Correlation of Sequence Predictions	96
3.10. The Effect of Ensemble Size on Multistate Design Calculations	98
3.11. Calculation Framework and Sequence Prediction Energetics	99
3.12. Calculation Framework and Sequence Prediction Diversity	102
3.13. Properties of On-Target and Off-Target Ensembles	104
3.14. Improvements to Ensemble Based Multistate Design of Protein Stability.....	108
3.15. Conclusions on the Study of Ensemble Generation Methodologies	110
Chapter 4. Rotamer Bias Methodologies	112
4.1. Structure and Sequence Space Relationship	114
4.2. Sequence Binning with SSD Calculations using Rotamer Optimized and Energy Minimized Templates.....	117
4.3. Sequence Enrichment with SSD Calculations using Rotamer Optimized and Energy Minimized Templates.....	124
4.4. Sequence and Structural Determinants of Rotamer Bias	130

4.5. Sequence Enrichment using Energetically Biased Single-state Design	143
4.6. Application of Rotamer Biased Templates to Multistate Analysis.....	154
4.7. Conclusions on the Study of Rotamer Bias Methodologies.....	160
Chapter 5. Applied Positive Multistate Design: Predicting the Substrate Multi- specificity of SMYD2 Methyltransferase	161
5.1. Lysine Methylation in Biology	163
5.2. Prediction of the SMYD2 Recognition Motif by Multistate Design	165
5.3. Prediction of the SMYD2 Recognition Motif by SPOT Peptide Array	175
5.4. Calculation Comparison with Experiment.....	179
5.5. Discovery of New SMYD2 Substrates.....	184
5.6. Conclusions on the Application of Positive Multistate Design	191
Chapter 6. Applied Negative Multistate Design: Predicting the Stability of Proteins and their Mutants	193
6.1. Accurate Prediction of Protein Stability.....	195
6.2. A State Independent Framework.....	198
6.3. Training Set Stability Correlation	201
6.4. Test Set Design and Correlation.....	215
6.5. Stability Predictions Using Amino Acid Reference State Energies.....	225
6.6. Stability Predictions for Alternate Protein Folds.....	230
6.7. Comparative Study of Alternate Stability Prediction Strategies.....	236
6.8. Features of Physical and Predictive Models	241
6.9. Conclusions on the Application of Negative Multistate Design.....	244
Chapter 7. Development and Application of <i>meta</i>-Multistate Design: the Rational Design of Protein Dynamics	245

7.1. The Importance of Protein Dynamics	247
7.2. The Design Objective, Hypothesis, and Rationale	249
7.3. Ensemble Preparation	261
7.4. Sequence Space and Calculation Results	266
7.5. Experimental Characterization of DANCER Proteins	275
7.6. Validation of DANCER Conformational States	287
7.7. Comparison of DANCER <i>meta</i> -Multistate Design Models and Solution NMR Structures	297
7.8. Implications of Mega Ensemble Model Space on <i>meta</i> -Multistate Design Prediction Outcomes	306
7.9. Conclusions on the Application of <i>meta</i> -Multistate Design	311
Chapter 8. Conclusions	313
8.1. Applications	314
8.2. Methods	318
8.3. Theories	321
8.4. Summary	323
References	325

LIST OF TABLES

Table 3.1. Potential energy and RMSG of backbone templates and ensembles.....	78
Table 3.2. Sequence enrichment of SSD and MSD calculations	85
Table 3.3. Sequence binning of SSD and MSD calculations	90
Table 3.4. Potential energy and RMSG of minimized ensembles	91
Table 3.5. Sequence binning results of MSD using energy minimized ensembles.....	92
Table 4.1. Sequence binning results for MIN and ROM prepared templates.....	122
Table 4.2. Sequence enrichment results for MIN and ROM prepared templates.....	129
Table 4.3. Similarity and identity of G β 1 benchmark sequences	131
Table 4.4. Sequence enrichment results for scROM and bbROM prepared templates.....	141
Table 4.5. Sequence enrichment results for energetically biased SSD.....	149
Table 4.6. Sequence binning predictions of energetically biased SSD calculations	151
Table 4.7. Sequence binning backbone drift predictions with ROM templates	156
Table 5.1. p53-SMYD2 interaction energy calculated by MSD	172
Table 5.2. k-Means cluster analysis of substitution interaction energy	173
Table 5.3. SMYD2 interactors containing the MSD and SPOT recognition motifs.....	185
Table 6.1. Experimental stabilities for G β 1 training set sequences.....	204
Table 6.2. MSD of G β 1 training set sequences using various ensembles.....	206
Table 6.3. Multiple linear regression for G β 1 training set sequences.....	209
Table 6.4. Calculated stabilities for G β 1 training set sequences	214
Table 6.5. Experimental and calculated G β 1 test set sequence stabilities.....	216
Table 6.6. Amino acid reference energies obtained from pentapeptide models	226

Table 6.7. Reference state energies for G β 1 training set and test set sequences	227
Table 6.8. Regression of training set sequence with reference state energies.....	228
Table 6.9. Calculated and experimental stabilities for alternate protein folds	232
Table 6.10. Stability regression parameters for alternate protein folds	236
Table 6.11. Structural characterization of backbone ensembles	242
Table 7.1. Seed tryptophan rotamer state geometries	263
Table 7.2. Properties of the total model ensemble structure space	266
Table 7.3. State conformation energies for sequences predicted stable	269
Table 7.4. G β 1 sequences predicted static and dynamic by <i>meta</i> -MSD	273
Table 7.5. Thermodynamic stability by thermal and chemical denaturation of test set proteins	282
Table 7.6. Kinetic parameters of DANCER proteins determined by ZZ-exchange NMR spectroscopy	286
Table 7.7. RMSD coordinate comparisons of solution NMR ensembles	290
Table 7.8. Ensemble tryptophan conformations for solution NMR structures and models of sequences H, K, and the major and minor species of DANCER-2	293
Table 7.9. Tryptophan conformations found in model <i>meta</i> -MSD sub-states	300
Table 7.10. RMSD coordinate comparisons of solution NMR ensembles against sub-state design models calculated by <i>meta</i> -MSD.....	304

LIST OF FIGURES

Figure 1.1. The twenty canonical alpha-amino acids.....	3
Figure 1.2. Dihedral angle descriptors for protein conformational space	5
Figure 1.3. The primary, secondary, and tertiary structure of the G β 1 fold	8
Figure 1.4. The inverse protein folding problem	12
Figure 2.1. An example of single-state CPD	25
Figure 2.2. Examples of multistate CPD	29
Figure 2.3. Single-state and multistate design.....	30
Figure 2.4. Sequence evaluation in positive and negative design	32
Figure 2.5. Frameworks for multistate analysis and design	36
Figure 2.6. Input templates for positive multistate design examples	38
Figure 2.7. Input templates for negative multistate design examples	46
Figure 2.8. Negative state selection and its impact on sequence evaluation	54
Figure 3.1. The single-state design artefact.....	60
Figure 3.2. The coordinate perturbation and energy minimization method	64
Figure 3.3. The coordinate perturbation and energy minimization algorithm	66
Figure 3.4. Tuning PertMin ensemble properties with minimization algorithm	67
Figure 3.5. Tuning PertMin ensemble properties with perturbation size	68
Figure 3.6. Tuning PertMin ensemble properties with system size	69
Figure 3.7. On- and off-target ensembles.....	71
Figure 3.8. G β 1 core residue design sequence space.....	72
Figure 3.9. G β 1 training set sequences.....	73
Figure 3.10. G β 1 backbone ensembles.....	75

Figure 3.11. Structural comparison of G β 1 backbone ensembles	77
Figure 3.12. Protein stability design tasks	82
Figure 3.13. Sequence binning of SSD and MSD calculations	88
Figure 3.14. ROC curves for binning results from MSD with all ensembles	94
Figure 3.15. Correlation of MSD calculation results.....	97
Figure 3.16. MSD with ensembles of various sizes.....	99
Figure 3.17. Calculation framework effect on predicted sequence energy	100
Figure 3.18. Calculation framework effect on binning predicted sequences	101
Figure 3.19. Calculation framework effect on predicted sequence diversity	103
Figure 4.1. Rotamer optimization followed by energy minimization	115
Figure 4.2. G β 1 benchmarking mutant sequences	118
Figure 4.3. Enrichment profiles for SSD including ROM-minimized rotamers	125
Figure 4.4. Enrichment profiles for SSD excluding ROM-minimized rotamers.....	127
Figure 4.5. ROM prepared backbone template RMSD comparisons.....	132
Figure 4.6. ROM prepared template side chain variability.....	135
Figure 4.7. Energy contact map for ROM _{XTAL} templates	138
Figure 4.8. Energy contact map for ROM _{NMR} templates.....	139
Figure 4.9. Sequence enrichment profiles for SSD with scROM and bbROM prepared templates	142
Figure 4.10. Sequence enrichment profiles for amino acid biased MIN _{XTAL} SSD	145
Figure 4.11. Sequence enrichment profiles for amino acid biased MIN _{NMR} SSD	146
Figure 4.12. Sequence enrichment profiles for configuration biased MIN _{XTAL} SSD	147
Figure 4.13. Sequence enrichment profiles for configuration biased MIN _{NMR} SSD	148

Figure 4.14. Seed sequence ranking by biased SSD calculations	152
Figure 4.15. Sequence space of backbone drift modeling with ROM _{XTAL} templates	158
Figure 5.1. The p53-SMYD2 peptide-enzyme complex.....	166
Figure 5.2. The PertMin ensemble of the p53 peptide.....	169
Figure 5.3. p53-SMYD2 peptide-enzyme interactions	170
Figure 5.4. The SMYD2 recognition motif predicted by MSD	174
Figure 5.5. SPOT assay recapitulates the methylation activity of SMYD2	175
Figure 5.6. Autoradiography of an example SPOT peptide array	177
Figure 5.7. Heat map representation of autoradiographic SPOT array results	178
Figure 5.8. The SMYD2 recognition motif predict by SPOT peptide array	179
Figure 5.9. Sequence binning the MSD predicted recognition motif	181
Figure 5.10. Comparison of motif predictions by SSD and MSD	182
Figure 5.11. SMYD2 interactors and selected gene ontology terms.....	186
Figure 5.12. Refined list of SMYD2 interactors.....	188
Figure 5.13. Discovery of novel SMYD2 substrates	189
Figure 5.14. SMYD2 methylates the recognition motif of novel substrates	190
Figure 6.1. Multistate design methodologies	199
Figure 6.2. G β 1 backbone ensembles.....	203
Figure 6.3. Correlation of MSD and experimental G β 1 training set stabilities.....	208
Figure 6.4. Training set sequence stabilities by positive and negative MSD	213
Figure 6.5. Far-UV CD spectroscopy of G β 1 test set sequences.....	218
Figure 6.6. NMR study of G β 1 wild type and mutant sequence H sequences	220
Figure 6.7. Chemical denaturation of G β 1 test set sequences.....	222

Figure 6.8. Prediction of G β 1 test set sequence stabilities	223
Figure 6.9. Regression of G β 1 test set sequence stabilities	224
Figure 6.10. Accuracy of positive and negative MSD predictions	225
Figure 6.11. Positive and negative state ensembles for alternate protein folds	231
Figure 6.12. MSD correlations of stability for alternate protein folds	235
Figure 6.13. MSD of G β 1 training set stability with backbone drift modeling.....	239
Figure 7.1. Tryptophan conformations for G β 1 design states.....	251
Figure 7.2. The tryptophan χ_1 and χ_2 potential energy landscape	253
Figure 7.3. Permitted exchange of tryptophan between the defined states	254
Figure 7.4. Optimization functions and the design of protein dynamics	257
Figure 7.5. The design of protein dynamics by <i>meta</i> -multistate design	259
Figure 7.6. The step-wise mega ensemble generation strategy	262
Figure 7.7. G β 1 dynamics design sequence space.....	267
Figure 7.8. Size exclusion chromatograms of test set proteins	276
Figure 7.9. Thermal denaturation of test set proteins	278
Figure 7.10. Chemical denaturation of test set proteins.....	281
Figure 7.11. Secondary structure of test set proteins by NMR chemical shift index analysis.....	283
Figure 7.12. ^1H - ^{15}N -HSQC NMR spectroscopy of DANCER proteins	285
Figure 7.13. ^1H - ^{15}N -HSQC NMR spectroscopy of static G β 1 proteins	285
Figure 7.14. Solution NMR ensembles of DANCER-2 protein.....	289
Figure 7.15. Test set tryptophan 43 ^{15}N -NOE correlations	291
Figure 7.16. Solution NMR ensembles of static test set proteins	295

Figure 7.17. Tryptophan 43 orientation relative to its ^{15}N -NOE correlations for the minor species of DANCER proteins.....	297
Figure 7.18. Comparative analysis between calculated and experimental tryptophan sub-state geometries.....	301
Figure 7.19. Ensemble structure space and <i>meta</i> -MSD predictions	308

LIST OF ABBREVIATIONS

BBAT2 – beta-beta-alpha tetramer 2

BMEC – backbone minimum energy configuration

CD – circular dichroism

Cdc42 – cell division control protein 42 homolog

Cm – concentration of chemical denaturant at the midpoint of denaturation

CPD – computational protein design

ciBR – conditional iterative branch relaxation

cMD – constrained molecular dynamics

df – degrees of freedom

DNA – deoxyribonucleic acid

FASTER – fast and accurate side chain topology and energy refinement

FMEC – FASTER minimum energy configuration

Gβ1 – protein G domain beta-1

GMEC – global energy minimum configuration

GTP – guanosine-5'-triphosphate

iBR – iterative branch relaxation

IPTG – isopropyl β-D-1-thiogalactopyranoside

ITSN – GTPase exchange factor named intersectin

MC – Monte Carlo

MD – molecular dynamics

MOE – molecular operating environment

MSA – multistate analysis

MSD – multistate design

NMR – nuclear magnetic resonance

NP-hard – non-deterministic polynomial time hard

mRNA – messenger ribonucleic acid

PDB – protein data bank

PertMin – coordinate perturbation followed by energy minimization

PKMT – protein lysine methyltransferase

PTM – post-translational modification

ROC – receiver operating characteristic

ROM – rotamer optimized and energy minimized

RMS – root mean square

RMSD – root mean square deviation

RMSG – root mean square gradient

SA – simulated annealing

SPR – surface plasmon resonance

SSD – single-state design

uMD – unconstrained molecular dynamics

WT – wild type

ABSTRACT

Traditional computational protein design (CPD) calculations model sequence perturbations and evaluate their stabilities using a single fixed protein backbone template in an approach referred to as single-state design (SSD). However, certain design objectives require the explicit consideration of multiple conformational states. Cases where a multistate framework may be advantageous over the single-state approach include the computer aided discovery of new enzyme substrates, the prediction of protein stabilities, and the design of protein dynamics. These design objectives can be tackled using multistate design (MSD). However, it is often the case that a design objective requires the consideration of a protein state having no available structure information. For such circumstances the multistate framework cannot be applied. In this thesis I present the development of two template and ensemble preparation methodologies and their application to three projects. The purpose of which is to demonstrate the necessary ensemble modeling strategies to overcome limitations in available structure information. Particular emphasis is placed on the ability to recapitulate experimental data to guide modelling of the design space. Specifically, the use of MSD allowed for the accurate prediction of a methyltransferase recognition motif and new substrates, the prediction of mutant sequence stabilities with quantitative accuracy, and the design of dynamics into the rigid G β 1 scaffold producing a set of dynamic variants whose tryptophan residue exchanges between two conformations on the millisecond timescale. Implementation of both the ensemble, coordinate perturbation followed by energy minimization (PertMin), and template, rotamer optimization followed by energy minimization (ROM), generation protocols developed here allow for exploration and manipulation of the structure space enabling the success of these applications.

ACKNOWLEDGEMENTS

There are many people whom I must express my sincere gratitude toward as the work presented in this thesis would not have been possible without their contributions.

To Dr. Roberto A Chica, I want to say how grateful I am to have received your supervision, particularly for your guidance regarding the preparation of presentations and manuscripts. That you gave me the opportunity to independently carry out this research and that you trusted me to freely explore and develop the topics in this thesis was so very important to me. I will miss our discussions, replete with excitement and argument, because the focus was always on the prospect of our next exciting scientific endeavour. I also look forward to seeing you at future conferences, because I know that they'll be full of great food and good beer.

To Dr. Natalie K Goto, I consider you to be the greatest mentor that I have met during the course of my graduate studies. Your support of my education improved it immeasurably so. If I could adopt your extreme scientific rigor and attention to detail, I think that I would be a better scientist.

To Adam M Damry and Christian K Euler whose laboratory education was entrusted to me, and to my fellow colleague Dr. Sylvain Lanouette, all of whom made essential contributions to the completion of the projects documented in this thesis. I would not have been able to successfully complete the task of validating my calculations in the laboratory on my own. That you all devoted so much of your time toward testing these hypotheses, was invaluable. I'm very excited to see what future projects you will all publish.

To my fellow colleagues: Curtis JW Walton, Antony D St.-Jacques, Matthew G Eason, Antonia T Pandelieva, Frédéric Thiebaut, Guido F Calderini, Samuel K Oteng-Pabi, and Dr. Christopher M Clouthier. I want to express my gratitude to you all of your support.

I would also thank friends and family members whose encouragement during the course of my studies was greatly appreciated. Particularly to my sister Sarah C Davey for being such an inspiring role model. Last, I would like to especially thank and acknowledge my wife, Julie M Levionnois, whose patience, compassion, and inquiring nature serves to continually enrich my life.

CHAPTER 1

Single-state Computational Protein Design Theory

As this thesis contains topics on the theories, methodologies, and applications of multistate computational protein design (CPD), I have included a brief introductory chapter detailing the theory of single-state design (SSD) calculation framework. First, the sequence-structure relationship of proteins is introduced. Next, the complexity of the protein folding problem is described. This is followed by a description of the inverse protein folding problem and the tools used to conduct CPD. Specifically, template model preparation, rotamer configurations, potential energy functions, and optimization algorithms are introduced. This chapter lays the ground work for the studies it precedes.

1.1. The Chemical Structure of Proteins

As this thesis is written on topics concerning the computational design of protein properties, it is essential that it is introduced with a review on the relationship between protein structure and sequence. This is because the primary function of CPD involves the search of candidate protein sequences and an evaluation of their ability to stabilize model protein structures. For CPD to be employed effectively, all design hypotheses rely on the validity of this relationship. It is only with a clear and comprehensive understanding of the protein structure and sequence space relationship, that the design of proteins for any property or desired complex function can be achieved. Thus, the organization of protein structures and the chemical identity and structural diversity of the amino acids that comprise protein structures is central to the formulation of the computational methodology.

Proteins are biologically synthesized organic (composed of hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S) atoms) macromolecules responsible for facilitating many of the physical interactions and chemistries essential to the existence of life. From the perspective of a molecular topological graph [Schultz, 1989; Schultz *et al.*, 1990], organismal proteins are linear polypeptide chains comprised of alpha-amino acid monomer units, referred to as residues. All alpha-amino acid residues, referred to henceforth simply as amino acids, are zwitterionic under physiologically relevant conditions (pH = 7.4), consisting of a central alpha-carbon bearing both amine (pKa $\approx$ 9.4) and carboxylic acid functional groups (pKa $\approx$ 2.2) [McMurray, 2004]. This set of atoms (N, C $_{\alpha}$, C, and O), common to all amino acids, are referred to as the backbone

atoms. Amino acids are distinguished from one-another according to a third alpha-carbon substituent referred to as the side chain (R-group). While there are a variety of chemically diverse amino acid side chains, the universal genetic code specifically encodes for twenty amino acids [Figure 1.1].

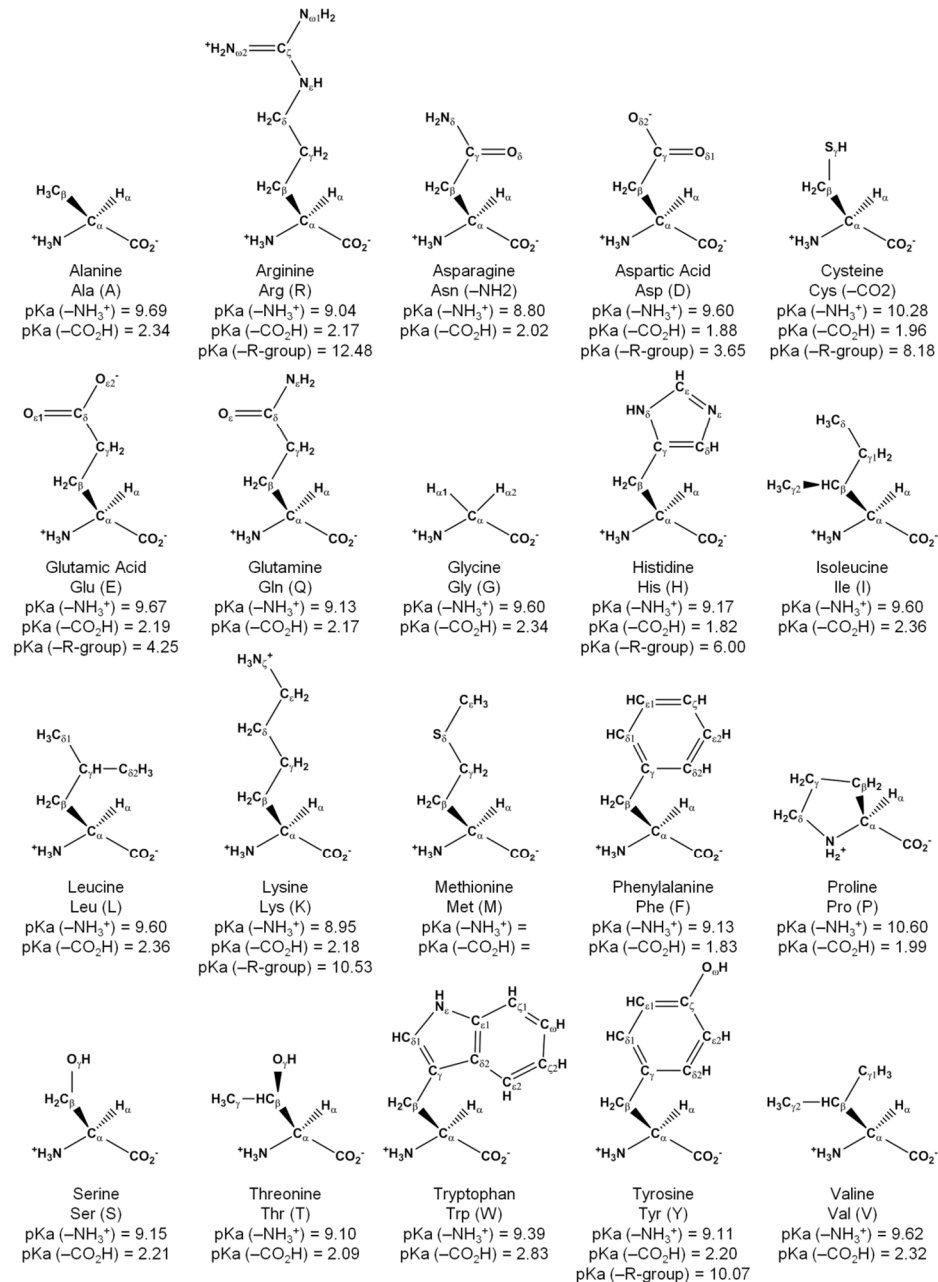


Figure 1.1. The twenty canonical alpha-amino acids. Chemical structures for the twenty canonical alpha-amino acids in their physiologically relevant state (pH = 7.4), shown in

alphabetical order with their three-letter code and single-letter code in parentheses. The pKa values for each amino acid is provided for the amino ($-\text{NH}_3^+$) and carboxyl ($-\text{CO}_2\text{H}$) termini, as well as their side chain ($-\text{R}$ -group) if it contains an ionisable functional group at or near physiological conditions [McMurray, 2004].

The twenty *canonical* amino acids are all alpha-carbon L-stereoisomers, with the exception of the achiral glycine amino acid [Clayden *et al.*, 2007]. Together, they comprise the *proteinogenic* building blocks from which all proteins are synthesized *in vivo* via the biological processes of deoxyribonucleic acid (DNA) transcription and messenger ribonucleic acid (mRNA) translation. During mRNA translation, a protein is synthesized in sequence via successive condensation reactions between the amine and carboxylic acid functional groups of two amino acids producing an amide bond covalently linking adjacent residues in the polypeptide chain. The termini of the polypeptide chain consists of the unreacted amine of the first residue incorporated into the chain and the unreacted carboxylic acid of the last residue incorporated into the chain, referred to as the protein N-terminal and C-terminal, respectively. The sequence of amino acid residues incorporated into the protein polypeptide chain is dictated by the transcribed mRNA. In this way, the evolution of protein sequence is made possible through the permutation, natural selection, and heritable transfer of genetic information between generations of organisms.

The three dimensional configuration adopted by a polypeptide chain can be described by a set of internal atomic coordinates, which include bond lengths, bond angles, and bond dihedral angles. Importantly, the configuration of backbone atoms are described by three bond dihedral angles [Figure 1.2.A] named phi (ϕ : $\text{C}-\text{N}-\text{C}_\alpha-\text{C}$), psi (ψ :

N-C $_{\alpha}$ -C-N), and omega (ω : C $_{\alpha}$ -C-N-C $_{\alpha}$). The phi and psi dihedral angles describe the bond torsion angles before and after the alpha carbon for each residue while the omega dihedral angle describes the bond torsion for the peptide bond joining adjacent residues. Side chain dihedral angle configurations, denoted and referred to as chi (χ) angles, are also provided with a systematic nomenclature numbered according to their distance from the alpha-carbon. For example, the side chain of lysine [Figure 1.2.B] has four bond torsion dihedral angles named chi₁ (χ_1 : N-C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$), chi₂ (χ_2 : C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$ -C $_{\delta}$), chi₃ (χ_3 : C $_{\beta}$ -C $_{\gamma}$ -C $_{\delta}$ -C $_{\epsilon}$), and chi₄ (χ_4 : C $_{\gamma}$ -C $_{\delta}$ -C $_{\epsilon}$ -N $_{\zeta}$).

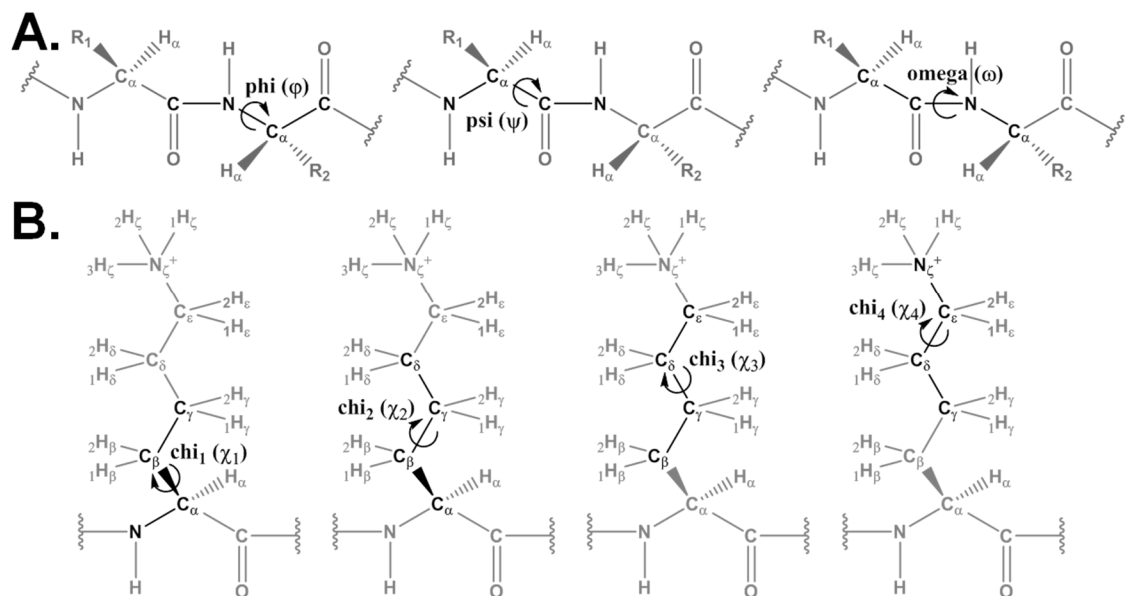


Figure 1.2. Dihedral angle descriptors for protein conformational space. Dihedral angles (black) are shown for backbone atoms (A) and side chain atoms (B).

Protein structure can be classified into a hierarchical schematic of increasing structural complexity. The chemical identity of the residues that make up the

polypeptide chain constitutes the primary structure of a protein, commonly referred to as the protein sequence. The primary structure of a protein can be organized into segments of secondary structure. Protein secondary structures are classified according to the coordinate configuration and non-covalent interactions of residue backbone atoms. The backbone dihedral angle ω , shared between adjacent residues, predominantly adopts an antiperiplanar conformation with the amine (N–H) and carbonyl (C=O) of the amide bond favoring the *trans* (antiperiplanar) over the *cis* (synperiplanar) configuration in approximately a 1000 to 1 ratio. The exception being proline, and other *non-canonical* N-substituted amino acids, favoring *trans* over the *cis* configuration in an approximately 3 to 1 ratio. The restriction to free rotation of the amide bond is due to its partial double bond character, principally the result of electronic delocalization of the nitrogen lone-pair electrons donated into the antibonding pi molecular orbital of the carbonyl ($n \rightarrow \pi^*$) [Kemnitz and Loewen, 2007]. The phi and psi backbone dihedral angles, representing bond torsions between the alpha carbon (sp^3) with the amine nitrogen (sp^2) and carbonyl carbon (sp^2) for each residue, respectively, can adopt a variety of conformations (bond torsions about sp^3 – sp^2 centers: 0° , $\pm 60^\circ$, $\pm 120^\circ$, and 180°) [Mayo *et al.*, 1990].

There are many secondary structures [Andersen *et al.*, 2002] that can be found across structural depositions in the protein data bank (PDB), including: helix, strand, loop, turn, and undefined random coil arrangements of the polypeptide backbone. Three helical arrangements include the 3_{10} -helix, alpha-helix, and pi-helix, which can be distinguished from one another as each contains a hydrogen-bonding interaction

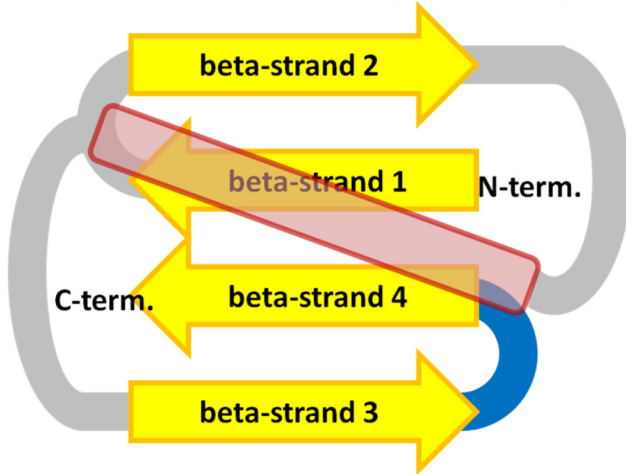
between the amine (N-H) and carbonyl (C=O) of backbone amides at intervals of three, four, and five residues apart along the segment of the helix secondary structure, respectively. The most common helical secondary structure arrangement, the alpha-helix, typically adopts ϕ and ψ backbone dihedral angles of -60° and -45° , respectively [Ramachandran *et al.*, 1963]. The beta-strand secondary adopts a twisted planar backbone configuration with ϕ and ψ backbone dihedral angles occupying conformations at approximately -135° and 135° , respectively [Ramachandran *et al.*, 1963]. In many proteins, beta-strands are typically found in planar sheet arrangements such that backbone hydrogen-bonding interactions between multiple distinct segments of beta-strands can be adopted. The N-terminal to C-terminal orientation of these beta-strands, relative to each other, defines the type of beta-sheet observed, with beta-strands sharing the same orientation referred to as parallel beta-sheets, and those in opposing orientations referred to as anti-parallel beta-sheets. Turns are a broad classification of secondary structure involving three to five residue segments having specific backbone dihedral angles and hydrogen-bonding arrangements [Hutchinson and Thornton, 1994; Venkatachalam, 1968]. The major feature shared between turn secondary structures is that residues positioned at the beginning and end of the segment adopt orientations in opposite directions. The relative three dimensional arrangements of a protein's secondary structure and the set sidechain non-covalent interactions define a protein's tertiary structure, often referred to as the protein's fold. Because of degree of conformational freedom across the polypeptide chain is vast, proteins have been observed to adopt a diverse number of intricate and complex tertiary structures [Bajaj

and Blundell, 1984]. The primary, secondary, and tertiary structures of the well studied globular *Streptococcal* protein G domain beta-1 (G β 1) fold [Gallagher *et al.*, 1994], used to benchmark the methodologies of this thesis, are illustrated in Figure 1.3.

A. Assignment of G β 1 Primary and Secondary Structure

->>>>>>----->>>>>>-----~~~~~~~~~~~~~~~~----->>>>>>LLLL>>>>>>-
 MTYKLIILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWYDDATKTFTVTE
 123456789 123456789 123456789 123456789 123456789 123456
 10 20 30 40 50

B. Cartoon of G β 1 Secondary and Tertiary Structure



C. Molecular Rendering of G β 1 Tertiary Structure

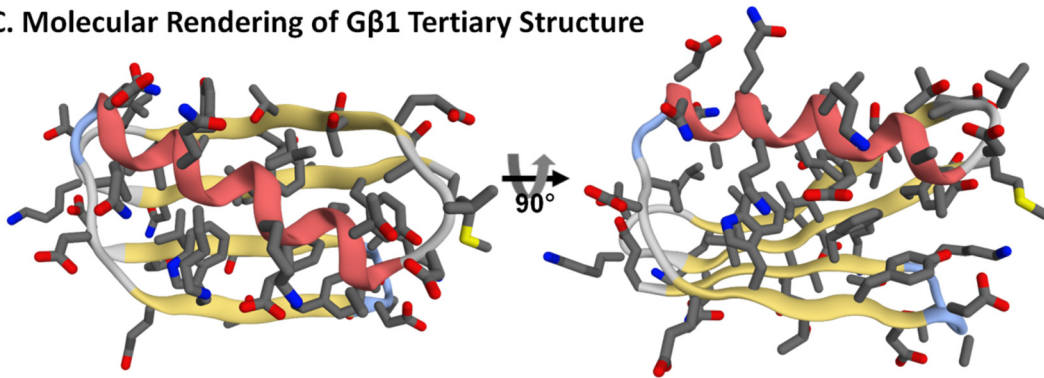


Figure 1.3. The primary, secondary, and tertiary structure of the G β 1 fold. The wild type sequence (A: primary structure) of the G β 1 is 56 residues in length with residues indicated using their single-letter nomenclature from N-terminal to C-terminal. These residues are assigned to secondary structures (A: secondary structure) indicating alpha-helix ($\approx$), beta-strand ($>$), turn (L), and undefined (-) secondary structures. These secondary structures are arranged in the GB1 fold (B: secondary structure) as depicted

in a cartoon representation with one alpha-helix (red bevelled rectangle), four beta-strands (yellow arrows), one four residue turn (blue arc), and three segments of undefined secondary structure (grey lines). The beta-strands are arranged in a sheet, with the first and second beta-strands, first and fourth beta-strands, and third and fourth beta-strands in anti-parallel, parallel, and anti-parallel configurations, respectively. The tertiary structure of the G β 1 fold (C: tertiary structure) is rendered depicting the arrangement of secondary structures and residue side chains.

The function and behaviour of protein folds can also vary considerably. For example, globular proteins generally adopt compact stable folds having an assortment of secondary structures, while the polypeptide chains of fibrous proteins assemble into quaternary structures (arrangements of polypeptide tertiary structures) to form structural scaffolds such as collagen, elastin, keratin, and fibroin [Voet and Voet, 2004]. Aberrant protein behaviour has also been observed whereby aggregation is involved in many disease pathologies [Greenwald and Riek, 2010]. Proteins can also be intrinsically disordered lacking stable tertiary structure [Dyson and Wright, 2005].

Despite the complexity and variety of protein folds, two general observations hold true for globular proteins. First, the locations of residues in the protein structure are generally distributed according to their polarity, with non-polar residues (valine, leucine, isoleucine, methionine, and phenylalanine) typically found in the interior of the protein fold sequestered from solvent, and polar residues (arginine, histidine, lysine, aspartic acid, and glutamic acid) typically found on the surface of the protein fold in contact with solvent [Voet and Voet, 2004]. While there are numerous exceptions, this partitioning of polar and non-polar residues from solvent is the primary driving force responsible for globular protein folding [Chandler, 2005]. Specifically, globular proteins adopt compact

tertiary structures that reduce the solvent accessible surface area of non-polar residues thereby decreasing the entropy of the protein-solvent system. Secondly, residues sequestered from solvent that occupy the protein core are typically arranged in a compact fashion with side chain atoms occupying conformational minima [Dill, 1999; Dill *et al.*, 1995].

1.2. The Complexity of Structure and Sequence Space

Having described the arrangement of the small 56 residue globular protein G β 1, it is evident that the task of modeling sequence and structural perturbations to even small wild type (WT) polypeptide sequences is not trivial due to the astronomical complexity of the problem. For example, if side chain degrees of freedom (*df*) were ignored and the conformations of backbone dihedral angles ϕ and ψ (with each occupying 6 local minima) were exclusively considered, the total number of potential atomic arrangements would total to 3.95×10^{85} conformations. Although many of these conformations will not correspond to globular protein folds, the task of directly evaluating the total structure space of the WT sequence alone is prohibitive. If the additional torsional flexibility of all protein sidechains configurations were also considered, the combined total *df* would be astronomically vast. The general equation:

$$df = \prod_{x=1}^n rot[x] \cdot \prod_{x=1}^{n-1} \psi[x] \cdot \prod_{x=2}^n \phi[x] \quad (\text{EQ 1.1})$$

allows for determination of the *df* for a sequence space searched across any given protein fold of *n* residues length, dependent on the number of side chain configurations

($rot[x]$, count of configurations across the set of amino acids) to be evaluated at each residue position (x), and the number of backbone dihedral configurations ($\phi[x]$ and $\psi[x]$) considered at each residue position.

The relationship between a protein's primary structure and its higher order secondary, tertiary, and quaternary structures is referred to as the protein folding problem. More precisely, the protein folding problem can be expressed in the form of a question to ascertain how a protein's sequence dictates its fold and function. Unfortunately, an exact solution to the protein folding problem remains elusive as the df associated with an exhaustive evaluation of the combinations of all sequences and folds for a given polypeptide chain is computationally prohibitive. To limit the size of the problem, an inverse approach to the solution of the protein folding problem can be applied [Pabo, 1983]. Instead of attempting to solve the fold of a given protein sequence, the inverse protein folding problem solves for the sequences tolerated on a given protein fold. By restricting the search and evaluation of sequences on a single backbone configuration, the df can be limited to the number of side chain configurations considered:

$$df = \prod_{x=1}^n rot[x] \quad (EQ 1.2)$$

resulting in a substantial reduction to the size of the problem. Thus, with the atomic coordinates for a given protein fold the set of sequences about the original WT sequence can be evaluated [Figure 1.4]. This *in silico* calculation is referred to as CPD.

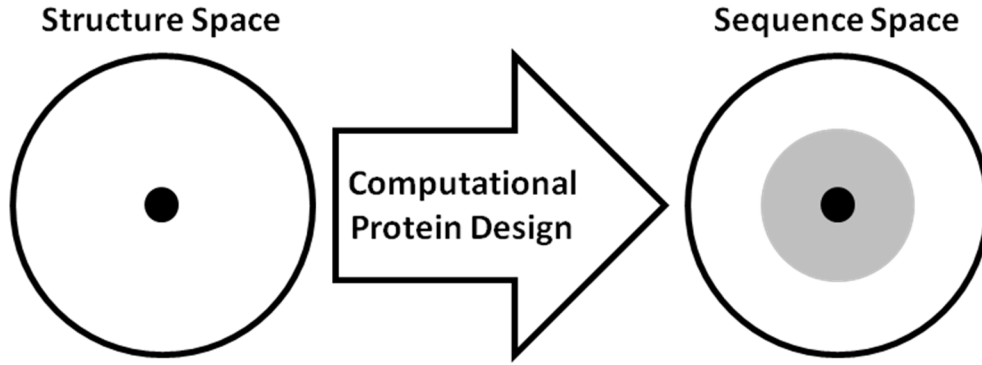


Figure 1.4. The inverse protein folding problem. The structure space specified by the atomic coordinates of the WT sequence (black) is used as a template for the evaluation of sequence perturbations (grey). The structure and sequence spaces are not shown to scale.

The objective of a CPD calculation is to compute the sequence that can best stabilize the input template from the set of user specified sequence perturbations. Traditionally, a search of sequence space is conducted using a single backbone template in a calculation framework referred to as SSD. To conduct a SSD calculation, the coordinates of the backbone template must be retrieved and prepared, a library of side chain configurations (rotamers) to be considered in the design calculation must be assembled, a method for evaluating sequence stability *in silico* (potential energy function) must be created, and an optimization function for searching the sequence space, are required. The goal of a SSD calculation is to solve the minimum energy rotamer configuration ($\min E_{seq}$) for the most stable sequence (*seq*) tolerated on the protein fold:

$$\min E_{seq} = \sum_{i=1}^n E_i[rot[i]] + \sum_{i=1}^{n-1} \sum_{j=i+1}^n E_{ij}[rot[i], rot[j]] \quad (\text{EQ 1.3})$$

where, energy is computed for the one-body (E_i) and two-body (E_{ij}) interactions between each rotamer (i) with the template and other rotamers (j), respectively, for the set of all rotamers (n) belonging to the sequence (seq).

The complexity of this CPD problem involving the evaluation of sequence configuration combinations is referred to as non-deterministic polynomial time hard (NP-hard) [Pierce and Winfree, 2002]. Specifically, CPD is an NP-hard problem because the objective ($\min E_{seq}$) of definitely solving the lowest energy rotamer configuration across the set of all configurations requires that all rotamer configurations be solved. Nonetheless, SSD can be conducted efficiently and effectively using stochastic algorithms to search and evaluate sequence space. Regardless of the method used to solve the design problem, the size and complexity of the problem scales with the number of positions and rotamers included in the calculation [Voigt *et al.*, 2000]. The purpose and implementation of backbone template preparation, side chain modelling with rotamer libraries, stability evaluation with potential energy models, and the search of sequence space are presented in the following four sections.

1.3. Template Preparation Methodology

Before initiating a SSD calculation, the coordinates for the template structure must be retrieved and prepared. There are a variety of sources and experimental atomic-resolution data. For example, the PDB contains x-ray crystallographic [Gallagher *et al.*, 1994] and solution nuclear magnetic resonance (NMR) spectroscopy [Gronenborn *et al.*, 1991] structures for the WT sequence G β 1 fold. It is also possible to produce a

template using homology modeling. Importantly, structures solved with particular methodologies (NMR) have proven less suitable at producing predictive input templates for CPD when compared with predictions made using crystallographic structures [Schneider *et al.*, 2009]. Regardless, the complete set of template coordinates required to model the design structure space in calculation must be considered. To prepare coordinates retrieved from the PDB, a means of interpreting the molecular system, for example the Molecular Operating Environment (MOE) program [MOE, 2011], must be available.

After retrieving the protein's atomic coordinates, and ensuring that all required atoms are present, hydrogens may need to be added to the structure. Several algorithms have been developed to tackle this computation including molProbity [Chen *et al.*, 2010] and Protonate3D [Labute, 2009]. For the structures prepared in this thesis, the Protonate3D algorithm as implemented in MOE [MOE, 2011] was employed. Following the addition of hydrogens, the template geometry is typically energy minimized, in internal redundant coordinate space, resulting in a prepared structure suitable for CPD.

1.4. Rotamer Libraries

After preparing the template structure, the design calculation must be defined and the mutant structure space of side chain configurations modelled on the template structure. A side chain configuration is referred to as a rotamer, while a collection of side chain configurations comprises a rotamer library. The design space encompasses

three types of objects present in the molecular system during calculation. Protein backbone and side chains that are not perturbed and remain fixed throughout the CPD calculation are referred to as template atoms. Residue positions that are allowed to explore their conformational state, but not change their chemical identity as indicated in the backbone template, must be modelled using the rotamer library. These rotamers are said to be floated because their conformation may change but the residue in the sequence remains the same. Lastly, residues that change in both their identity (mutation) and configuration are referred to as design residues. Rotamers that are placed onto the backbone template in the design structure space are said to be threaded onto the template. It is important to note, that not all protein side chains present on the template structure must be floated or designed and that the sequence space to be searched can be further reduced.

Side chain configurations can be built or collected from a variety of sources. The penultimate rotamer library developed by Lovell *et al.* [Lovell et al., 2003] for example contains 152 side chain conformations for all twenty of the canonical amino acids. The rotamer library was collected from crystallographic structures depositions in the PDB regardless of their progenitor residue's secondary structure. As a result, the rotamer library is referred to as a *backbone independent* library. The work in this thesis employs the Dunbrack *backbone dependent* rotamer library, which expands on the rotamer library generation procedure (approximately 350 total rotamers) creating separate collections of rotamers grouped according to their respective ϕ and ψ backbone dihedral angles [Dunbrack and Cohen, 1997]. Use of a backbone dependent rotamer

library exploits the correlated nature of sidechain and backbone configurations thereby improving the accuracy of sidechain prediction while reducing the required conformational space to be sampled [Dunbrack and Karplus, 1993]. However, for either rotamer library, bond lengths and angles are held fixed at ideal values and rotamer configurations vary only by their χ dihedral angles. The configurations of the first dihedral angle, χ_1 , are similar for all canonical amino acids as the single bond between sp^3 hybridized C_α and C_β atoms occupy configurations about three minima at $\pm 60^\circ$ and 180° , with slight preference for the *gauche minus* (-60°) conformation.

1.5. Energy Functions

Following the preparation of the design model, involving the threading of rotamers onto the backbone template, the stabilities of the designs (the set of all rotamer combinations threaded onto the template) must be evaluated. In CPD, stability is typically approximated with the use of a potential energy model or a heuristic scoring function [Boas and Harbury, 2007] and reported as a score, with lower energy or score value indicating more stable interactions than those having higher energies or scores. The CPD calculations conducted in this thesis utilize a simple physics based potential energy model [Mayo *et al.*, 1990] to compute the interaction energy between rotamers and their template (one-body energies), and between pairs of rotamers at different positions (two-body energies). The energy for a given combination of rotamers threaded onto a template is computed from the sum of all one-body and two-body energies (EQ 1.3).

The energy computed for one- and two-body interactions are the sum of three non-covalent molecular mechanics potential energy terms, describing: van der Waals (E_{vdW}), electrostatic (E_{ele}), and hydrogen-bonding (E_{hbd}) interactions. The van der Waals interaction energy computed between each atom, k and l , belonging to the set of atoms of two rotamers, $N[rot[i]]$ and $N[rot[j]]$, at a distance of r_{kl} has the form:

$$E_{vdW} = \sum_{k=1}^{N[rot[i]]} \sum_{l=1}^{N[rot[j]]} D_{kl} \left[\left(\frac{\alpha \cdot R_{kl}}{r_{kl}} \right)^{12} - 2 \left(\frac{\alpha \cdot R_{kl}}{r_{kl}} \right)^6 \right] \quad (EQ 1.4)$$

corresponding to a 12–6 Lennard-Jones potential energy curve [Lennard-Jones, 1924]. The minimum on the potential energy curve is found at a distance (R_{kl}) calculated as the geometric mean of the van der Waals radii for atoms k and l (EQ 1.5) while and the depth of the well (D_{kl}) is computed as the geometric mean of the well depth parameters of both atoms k and l (EQ 1.6) [Mayo *et al.*, 1990]. The geometric mean is computed as follows for the two parameters:

$$R_{kl} = \sqrt{2(R_k \cdot R_l)} \quad (EQ 1.5)$$

$$D_{kl} = \sqrt{D_k \cdot D_l} \quad (EQ 1.6)$$

The van der Waals interaction energy contains a third parameter (α) that allows for the attenuation of the geometric atomic radii term (R_{kl}). Scaling the α parameter, typically set at a value of 0.9, readily enables modulation of the repulsive and attractive shape of the 12–6 curve.

The second interaction energy term calculated accounts for the electrostatic energy (E_{ele}) between two atomic partial or point charges (q_k and q_l) [Gasteiger and

Marsili, 1980]. The electrostatic term is typically computed with a distance dependent dielectric (ϵ) between 8 and 40, to approximate the internal environment of a protein [Boas and Harbury, 2007], having the form shown in equation 1.7.

$$E_{\text{ele}} = \sum_{k=1}^{N[\text{rot}[i]]} \sum_{l=1}^{N[\text{rot}[j]]} \left(\frac{q_k q_l}{\epsilon r_{kl}} \right) \quad (\text{EQ 1.7})$$

The last interaction energy term computes the hydrogen-bonding interaction energy (E_{hbd}). This term is calculated using three atoms, with one atom (k) belonging to the hydrogen bond acceptor ($\text{rot}[i]$), and two atoms (l and m) belonging to the hydrogen-bond donor ($\text{rot}[j]$). The use of three atoms in the computation allows a more thorough geometric treatment of the hydrogen bonding interaction by considering the angle between the donor and acceptor atoms (θ). The angle function ($F[\theta]$) employed is specific to the hybridization of the donor and acceptor atoms [Dahiyat and Mayo, 1997], but the hydrogen-bonding term has the general form:

$$E_{\text{hbd}} = D_0 \left[5 \left(\frac{R_0}{r_{km}} \right)^{12} - 6 \left(\frac{R_0}{r_{km}} \right)^{10} \right] \cdot F[\theta] \quad (\text{EQ 1.8})$$

with well depth (D_0) and hydrogen-bond equilibrium distance (R_0) parameters set typically to $8.0 \text{ kcal}\cdot\text{mol}^{-1}$ and $2.8 \text{ \AA}$, respectively.

Additional energy terms can be appended to the potential energy function, including reference energies intended to reflect the stability of the unfolded state using amino acid identity specific penalties [Ali *et al.*, 2005; Busch *et al.*, 2008] and solvent penalties calculated using the solvent exposed surface area of rotamers [Lazaridis and Karplus, 1999; Street and Mayo, 1998]. Notably, the evaluation of rotamer interactions

described here excludes the consideration of backbone configuration energetics. Thus, this theoretical treatment asserts an assumption that all backbone configurations are energetically degenerate. Furthermore, the evaluation of the restriction of state is not computed, thus the entropy of sequence perturbations to the system are unconsidered.

1.6. Sequence Optimization Algorithms

Having evaluated the energies of rotamers against the template (one-body) and rotamers at alternate design positions (two-body) the SSD calculation can be concluded with the optimization of the rotamer space. Several algorithms have been developed, including those that are deterministic, like dead end elimination algorithms that provably solve the global energy minimum configuration (GMEC) [Desmet *et al.*, 1992; Gordon *et al.*, 2003]. However, the size and complexity of many sequence optimization problems can prove overwhelming for brute force optimization using deterministic algorithms, leading to issues of poor convergence and efficiency [Traore *et al.*, 2016]. The need for rapid and reliable solution of large design sequence spaces has lead to the development of several stochastic algorithms, including the use of Monte Carlo (MC) [Metropolis *et al.*, 1953] followed with simulated annealing (SA) [Kirkpatrick *et al.*, 1983] and the development of genetic algorithms [Jones, 1994].

Sequence optimization calculations presented in this thesis were calculated using an improved version of the fast and accurate side chain topology and energy refinement (FASTER) algorithm [Desmet *et al.*, 2002] referred to as Way-FASTER [Allen and Mayo, 2006]. The original FASTER algorithm consist of a series of protocols, each acting on the

converged set of rotamer configurations from the previous step. To begin, FASTER initiates a search of the sequence space by deterministically identifying rotamers with the lowest one-body energy at each design position. The protocol and the input rotamer configurations solved by the protocol are both referred to as the backbone minimum energy configuration (BMEC). Thus, the BMEC protocol ignores any two-body interactions in its search of the rotamer space. Next, a second deterministic protocol, referred to iterative branch relaxation (iBR), is initiated from the BMEC set of rotamers. The iBR protocol involves the solution of the lowest energy (sum of the one- and two-body terms) rotamer configuration for a single design position while the remaining design positions are held fixed. Thus, application of the iBR protocol solves the lowest energy rotamer with respect to all other positions in the design. The iBR protocol is repeated, simultaneously updating all positions until convergence is achieved (no lower total energy solution can be found).

Following convergence of the iBR protocol, the next sequence search protocols to be completed by FASTER are stochastic. First, conditional iBR (ciBR) is initiated from the converged iBR configurations and implements the same search protocol, with the exception that each new configuration has a probability ($P = 0.8$) of being kept. ciBR is repeated until convergence is achieved (a maximum number of non-productive cycles has been reached). The next two protocols, single and double perturbation relaxation (sPR and dPR, respectively), are the most time consuming search protocols in the FASTER algorithm. The perturbation relaxation protocol sets configurations of rotamers at one (sPR) or two (dPR) design positions and attempts to solve the lowest energy

configuration of rotamers at the remaining design positions. In this way, rotamer relaxations of the perturbation relaxation protocol can be thought of as the inverse procedure to the iterative batch protocol. Repeating the set of stochastic protocols in the FASTER algorithm yields a set of trajectories whose convergence to the same sequence is referred to as the FASTER minimum energy configuration (FMEC).

Two major improvements have been made to the FASTER algorithm in an implementation known as Way-FASTER [Allen and Mayo, 2006]. The first improvement involved replacement of the BMEC protocol, demonstrated to produce rotamer configurations with high two-body energies (because they are ignored by the BMEC protocol) with a series of MC-SA trajectories. The resulting rotamer configurations have lower total energies and improve the diversity of configurations when initiating the iBR protocol. The second improvement involved restricting the relaxation of rotamers by sPR and dPR to only those that interact strongly with the perturbed rotamer(s). Application of these protocol changes lead to substantial improvements in algorithm reliability and efficiency at solving FMEC when tested in SSD against four test proteins, including G β 1.

1.7. Protein Engineering Examples Aided by Single-state Design

The application of SSD using methods similar to those described above, have lead to a variety of successful protein engineering efforts. By stabilizing sequences in the context of a protein fold, SSD has been used to improve protein stability [Dantas *et al.*, 2003; Malakauskas and Mayo, 1998; Nauli *et al.*, 2001; Shah *et al.*, 2007], design protein

conformational switching [Ambroggio and Kuhlman, 2006], and create the first *de novo* protein fold (Top7) [Kuhlman *et al.*, 2003]. By stabilizing interactions found in quaternary protein structures, specificities have also been engineered using SSD for small molecules [Allert *et al.*, 2004; Lilien *et al.*, 2005], peptides [Shifman and Mayo, 2003], and proteins [Huang *et al.*, 2007b; Kapp *et al.*, 2012; Steed *et al.*, 2003]. The approach has also been applied to the design of esterase [Bolon and Mayo, 2001], Kemp eliminase [Privett *et al.*, 2012; Röthlisberger *et al.*, 2008], retro-aldolase [Jiang *et al.*, 2008], and Diels-Alderase [Siegel *et al.*, 2010] activities into protein scaffolds, by evaluating sequence stabilities on templates that include catalytically relevant model structures. In the next chapter, I will describe multistate design (MSD), an emerging methodology in CPD that expands upon the traditional SSD framework by including multiple protein backbone configurations and states into the CPD framework, and introduce the hypothesis and rationale of this thesis.

CHAPTER 2

Multistate Computational Protein Design Theory

Traditional CPD calculations search and rank sequences using a single fixed protein backbone template in an approach referred to as SSD, as discussed in the previous chapter. While SSD has enjoyed considerable success, certain design objectives require the explicit consideration of multiple conformational and/or chemical states. Cases where a multistate approach may be advantageous over the SSD approach include the design of conformational changes into proteins, using native ensembles to mimic backbone flexibility, and the design of ligand binding or oligomeric association specificity. These design objectives can be efficiently tackled using MSD, an emerging methodology in CPD that considers any number of protein conformational or chemical states as inputs instead of a single protein backbone template, as in SSD. In this chapter, recent examples of the successful design of a desired property into proteins using MSD are described. These studies employing MSD are divided into two categories – those that utilize multiple conformational states, and those that utilized multiple chemical states. These examples serve as the basis from which the multistate methodologies proposed in this thesis were developed.

Some of the language and figures in this chapter were adapted from a review article and a book chapter both coauthored with Dr. Roberto A. Chica.

Davey JA and Chica RA (2012) Multistate approaches in computational protein design. *Protein Science*, 21(9):1241-1252 [Davey and Chica, 2012].

Davey JA and Chica RA (2015) Multistate computational protein design with backbone ensembles. *"Methods in Molecular Biology: Computational Protein Design"*, Ilan Samish, *book in production*.

2.1. Single-state Design Challenges

In the previous chapter, I introduced the traditional single-state CPD framework involving the optimization of amino-acid sequences on coordinates from a single, fixed protein backbone template. The SSD calculation framework searches sequence space with the objective of identifying sequences satisfying a single specific design objective that is explicitly represented by the template [Figure 2.1] in an approach referred to positive design [Lassila, 2010]. The result of a positive SSD calculation is a list of sequences rank-ordered according to their predicted ability to stabilize the template and thus likely to exhibit the intended property or function. Thus, by allowing for the efficient search of sequence space *in silico* across an astronomically vast scale that is experimentally impossible to achieve [Chica *et al.*, 2005], CPD serves as a useful tool to aid protein engineers who strive to create new proteins to solve specific chemical problems.

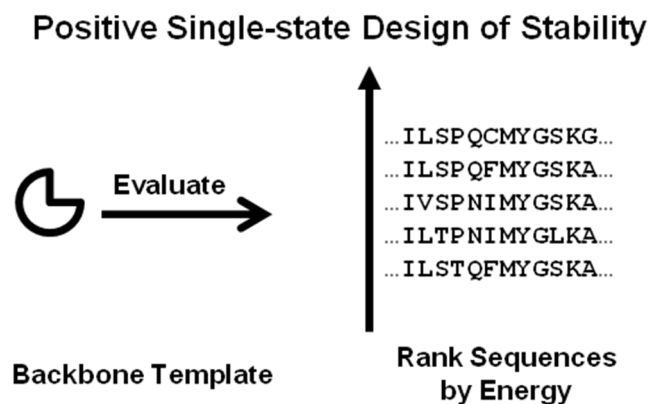


Figure 2.1. An example of single-state CPD. Sequence stability is evaluated on a target fold represented by a single-fixed backbone template and ranked according to their energy.

In practice however, a number of challenges arise as a result of the approximations made when modeling proteins as single-state entities. These challenges can be subdivided into two categories, specifically, the evaluation and identification of the minimum energy atomic configuration for a given protein sequence, and the consideration and evaluation of protein sequences in the context of multiple competing states. The first challenge results from the fact that SSD is susceptible to false negative predictions that result from the combined use of a fixed protein backbone template and a set of rigid rotamers in order to model mutant protein structures. This *fixed backbone approximation* leads to the incorrect rejection of desirable sequences that would be otherwise accepted if the backbone geometry was allowed to relax or if a slightly different rotamer configuration was allowed [Choi *et al.*, 2009].

To address the fixed backbone approximation, several strategies have been developed including the use of softer repulsive potential energy terms [Dahiyat and Mayo, 1997; Grigoryan *et al.*, 2007; Kellogg *et al.*, 2011], flexible backbone algorithms [Murphy *et al.*, 2012; Ollikainen *et al.*, 2013; Smith and Kortemme, 2011], *post-hoc* application of energy minimization [Gainza *et al.*, 2016], iterative energy minimization [Borgo and Havranek, 2012; Wang *et al.*, 2005], and continuous rotamer optimization [Gainza *et al.*, 2012]. The second challenge, involves the evaluation of sequences in the context of multiple competing states. For example, in SSD undesired states, such as the unfolded and aggregated states are either ignored or implicitly designed against through the use of penalties applied to the potential energy function during the evaluation of sequences.

While CPD originated initially as a means to understand the forces that govern protein stability [Malakauskas and Mayo, 1998; Street and Mayo, 1999], the method has since expanded to facilitate the design of numerous protein properties. For instance, the application of SSD to improve protein stability has had considerable success [Dantas *et al.*, 2003], leading to several hyperthermophilic variants of both the *Drosophila melanogaster* engrailed homeodomain [Shah *et al.*, 2007] and the well-studied β 1 domain of *Streptococcal* protein G [Malakauskas and Mayo, 1998; Nauli *et al.*, 2001]. The method has been extended to the design of protein conformational switching [Ambroggio and Kuhlman, 2006] as well as the creation of the first novel protein fold (Top7) with atomic-level accuracy [Kuhlman *et al.*, 2003].

SSD has also been applied in protein design efforts involving the introduction of new or altered binding specificity for metals [Calhoun *et al.*, 2003; Marvin and Hellinga, 2001; Summa *et al.*, 2002], small molecules [Allert *et al.*, 2004; Lilien *et al.*, 2005], peptides [Shifman and Mayo, 2003], proteins [Huang *et al.*, 2007b; Kapp *et al.*, 2012; Steed *et al.*, 2003], and DNA [Ashworth *et al.*, 2006]. The approach has also been used in the design and introduction of new catalytic activity into proteins. For example, esterase [Bolon and Mayo, 2001], Kemp eliminase [Privett *et al.*, 2012; Röthlisberger *et al.*, 2008], retro-aldolase [Jiang *et al.*, 2008], and Diels-Alderase [Siegel *et al.*, 2010] activities have all been successfully designed *de novo* using SSD.

2.2. Multistate Approaches to Computation Protein Design

For the examples provided above, SSD has had success for two major reasons: the first is that the coordinates for the protein backbone template were suitable for the desired function and redesigned sequence, and the second, is that the simulation goals could be accomplished with the use of a positive design approach. However, in cases where multiple desired conformational or chemical states for a protein of interest, or cases where desired and undesired states (i.e. negative design) must be considered to design the desired property, SSD is expected to have reduced success. Cases where a multistate approach may be advantageous over the SSD approach include designing conformational changes into proteins, using native ensemble to mimic backbone flexibility, and designing ligand specificity or oligomeric association.

Figure 2.2 illustrates specific design objectives suitable for solution by MSD, an emerging methodology in CPD that considers any number of protein conformational or chemical states as inputs instead of a single protein backbone template, as with SSD. MSD allows sequence selection to be driven by the energetic contributions of multiple conformational or chemical states simultaneously. In truth, all protein design problems are MSD problems because proteins are dynamic by nature since they are capable of adopting alternate local minima conformations [Bouvignies *et al.*, 2011; Henzler-Wildman and Kern, 2007] and these conformations can be further influenced and potentially stabilized by the binding of allosteric modifiers, small molecules, and other proteins [Lee and Craik, 2009]. Additionally, proteins that bind ligands can exist in various chemical states such as the unbound and bound states.

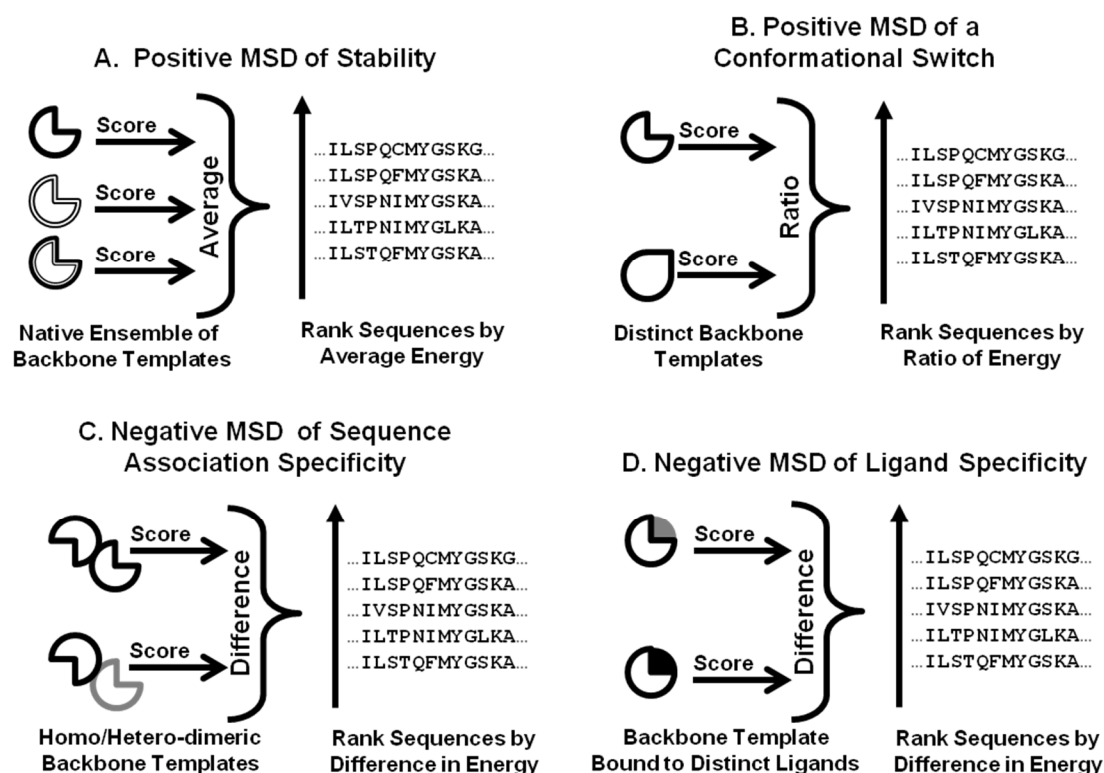


Figure 2.2. Examples of multistate CPD. Examples of design calculations include: positive MSD of sequence stability (A), positive MSD of a conformational switch (B), negative MSD of sequence association specificity (C), and negative MSD of ligand binding specificity (D).

In the examples illustrated in Figure 2.2, sequence energy must be recombined and evaluated in the context of the particular design objective using a fitness function. For example, MSD of sequence stability can be implemented through the evaluation of sequence stability across an ensemble of backbone templates representing the target fold [Figure 2.2.A]. In this instance, sequence stability is reported as a Boltzmann weighted average. In the example of the design of a conformational switch [Figure 2.2.B], sequences are evaluated on two distinct backbone templates, each representing a specific protein conformational state, and reported as the average energy between the two states for each sequence. Oligomeric association of sequences is designed by

evaluating sequence pairs on both homo- and heterodimeric templates and taking the difference in energy between the desired and undesired association states [Figure 2.2.C]. The design of ligand binding specificity favoring one small molecule over another involves the evaluation of sequences across templates incorporating each ligand and computing the difference in energy between the desired and undesired states [Figure 2.2.D]. The first two examples presented (panels A and B) utilize a positive design approach; the last two examples presented (panels C and D) utilize a negative design approach.

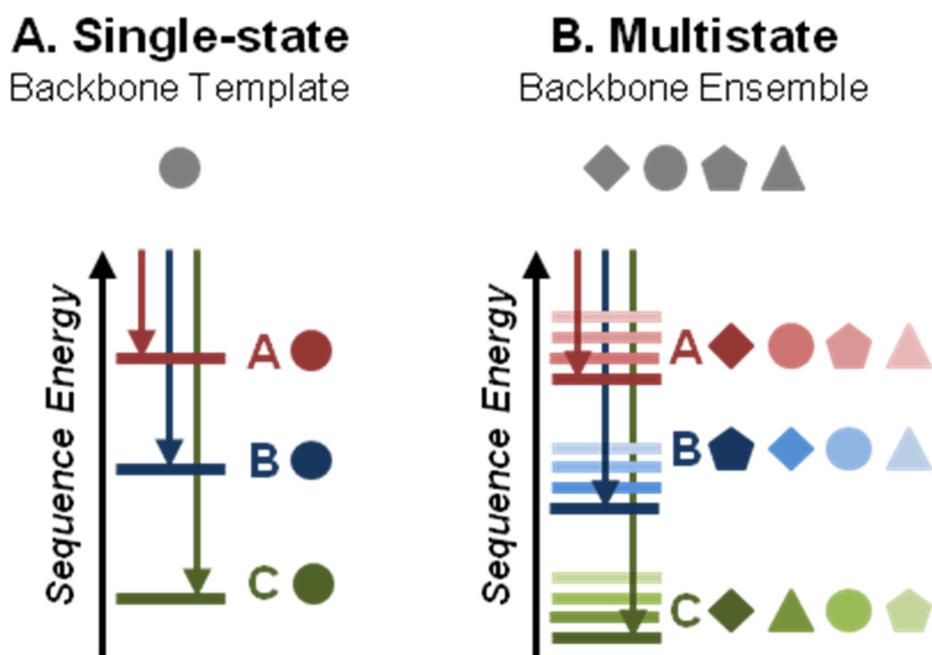


Figure 2.3. Single-state and multistate design. In SSD (A), a single backbone template (circle) is used to score and rank sequences (A: red, B: blue, and C: green) according to their predicted stability (arrow) with respect to a fixed reference energy (i.e. 0 energy units). In MSD (B), an ensemble of four backbone structures (diamond, circle, pentagon, and triangle) is used to score and rank sequences. Predicted stability (arrow) is computed as a specific function of ensemble member energy.

Thus, MSD differs from SSD by its use of an energy combination function to compute *sequence fitness* and a modified search algorithm to find optimal sequences in the context of multiple backbone templates. Importantly, the fixed backbone approximation that results from implementation of the SSD framework can be effectively alleviated by MSD [Figure 2.3]. This is principally the result of finding more stable template-rotamer combinations with which to evaluate a given set of protein sequences.

2.3. Positive and Negative Multistate Design

As described above, MSD accounts for alternate conformational and/or chemical states for all specified sequences during the CPD simulation. Alternate conformational states are included in CPD simulations by searching the same sequences on multiple different backbone templates, such as distinct folds or native ensembles. This approach can be contrasted with the MSD of alternate chemical states which involves the simulation of the same sequences in search of alternate functions, for example, comparing sequence specificity for multiple ligands or for oligomeric association. MSD simulations utilizing multiple conformational and chemical states lend themselves toward two specific CPD strategies referred to as positive and negative design [Figure 2.4].

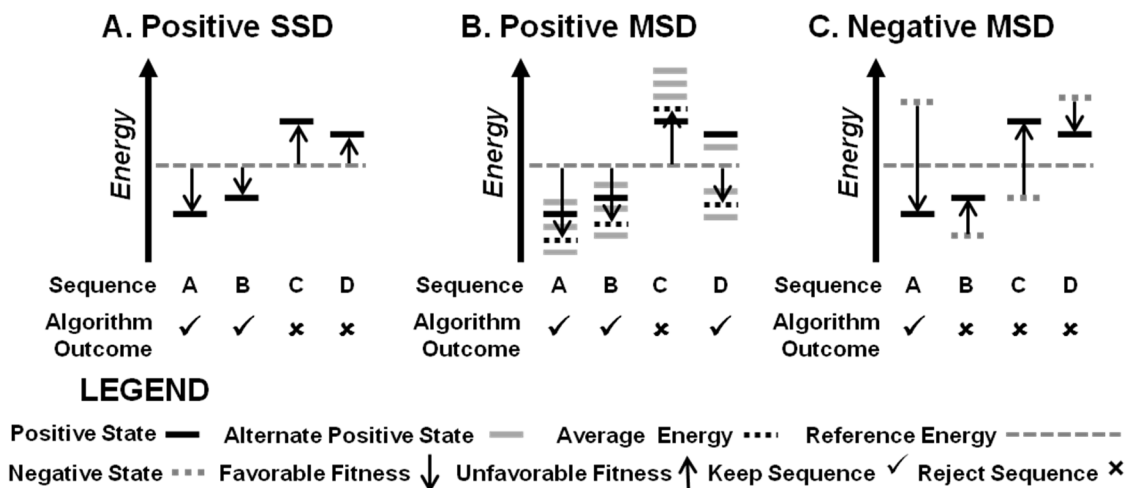


Figure 2.4. Sequence evaluation in positive and negative design. Sequence ranking and selection by positive SSD (A), positive MSD (B), and negative MSD (C). Reference energy can be arbitrary assigned, based on the evaluated stability of a known stable sequence, or using reference energies.

As shown above in figure 2.4, the outcome of sequence ranking and selection on the predicted stability of four hypothetical sequences (A, B, C, and D) is dependent on both the calculation approach and strategy employed during CPD. In this example, calculation by positive SSD [Figure 2.4.A] keeps and ranks sequence A as the best sequence because it has the largest favorable difference in energy between the single positive state and the reference energy. Sequence B is also kept although its difference in energy from the reference value ranks it lower than sequence A, while sequences C and D are both rejected as their energy on the positive state with respect to the reference is positive and thus unfavorable. The ranking of sequences for this hypothetical positive SSD calculation example is: $A > B$. The next hypothetical calculation, performed using positive MSD [Figure 2.4.B], evaluates and ranks sequences using a Boltzmann weighted average of sequence energies over an ensemble of

backbone templates, which includes the original template from SSD. In this example, sequences A and B are both kept as their ensemble energy is less than that of the reference energy. Sequence C is rejected as its ensemble Boltzmann weighted average energy lies above the reference energy and sequence D is kept as its ensemble Boltzmann weighted average energy is less than the reference energy, even though some of the states belonging to the ensemble evaluate the sequence with an energy greater than that of the reference. The ranking of sequences kept for this example is: A > B > D. The last example details sequence selection and ranking by negative MSD [Figure 2.4.C], which involves computing the difference in energy for each sequence evaluated in the context of a desired positive state, and undesired or competing negative state. In this manner, sequences favoring the positive state are ranked and evaluated favorably, while those favoring the negative state are ranked and scored poorly. For this example, sequence A is kept, as it is the only sequence whose difference in energy between the positive and negative state is favorable and its positive state energy falls below that of the reference energy. The remaining sequences (B, C, and D) are rejected.

2.4. The Practical Implementation of Multistate Design

As discussed previously, while SSD is well suited to positive design simulations for a single desired state, MSD is better suited to the application of simulations which require the explicit consideration of multiple adoptable states during sequence optimization. CPD simulations traditionally require the following: (1) a protein backbone

template, (2) a rotamer library containing a discrete set of conformations for all amino acid side chains to be tested, (3) a scoring function to rank rotamers in order of stability or desired function, and (4) an optimization algorithm to search through the combinations of rotamers to return the sequences with the best predicted energies. Backbone templates are often prepared from protein crystal structures; however solution NMR or molecular dynamics (MD) structures have also been used [Boas and Harbury, 2007; Schneider *et al.*, 2009].

Generally, hydrogens are explicitly modeled in the design simulation and solvent is implicitly accounted for using a distance-dependent dielectric, a simplified surface area model [Boas and Harbury, 2007], or an occlusion based solvent model [Chica *et al.*, 2010; Lazaridis and Karplus, 1999]. The preparation of template backbones can be completed with a coarse energy minimization to alleviate van der Waals clashes introduced as a result of the hydrogen addition process or those already present in the deposited crystal structure. After the backbone template has been prepared, the CPD calculation can begin by threading sets of discrete rotamers onto the template at specified positions. Following sequence optimization, the output returned by CPD algorithms is a list of ranked sequences based on their evaluated energy.

From a computational standpoint, the major difference between MSD and SSD simulations is amino-acid sequence optimization. A multistate CPD simulation can be viewed as a collection of multiple independent single-state calculations whereby rotamers for a specific amino-acid sequence are optimized in the context of each of the conformational and/or chemical states used as input templates. This means that an

amino-acid sequence will not adopt the same side chain conformations in all states. In each of these single-state calculations, rotamer combinations are scored using typical scoring functions for each amino-acid sequence. This fitness value is a single value ranking for that amino-acid sequence across all states which reflects how well the sequence stabilizes the positive state(s), and in the context of negative design, how it also destabilizes the negative state(s). MSD optimization algorithms then attempt to optimize this fitness value as a function of amino-acid sequence. Many of the common optimization algorithms, such as MC with SA [Ambroggio and Kuhlman, 2006], genetic algorithms [Pokala and Handel, 2005], and FASTER [Allen and Mayo, 2010], as well as deterministic algorithms such as dead-end-elimination [Yanover *et al.*, 2007], have been adapted for use in MSD calculations.

2.5. Multistate Analysis and Multistate Design

An alternate approach to MSD that can be used to evaluate fitness of amino-acid sequences across multiple backbone templates is *multistate analysis* (MSA). MSA involves the combination of scores obtained from parallel SSD simulations into a single fitness value for each sequence that is computed post-CPD. The resulting fitness values are then used to re-rank sequences [Figure 2.5]. By employing alternate backbones as input templates to these parallel SSD calculations, MSA can be used to identify the most favorable template to score each sequence [Howell *et al.*, 2014] or to evaluate how well each sequence stabilizes an ensemble of backbone templates. MSA differs from MSD by its sequence optimization procedure, which is not informed by the energetic

contributions of multiple backbones. Instead, sequence optimization in MSA is performed as in SSD and only the use of an energy combination function to compute sequence fitness distinguishes it from SSD. Because of this, MSA has the benefit of being less computationally demanding than MSD but has the drawback of potentially constraining explored sequence space since sequence optimization is not guided by multiple states.

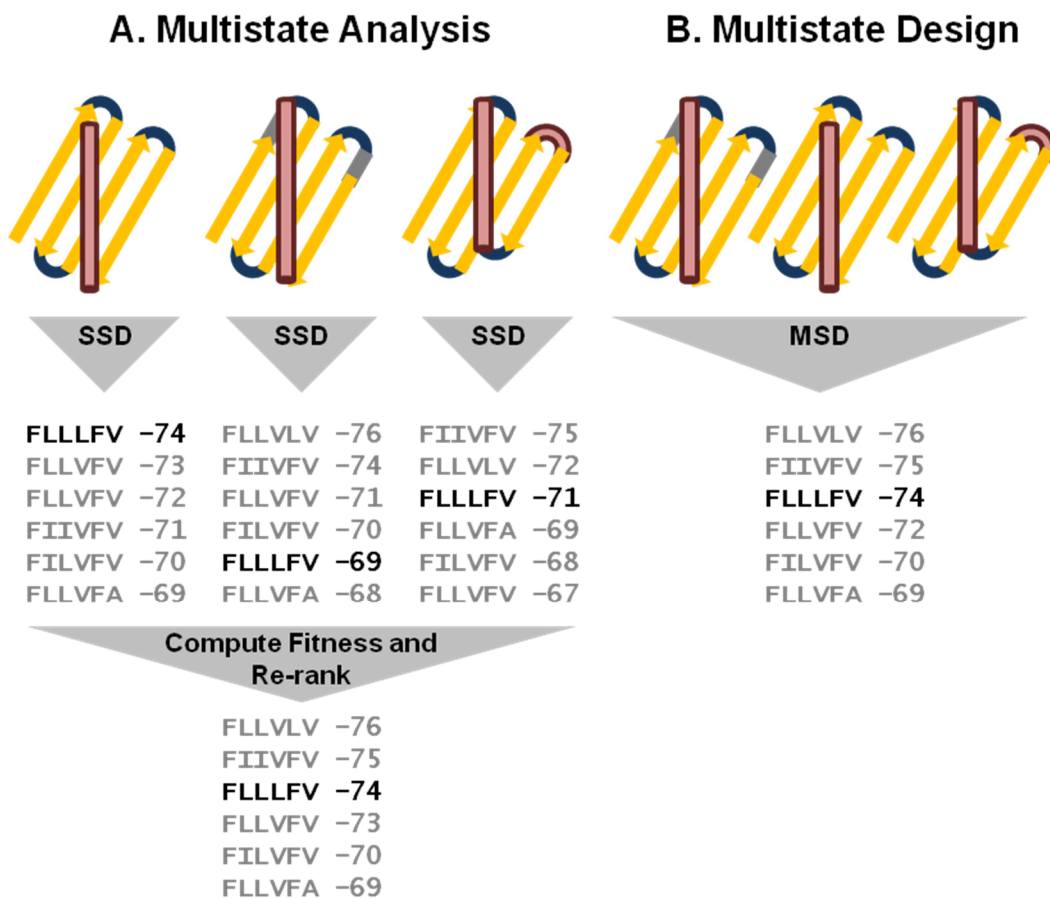


Figure 2.5. Frameworks for multistate analysis and multistate design. In MSA (A), multiple independent SSD calculations are performed in parallel using alternate backbones as input templates. The sequence energies obtained from each SSD calculation are recombined post-CPD into a fitness value for each sequence across the ensemble of three structures. Sequences are then re-ranked based on their fitness

values, generating a new ranked list of evaluated sequences. In MSD (B), sequence optimization is guided by the energetic contributions of multiple protein structures simultaneously. Thus, sequence optimization and evaluation are performed concertedly, resulting in a list of sequences that are ranked according to their predicted stability across the ensemble of three structures.

2.6. Protein Engineering Examples with Multiple Conformational States

MSD can be used to design protein sequences that undergo large conformational changes depending on experimental conditions, in effect leading to protein switches whose conformation can be controlled via a desired stimulus. For example, in 2006, Ambroggio and Kuhlman engineered a conformational switch, referred to as Sw2, capable of reversibly adopting two distinct folds depending on the presence of transition metals [Ambroggio and Kuhlman, 2006]. The first state, resembling a 2Cys-2His zinc finger fold, was stabilized in the presence of Zn(II) while the second state, involving assembly of the peptide into a trimeric coiled-coil fold, was favored in the absence of Zn(II) [Figure 2.7.A]. The engineering of Sw2 employed a MSD approach in which sequences were searched on two distinct protein backbones. These templates consisted of residues 13-44 of hemagglutinin from *H. influenzae* and residues 3-33 of the Zif268 zinc finger-DNA complex from *M. musculus*, corresponding to the desired coiled-coil and zinc finger conformations respectively. Rotamers were scored using the standard Rosetta energy function to which an additional scoring term based on the probability of finding specific amino acids at each position in multiple sequence alignments was added, and these scores were in turn used to compute the fitness of each sequence. The fitness of each sequence was computed as the score average for

each sequence on both conformational templates. By selecting for sequences that stabilized both conformations, Ambroggio and Kuhlman performed a positive MSD simulation. The design objective for this particular example requires a MSD simulation because sequence optimization for a single conformation in a SSD approach does not guarantee the stabilization of the alternate, unconsidered, conformation.

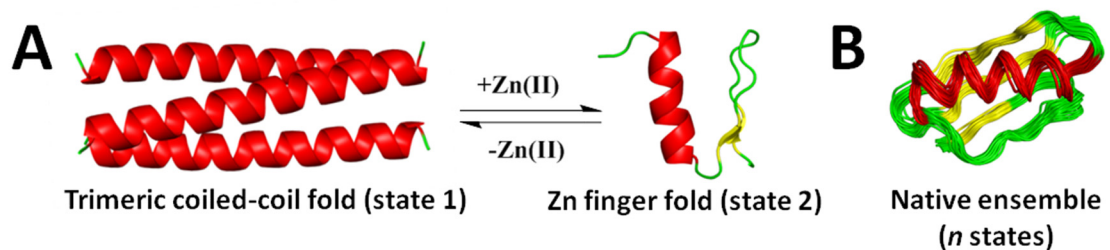


Figure 2.6. Input templates for positive multistate design examples. Two distinct protein folds (trimeric coiled-coil and a zinc finger) were used as inputs to design a conformational switch (A) [Ambroggio and Kuhlman, 2006]. Native ensembles of G β 1 consisting of multiple backbone templates were used as inputs for the design of sequences with improved stability (B) [Allen *et al.*, 2010].

To demonstrate the existence of the conformational switch, circular dichroism (CD) spectroscopy was employed to monitor the disappearance of the coiled-coil conformation in the presence of increasing concentrations of Zn(II). In corroboration with analytical ultracentrifugation data and CD spectroscopy experiments, the authors were able to support that the trimeric coiled-coil conformation of the designed Sw2 did exist in the absence of Zn(II). To conclusively demonstrate that the zinc finger conformation was achieved, cobalt absorption spectroscopy was used to show that the engineered Sw2 sequence did indeed bind Co(II) in a monomeric zinc finger fashion. Cobalt-bound Sw2 showed peaks at 310, 340, and 640 nm, consistent with what is

observed for other Cys2-His2 zinc fingers. However, analytical ultracentrifugation experiments showed that Sw2 began to associate in a nonspecific manner at higher concentrations ($> 100 \mu\text{M}$). This is an important observation as it demonstrates the consequence of not explicitly modelling undesired states in a negative design approach, in this case an aggregated state, during the CPD simulation. Another important consideration about this MSD example is that the sequences ability to undergo the conformational change itself was not explicitly designed for. Instead, the conformational change was expected to occur because the designed sequences were predicted to be stable in both conformational states. Despite the successful design of a conformational switch, the authors note that the current limitation of this approach lies in the ability to accurately compute the relative free energy values between the two conformations.

Another example of MSD applied to multiple conformational states was published by Allen *et al.* [Allen *et al.*, 2010] who used native backbone ensembles approximating the flexibility of proteins to design sequences that could improve the stability of a single fold belonging to the protein domain G β 1 [Figure 2.7.B]. The overall goal was to test whether the use of native backbone ensembles in a MSD approach could predict higher quality sequences than the use of a single fixed backbone in a SSD approach. To test this, the authors used a variety of structural ensembles, including solution NMR structures as well as templates derived from constrained molecular dynamics (cMD) and unconstrained molecular dynamics (uMD) simulations. These backbone ensembles were compared alongside a typical SSD simulation using a single crystallographic structure of WT G β 1. During the positive MSD simulation, each sequence was searched across all

backbone templates and the energy of each sequence returned. To compute the fitness of each sequence across the ensemble, a Boltzmann weighted average function was used. Employing this averaging function ensured that sequences which were favored in the majority of states were not penalized by the few states which disfavored them. Stability of the sequence variants was then experimentally determined using a 96-well plate chemical denaturation assay.

It was found that significant improvement over the control SSD simulation was afforded only by the cMD backbone ensemble which gave sequences of similar or better stability to the WT and returned no destabilized unfolded variants. The NMR solution structure ensemble gave 6 sequences with stabilities similar or identical to that of the WT, and 18 which were highly destabilized. The cMD ensemble performed the poorest as all 24 designed sequences failed to produce any significant fluorescence signal change during chemical denaturation, indicating that the chemical environment of the Trp fluorescence reporter is very different from that of the target structure. This observation, coupled with the low levels of protein expression for members of this library, suggests that the members of the uMD library likely assume structures that differ from the target structure. It is important to note that the backbone RMSD for the NMR ensemble, as well as the unconstrained and the cMD ensembles were 0.25, 0.84 and 0.12 Å, respectively. [Allen *et al.*, 2010] The uMD ensemble had the greatest RMSD from crystallographic structure and gave the poorest predictions of the three ensembles suggesting that larger deviations from the crystal structure may be responsible for predicting sequences that are incompatible with the target structure. There still remains

a difference in the predictive capability between the NMR and the cMD ensembles which may be due, in part, to their respective deviation from the input crystallographic structure. These predictive differences may also arise because not all backbones included in NMR ensembles may be suitable to use as templates for CPD calculations [Schneider *et al.*, 2009].

In each case, regardless of MSD ensemble or SSD template, the experimental stability and the simulation score for each sequence were not correlative. This suggests, once again, that there may be many more factors influencing the outcome and success of a MSD simulation than just the ensemble's RMSD from the input crystallographic structure. Nonetheless, this particular example demonstrates the improved predictive power that MSD can afford over SSD for the engineering of protein stability.

Backbone ensembles have also been used for the computational design of a pair of interacting proteins that is orthogonal to the WT proteins from which they were derived and that is capable of mediating complex biological processes within cells. In this study [Kapp *et al.*, 2012], two interacting proteins, the GTPase cell division control protein 42 homolog (Cdc42^{WT}) and its partner, the GTPase Exchange Factor named intersectin (ITSN^{WT}), were designed such that the new sequence variants, referred to as *ortho*Cdc42 and *ortho*ITSN respectively, would preferentially associate together over their WT parents. Following a computational alanine scan [Kortemme and Baker, 2002], position 56 in Cdc42 was identified as the main candidate for affecting ITSN binding without perturbing other binding interactions that are required for the biological activity of Cdc42. Position 56 of Cdc42^{WT} and four positions on ITSN^{WT} (1369, 1373, 1376, and

1380) were then mutated *in silico* to introduce a specific interaction between the *ortho*Cdc42 and its mutated binding partner, *ortho*ITSN. Initially, the authors performed these *in silico* mutations using a positive SSD approach, however, Kapp *et al.* found that the use of a single fixed backbone template could not correctly predict mutations in ITSN that were specific to the identity of the mutated amino acid at position 56 of Cdc42.

Failure to predict precisely the sidechain-sidechain interactions in a protein binding interface likely resulted from the use of a single fixed backbone, presumably by biasing the choice of rotamer during the simulation, a known artifact in CPD. To address this issue, a backbone ensemble generated using backrub motions [Davis *et al.*, 2006; Friedland *et al.*, 2008] was used in design calculations where the identity but not the conformation of the mutated residue 56 was fixed while the four neighboring positions in ITSN were designed. This approach allowed for the identification of glutamic acid at position 1373 in ITSN when an arginine residue is found at position 56 in mutated Cdc42, demonstrating that MSD yielded a specific interaction in the orthogonal binding site.

To validate the designed *ortho*Cdc42 and *ortho*ITSN pair, *in vitro* experiments were employed to examine the catalytic activity and thermal stability of the individual proteins, as well as the dissociation constant of the Cdc42^{WT}/ITSN^{WT}, Cdc42^{WT}/*ortho*ITSN, *ortho*Cdc42/ITSN^{WT}, and *ortho*Cdc42/*ortho*ITSN pairs. The first experiment monitored the catalysis of nucleotide exchange by ITSN using a fluorescently-labelled GDP analog. By measuring the loss of fluorescence as a function

of time, Kapp *et al.* found that the mutant *orthoITSN* catalyzed nucleotide exchange in *orthoCdc42*, but not in $Cdc42^{WT}$, while $ITSN^{WT}$ did not catalyze exchange in *orthoCdc42*, demonstrating that the designed pair maintained its activity. Thermal stability of each protein was monitored using CD spectroscopy. Thermal melts of the two WT and mutant proteins indicated a similar stability. The association specificity of *orthoCdc42/orthoITSN* was evaluated using surface plasmon resonance (SPR) spectroscopy to determine the dissociation constant (K_D) for the four possible pairs. SPR results indicated that the orthogonal pair, *orthoCdc42/orthoITSN*, associated specifically and preferentially ($K_D = 478 \pm 22$ nM) while the non-cognate pairs, $Cdc42^{WT}/orthoITSN$ and *orthoCdc42/ITSN^{WT}*, were not observed to associate. While specific, the designed orthogonal complex has approximately 16-fold weaker binding affinity than the WT $Cdc42^{WT}/ITSN^{WT}$ complex ($K_D = 29 \pm 2$ nM). Additional validation was done in an *in vitro* reconstituted system as well as *in vivo* in mammalian cells to demonstrate that the signal pathway's function was unimpeded when the WT $Cdc42^{WT}/ITSN^{WT}$ pair was replaced with the designed *orthoCdc42/orthoITSN* pair.

The examples provided for MSD using multiple conformational states illustrate different ways of exploiting structural data for different design purposes. In one case, the multiple conformations were used to create a protein capable of undergoing a desired conformational change when exposed to a stimulus. In the other, the use of native backbone ensembles during a CPD simulation leads to a substantial improvement in the quality of output variants by simulation. In this latter example, improvements likely result from decreasing the number of false negatives, which occur because many

rotamers, which could be compatible with slight changes in the backbone, would be considered sterically incompatible and be discarded because the protein backbone is not allowed to relax after rotamer placement [Desjarlais and Handel, 1999]. Although researchers have focused on remodeling the backbone during the CPD simulation [Saunders and Baker, 2005], in effect creating a flexible backbone that would be capable of tolerating rotamers that cause slight steric clashes, MSD with native ensembles that mimic backbone flexibility can be used as an alternative.

For example, native backbone ensembles derived from MD simulations, backrub motions, and kinematic closure refinement protocols have been used as inputs in MSD to recapitulate antibody–antigen interface amino acid residues that were experimentally observed by phage display [Babor *et al.*, 2011]. In the future, MSD with native ensembles could be combined with the design of desired properties such as ligand binding and oligomeric association to improve the quality of designs.

2.7. Protein Engineering Examples with Multiple Chemical States

Multiple chemical states, such as protein/ligand complexes or oligomeric association of subunits, can be used in MSD to explicitly design for specificity using a negative design approach. One example of the application of MSD to modify ligand binding specificity is a study by Frey *et al.* who examined the effect of mutations to the antibiotic target dihydrofolate reductase of the methicillin resistant *S. aureus* on both inhibitor binding and catalytic activity [Frey *et al.*, 2010]. In this article, Frey and coworkers sought to investigate whether or not mutations that conferred antibiotic

resistance while maintaining catalytic activity could be predicted in an attempt to assist the drug design process. In this manner, MSD was applied to screen for sequences which preferentially bound the natural substrate dihydrofolate (chemical state 1) over a propargyl-linked antifolate inhibitor (chemical state 2) using a single backbone template [Figure 2.8.A]. Even though a single protein backbone was used, this is an example of MSD as two templates with different ligands were considered during sequence optimization. The fitness of each sequence, describing the binding preference, was calculated by taking the ratio of the score describing the binding affinity for the substrate over the score of the binding affinity for the inhibitor. Thus, the authors sought to identify sequences that scored highly for dihydrofolate and poorly for the inhibitor. Their MSD simulation returned 105 mutants with a score ratio of infinity. Four double mutant sequences, which exhibited a significantly better dihydrofolate score than the other sequences, were experimentally tested. Three of the experimentally tested double mutants (V31Y/F92I, V31Y/F92S, V31F/F92L) conferred antibiotic resistance (resulting in 18, 8.7, and 13 fold increases to the original K_i of 10 nM, respectively) while maintaining sufficient catalytic activity to maintain cell viability (36, 107, and 306 fold decreases to the original k_{cat}/K_M of $2.14 \mu\text{M}^{-1}\text{s}^{-1}$, respectively). It is important to note that the MSD design procedure required the X-ray crystallographic structures of dihydrofolate reductase bound separately to both the substrate and inhibitor. Having both templates eliminated the need for translation and rotation of either substrate or inhibitor in the active site of the protein for each sequence solution to be scored.

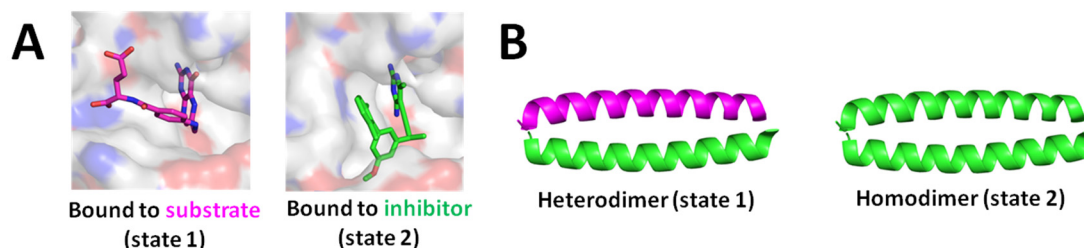


Figure 2.7. Input templates for negative multistate design examples. Enzyme structures bound to either a substrate or an inhibitor were used as inputs to design for antibiotic resistance (A) [Frey *et al.*, 2010]. Coiled coil templates were used as inputs to design oligomeric association specificity (B) [Havranek and Harbury, 2003].

MSD can also be used for the design of oligomeric association of protein subunits. As many biological processes are mediated and controlled by protein-protein interactions, the ability to design for protein-protein binding specificity is of paramount importance. A recent review by Chen and Keating [Chen and Keating, 2012] details an integrated approach to the design of protein-protein interaction specificity using computational design and experimental library screening methodologies. Here, we focus on publications employing a MSD approach to tackle the design of oligomeric association specificity and the challenge associated with the simultaneous design of specificity and stability.

In a pioneering example of MSD, Havranek and Harbury developed and experimentally validated this approach for the redesign of GCN4 dimeric coiled-coil association specificity [Havranek and Harbury, 2003]. By considering both the homo and heterodimer form of each coiled-coil sequence pair in the MSD simulation [Figure 2.8.B], Havranek and Harbury could direct formation to prefer one specific oligomeric association (homodimer) in positive design fashion while searching the same sequences

against formation of the other oligomeric association (heterodimer) in a negative design fashion and vice versa. Two additional states, the unfolded and aggregated states, were also explicitly designed against. The score for each sequence in each state was the computed free energy. The fitness of each sequence was evaluated by the difference in free energy for the target state from the ensemble of competing states. Havranek and Harbury used the *S. cerevisiae* GCN4 homodimeric coiled-coil structure as their design scaffold for both the target dimeric positive state(s) and the competing dimeric negative state(s). The unfolded state was modelled using AGADIR parameters [Lacroix *et al.*, 1998] and the aggregated state was determined by re-evaluating the stability of the target dimer with an adjusted dielectric constant to reflect the environment of an aggregated protein. Experimental validation of the association specificity and stability predicted from MSD simulation was carried out using a disulfide-exchange assay and a urea denaturation assay, respectively. Havranek and Harbury were able to produce 8 new individual sequence variants which preferentially associated as homodimers and 4 new sequence pairs which preferentially associated as heterodimers. However, most of the designed sequences were destabilized when compared to the WT. Furthermore, Havranek and Harbury's results demonstrated that omission of any structures from the ensemble of negative states (i.e. the competing homo/hetero-dimer, unfolded or aggregated states) was detrimental to the performance and outcome of their simulation. For example, omission of the unfolded and aggregated states during the MSD simulation gave 2 sequence pairs which were predicted to associate specifically but

also to be unstable, while omission of either the competing homo/hetero-dimer state resulted in a total loss of predicted association specificity.

Another example of MSD using chemical states was the redesign of the *H. influenzae* SspB adaptor protein. In this study, Bolon *et al.* designed the WT sequence, which associates as a homodimer, into mutant variants that preferentially associate as heterodimers [Bolon *et al.*, 2005]. The authors employed and compared both a SSD and MSD approach to their CPD objective. The first approach, SSD, involved stabilization of the heterodimer in a positive design fashion without explicit consideration of the homodimer, while the second approach, MSD, explicitly included the competing homodimer state. Two pairs of sequence variants at positions 12, 15, 16 and 101, found at the dimer interface, were produced. The first pair, containing Phe12/Ala15/Phe16/Ile101 (FAFI) in one subunit and Leu12/Ala15/Leu16/Ile101 (LALI) in the other subunit, was found after SSD while the second pair, containing Leu12/Ser15/Leu16/Ala101 (LSLA) in one subunit and Tyr12/Gly15/Phe16/Met101 (YGFM) in the other subunit, was found using the MSD approach. For both the SSD and MSD approaches, the score for each dimer sequence pair was computed using the standard Dreiding force field terms [Mayo *et al.*, 1990]. In the case of SSD, the lowest energy sequence was selected while in the case of MSD, the fitness for oligomeric association specificity was computed by taking the difference in energy between the target heterodimer state and the competing homodimer states ($\text{fitness} = 2 \times E_{AB} - E_{AA} - E_{BB}$, where E is the energy and A and B are different monomeric subunits). To experimentally validate their designs, Bolon *et al.* employed ion-exchange

chromatography to isolate homodimer and heterodimer species and a urea denaturation assay to examine stability. These experiments allowed the authors to calculate the free energy of dissociation/unfolding at 30 °C which demonstrated that their SSD approach could not produce sequences (sequence pair 1 variants (A) FAFI and (B) LALI) which favored either the homo or heterodimer state ($\Delta G_{AA} = 24.3$, $\Delta G_{BB} = 25.6$, $\Delta G_{AB} = 25.6$ [kcal·mol⁻¹]) while the MSD approach allowed for the design of sequences (sequence pair 2 variants (A) LSLA and (B) YGFM) that favorably formed heterodimers over homodimers ($\Delta G_{AA} = 14.5$, $\Delta G_{BB} = 17.5$, $\Delta G_{AB} = 20.1$ [kcal·mol⁻¹]). While the MSD approach allowed for the successful design of specificity, the stability of the heterodimer relative to the WT sequence was reduced (WT: LAYV, $\Delta G_{WT-WT} = 23.6$ kcal·mol⁻¹ vs 20.1 kcal·mol⁻¹ for the LSLA/YGFM heterodimer).

The computational redesign of oligomeric association using MSD has also been successfully accomplished by Ali *et al.* who redesigned a previously engineered homotetramer, comprised of four $\beta\beta\alpha$ motif peptides referred to as beta-beta-alpha tetramer 2 (BBAT2), to prefer association in a heterotetrameric fashion [Ali *et al.*, 2005]. In this MSD example, the positive state consisted of the heterotetrameric assembly (ABAB, where A and B are different monomeric subunits) of two previously designed mutants of BBAT2, which was generated using symmetry operations on their crystal structures (PDB codes 1SNA and 1SNE). In addition, four negative states consisting of the two homotetramers (AAAA and BBBB, where A and B are different monomeric subunits) and the unfolded state for each peptide ($A_{Unfolded}$ and $B_{Unfolded}$) were included. The unfolded state energy was evaluated by the sum of the energy of all amino acids,

belonging to the designed sequence, in the context of a Gly-Gly-Xaa-Gly-Gly pentapeptide model. All energies were computed using a modified CHARMM19 force field. [MacKerell *et al.*, 1998] Stability of the tetramer was computed as the difference between the unfolded state energy and the heterotetramer energy ($E_{\text{Unfolded}} - E_{\text{ABAB}}$) while the specificity of the complex was calculated as the energy difference between two heterotetramers and the respective homotetramers ($E_{\text{AAAA}} + E_{\text{BBBB}} - 2 \times E_{\text{ABAB}}$). Sequences were ranked using both stability and specificity fitnesses. Those that possessed high fitness values for both stability and specificity were further minimized and rescored. It was found after MSD that monomers having Glu and/or Asp mutations at positions 11, 13, and 18, and monomers having Arg and/or Lys at opposing sites on adjacent subunits (positions 11 and 13 from one monomer interact with positions 18 and 13 on the other monomer, respectively) conferred the best fitness values for directing specificity of the complex to the heterotetramer. Two heterotetramers were thus designed — BBAhetT1 and BBAhetT2, composed of 2 subunits each of individual peptides A-Ala and B-Phe, or A-Abu and B-Phe, respectively. Each of the designed monomers (A-Ala, A-Abu, and B-Phe) displayed very weak CD signal between 200 and 300 nm when tested individually, indicating that they have very little secondary structure. However, equimolar mixtures of A-Ala/B-Phe (BBAhetT1) and A-Abu/B-Phe (BBAhetT2) gave rise to a significant increase in ellipticity indicative of interhelical association. The heterotetrameric association specificity of BBAhetT1 and BBAhetT2 was also demonstrated by fluorescence quenching experiments (the A-Ala and A-Abu monomers were synthesized with a quencher while the B-Phe monomer was

synthesized with a fluorophore) and analytical ultracentrifugation. Finally, the crystal structure of BBAhetT1 was solved and confirmed the designed C₂-symmetric heterotetramer assembly. Thermal denaturation experiments of the two designed heterotetramers demonstrated that they are less stable than the parent BBAT2 homotetramer. These results again demonstrate that the sequences identified by MSD, although displaying different oligomeric association specificity, also display decreased stability, similar to what Bolon and coworkers [Bolon *et al.*, 2005] observed in their design of the SspB adaptor protein.

This trade-off between increased specificity and decreased stability arises from the fact that sequence optimization for the design of oligomeric association specificity in the previous examples involved a negative design approach which did not explicitly attempt to increase the stability of the positive state. To address this issue, Grigoryan *et al.* developed a landmark computational framework, referred to as cluster expansion and linear programming-based analysis of specificity and stability (CLASSY) [Grigoryan *et al.*, 2009]. The CLASSY framework is initiated by designing for stability of the positive state, i.e. the design–target heterodimer, without consideration of specificity. This first positive design calculation yields a sequence with maximum affinity for the target but does not necessarily confer specificity over competing states, referred to as off-targets. This is followed by a negative design calculation whereby sequences are optimized by computing the difference in energy between the positive state (design–target) and the best ranking competing state (design–off-target). By sequentially introducing a larger specificity constraint (Δ) during sequence optimization ($\Delta = E_{\text{design–target}} - E_{\text{design–off-target}}$),

specificity can be gradually increased. This procedure, referred to as a specificity sweep, allows for the optimization of sequences maximizing the specificity of the desired interaction while minimizing the resulting decreased stability of the target state.

To validate their CLASSY framework, Grigoryan *et al.* designed and tested 48 peptides to bind representative members across the 20 families of the human basic-region leucine zipper (bZIP) transcription factors, known to associate as homo- and/or heterodimeric parallel coiled-coils. Experimental screening of their designs was completed using a micro array assay where the target bZIPs were immobilized onto the surface of the array and the plate was washed with designed peptides bearing a fluorescent dye [Newman and Keating, 2003]. The authors showed that of the 48 designed peptides, 40 bound to their intended target bZIP. The 7 designed peptides that showed the highest specificity were further characterized by thermal denaturation monitored with CD spectroscopy. Each of the 7 designed peptides were denatured in the presence of either their intended target bZIP, the next-best interaction partner reported by the array experiment, a protein closely related by sequence identity to the target bZIP, or the design peptide itself, giving a total of 28 peptide mixtures. Thermal melts from 18 of the 21 mixtures containing undesired design–off-target complexes showed no increase in temperature of denaturation compared to that of the average of the individual components, indicating that the designed specificity was achieved. Thus, Grigoryan *et al.* showed that the problem involving the sacrifice of stability to achieve enhanced specificity could be in part circumvented by their CLASSY framework.

In the design examples described above, MSD was used to design for ligand binding or oligomeric association specificity. Since the design of specificity requires the evaluation of alternate competing states, it is not surprising that MSD in a negative design approach was required to achieve the design goal. Indeed, the optimization of sequences for one state using SSD does not guarantee that the same sequence will be detrimental to the other state(s). In the future, this approach might enable the application of positive MSD using multiple chemical states for the design of desired small molecule binding specificity. For example, enzyme/substrate complexes could be used as inputs for MSD to explicitly design for broad specificity or multi-substrate enzymes.

2.8. Multistate Computational Protein Design Challenges

The examples described above illustrate how MSD can improve the quality of designs, either by identifying sequences that are compatible with desired states and incompatible with undesired ones or by keeping sequences that would have been discarded in the context of a single fixed backbone template. Although MSD can provide an improved and successful avenue over SSD for CPD, there remains a challenge to overcome in order to improve its usefulness for the design of any desired protein property. This challenge is the accurate modelling of energetic effects arising from destabilizing mutations in competing states during negative design.

In negative design calculations, the selection of relevant sequences for the negative state can be difficult. This difficulty arises from the fact that the score values

for the negative state may not be meaningful. Consider the example shown in figure 2.9, where the same hypothetical five amino acid sequence (Tyr-Ser-Trp-Ala-Ala) was scored in the context of two negative state backbones. Using negative state backbone A, a severe steric clash between the sidechains of Tyr and Trp leads to an overinflated energy for the negative state. As a result, this sequence would be preferred during a negative MSD calculation as its fitness, i.e. the difference in energy between the positive and negative states, would be very high. However, although steric clashes are likely destabilizing in the context of a real protein, backbone motions can alleviate them through conformational rearrangements. Thus, negative states that contain multiple steric clashes should be preferred in negative MSD since they are likely to destabilize proteins more efficiently.

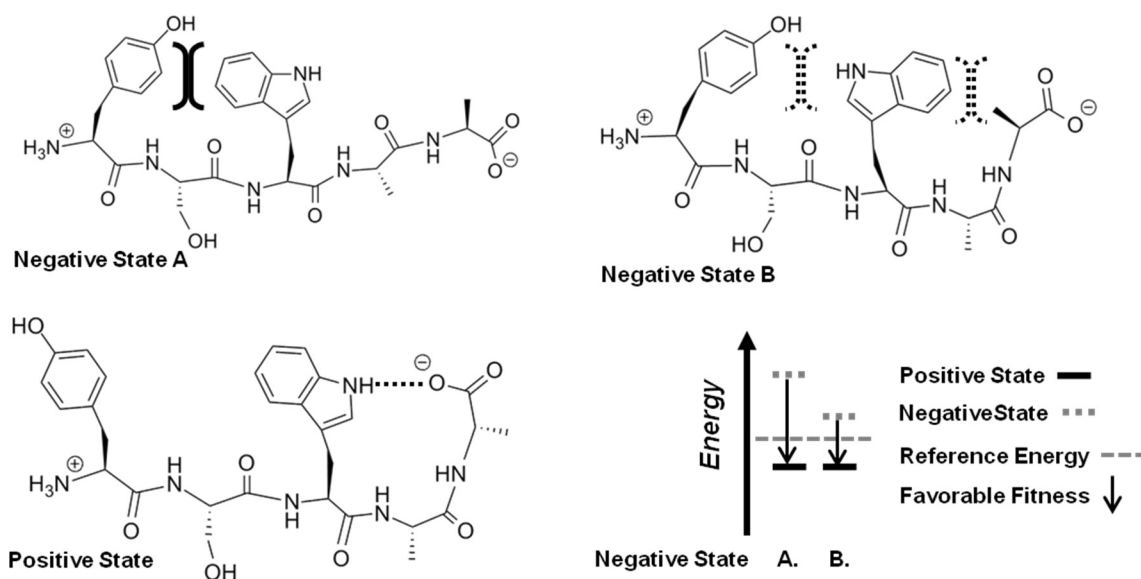


Figure 2.8. Negative state selection and its impact on sequence evaluation. A sequence of a hypothetical pentapeptide (Tyr-Ser-Trp-Ala-Ala) is modeled and scored in the context of a single desired positive state and one of two competing negative states. Negative state **A** has a major unfavorable interaction (thick curved lines) between the

side chains of Tyr and Trp, while negative state **B** has two minor unfavorable interactions (thin curved lines) between the side chain of Trp and the side chains of Tyr and Ala. Meanwhile, the positive state has the best energy due to a favorable electrostatic interaction between the side chain of Trp and the terminal carboxylate group (dotted line). The positive state also lacks the unfavorable steric interactions present in the negative states. Although negative state **A** is physically less relevant than negative state **B**, it results in a more favorable ranking of the sequence due to a large difference in energy with the positive state.

Bolon *et al.* recognized this observation and came up with a workaround to select sequences exhibiting many steric clashes in the negative state. To do this, Bolon *et al.* restricted the pairwise interaction energy between two rotamers exhibiting unfavorable steric clashes at a maximum value to approximate conformational relaxation, thereby giving preference to sequences having multiple smaller steric clashes, considering them to be more acceptable than sequences having a few major steric clashes [Bolon *et al.*, 2005]. For the hypothetical example in figure 2.9, negative state backbone B contains two smaller steric clashes between the sidechain of Trp and the sidechains of Tyr and Ala. Although the fitness of the sequence is lower because the difference in energy between the negative state and the positive state is lower, the sequence can now be selected even in the presence of sequences that contain major steric clashes. Although this strategy helps to address the issue of conformational relaxation in the negative state, it is still unknown whether the modelling of the negative state in this fashion is physically relevant. Another approach to address the issue of relevant scoring of the negative state is the use of more rigorous statistical mechanics methods to score the negative states [Boas and Harbury, 2008]. For this approach to be successful, accurate atomic models of the negative states are required. However, methods to generate

accurate atomistic models for aggregated and unfolded states have not yet been developed and validated, limiting the use of this approach in negative design.

2.9. Thesis Hypothesis and Rationale

The set of MSD experiments reviewed in this chapter serve as examples of how to apply and implement the multistate framework in the design of complex protein properties and functions. Importantly, the methodology and the success of these calculations rely on the generation of suitable ensemble states capable of representing these protein properties and behaviours. The ideal MSD calculation would incorporate all possible chemical and conformational states allowing for the consideration of any protein property or function in calculation. However, because the landscape of a protein is conformationally vast, such an ensemble cannot be considered.

Instead, alternative ensemble representations of chemical and conformational states must be developed. The challenge here lies in the fact that many of the relevant protein states cannot be directly observed. For example, a general purpose explicit full atom model of the unfolded state that could be applied in the negative design of a protein's stability does not exist. Hence, the goal of this body of work is to improve upon the understanding of the relationship between the ensembles used in MSD and their utility to predict protein properties. Particular emphasis will be placed on the use of and ability to recapitulate experimental data to guide modelling of the design structure space.

Having introduced the theoretical frameworks of SSD and MSD, the next five chapters include the development of two new ensemble and template generation protocols, and their application to three protein design projects involving the prediction of new methyltransferase substrates, the prediction of protein sequence stability for WT and mutant G β 1 sequences, and the design of conformational exchange into the involving the lone tryptophan side chain in the G β 1 fold. For each project, discussions on the experimental validation of the MSD calculations are included.

CHAPTER 3

Ensemble Generation Methodologies

MSD with backbone ensembles approximating conformational flexibility can predict higher quality sequences than SSD with a single fixed backbone. However, it is unclear what characteristics of backbone ensembles are required for the accurate prediction of protein sequence stability. In this chapter, we demonstrate improvements to the accuracy of protein stability predictions made with MSD by using a variety of backbone ensembles to recapitulate the experimentally measured stability of 85 *Streptococcal* protein G domain β 1 sequences. Ensembles tested here include an NMR ensemble as well as those generated by MD simulations, by Backrub motions, and by PertMin, a new method that we developed involving the perturbation of atomic coordinates followed by energy minimization. MSD with the PertMin ensembles resulted in the most accurate predictions by providing the highest number of stable sequences in the top 25, and by correctly binning sequences by their calculated stability relative to the WT sequence, as stable or unstable with the highest success rate ($\approx 90\%$) and the lowest number of false positives. The performance of PertMin ensembles is due to the fact that their members closely resemble the input crystal structure and have low potential energy. Conversely, the NMR ensemble as well as those generated by MD simulations at 500 or 1000 K reduced prediction accuracy due to their low structural similarity to the crystal structure. The ensembles tested herein thus represent on- or

off-target models of the native protein fold and could be used in future studies employing both positive and negative design strategies.

Some of the language and figures in this chapter were adapted from a research article and a book chapter both coauthored with Dr. Roberto A. Chica.

Davey JA and Chica RA (2014) Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins: Structure, Function, and Bioinformatics*, 82(5):771-784 [Davey and Chica, 2014].

Davey JA and Chica RA (2015) Multistate computational protein design with backbone ensembles. *"Methods in Molecular Biology: Computational Protein Design"*, Ilan Samish, book in production.

3.1. Issues with Single-state Design

CPD is a useful tool for protein engineers. It facilitates the identification of proteins displaying a desired property by enabling the *in silico* evaluation of amino acid sequences on a scale that is experimentally impossible to achieve [Chica *et al.*, 2005]. As discussed previously (Chapter 1), CPD simulations are typically performed using a SSD approach, whereby discrete amino acid side chain conformations, called rotamers, are threaded onto a single fixed protein backbone template. Interaction energies between each rotamer and the template as well as between each pair of rotamers are then calculated using an empirical energy function [Boas and Harbury, 2007]. Finally, rotamer combinations are optimized using a search algorithm to explore both rotamer and sequence space to give a list of sequences ranked by their *in silico* score, which reflects their predicted stability.

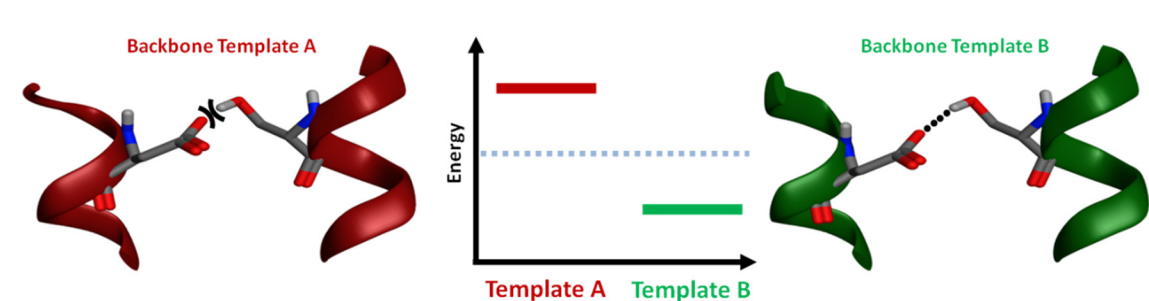


Figure 3.1. The single-state design artefact. Two backbone templates (A: red, B: green) consisting of two helical structures varying in geometry and relative orientation, are used to evaluate a pair of rotamers having identical configuration at the same position on either template. When evaluated on the template A, the rotamer pair are scored unfavorably, penalized as the result of an unfavorable steric interaction (indicated with bold curves). The same rotamer pair is scored favorably, the result of a hydrogen bonding interaction (indicated with a bold dotted line) that is present when evaluated on template B.

While SSD has successfully been used to design a variety of protein properties and functions [Allert *et al.*, 2004; Ashworth *et al.*, 2006; Bolon and Mayo, 2001; Kuhlman *et al.*, 2003; Röthlisberger *et al.*, 2008; Shah *et al.*, 2007; Shifman and Mayo, 2003], the methodology is prone to false negative predictions, resulting in the omission of desirable sequences. This known artefact of SSD [Figure 3.1] can be attributed to the combined use of a fixed protein backbone template with rigid rotamers, which can result in the rejection of favourable amino acid sequences that would have been accepted on a slightly different backbone geometry or with a slightly different rotamer configuration [Choi *et al.*, 2009]. To address the rigid rotamer and fixed backbone approximations, CPD can be performed using continuous rotamer minimization [Gainza *et al.*, 2012] or flexible backbone design, respectively [Murphy *et al.*, 2012; Ollikainen *et al.*, 2013; Smith and Kortemme, 2011]. Alternatively, the fixed backbone approximation can be addressed by optimizing sequences on an ensemble of fixed protein backbone templates in an approach referred to as MSD [Davey and Chica, 2012].

MSD simulations consist of multiple independent single-state calculations in which rotamers for a specific amino acid sequence are optimized on each of the ensemble members used as input templates. However, unlike in SSD, most or all of the single-state scores are combined into a fitness value representing the predicted stability of the sequence across the ensemble. For the work presented in this chapter, the Boltzmann weighted average function served as the fitness function of choice (see EQ 3.1 and EQ 3.2). As a result, combinations of rotamers that would be rejected in SSD because they cause steric clashes on a single fixed backbone template could be

accepted in MSD if they have a stabilizing effect in the context of some or a majority of the ensemble backbone templates. Thus, the use of native backbone ensembles in MSD may help to reduce the number of false negatives within the predicted sequences.

MSD with native backbone ensembles has been shown to predict higher quality sequences than SSD with a single fixed backbone [Allen *et al.*, 2010]. Using a backbone ensemble derived from a cMD simulation, Allen *et al.* were able to predict 24 mutant sequences of G β 1 that displayed equivalent or improved stability relative to the WT. However, ensembles derived from an uMD simulation or from solution NMR structures resulted exclusively in the prediction of sequences which adopted a non-native fold or predominantly in the prediction of mutant sequences that were either less stable than WT or unfolded, respectively. These results demonstrated that the choice of input backbone ensemble significantly affected the predictive capability of the MSD simulation. However, it is currently unclear what ensemble properties and characteristics are required to provide a robust ensemble for the accurate prediction of protein sequence stability.

In this chapter, we intend to demonstrate how to improve the accuracy of protein stability predictions made with MSD by using a variety of different backbone ensembles to correctly predict the experimentally validated stability of 85 G β 1 sequences [Allen *et al.*, 2010]. Ensembles that were tested include an NMR ensemble as well as those generated by MD simulations run at different temperatures, by Backrub motions [Davis *et al.*, 2006; Lauck *et al.*, 2010], and by a new method that we have developed for the rapid generation of backbone ensembles called PertMin. MSD with

ensembles created by the PertMin method, which involves the perturbation of atomic coordinates followed by energy minimization, resulted in the most accurate predictions. This is due to the fact that PertMin ensemble members closely resemble the input crystal structure and have low potential energy. Similarly, we found that ensembles created with Backrub motions and a MD simulation at 100 K had a high degree of structural similarity with the PertMin ensembles, and also resulted in accurate predictions. Conversely, ensembles generated by MD simulations at 500 or 1000 K reduced prediction accuracy compared to SSD with the crystal structure. This was due to their low structural similarity to the crystal structure and high potential energy. Additionally, MSD with the NMR ensemble, which also exhibits low structure similarity to the crystal structure, resulted in the least accurate predictions. The ensembles tested in this chapter thus represent on- or off-target models of the native protein fold that could be used in future studies to design for desired properties other than stability, or as positive and negative states in negative design.

3.2. The Perturbation Followed by Energy Minimization Algorithm

The PertMin procedure involves the application of small coordinate perturbations into a starting protein structure to generate a set of randomly perturbed structures. An energy minimization procedure is then applied to the perturbed structures, which minimize to different local minima that become accessible because of diverging trajectories [Figure 3.2.A and Figure 3.2.B]. PertMin thus exploits the initial condition sensitivity of energy minimization [Williams and Feher, 2008]. A benefit of the PertMin

protocol is that structural deviation from the input structure and ensemble diversity (i.e. structural deviation between ensemble members) can be controlled by the number of minimization steps [Figure 3.2.C and Figure 3.2.D].

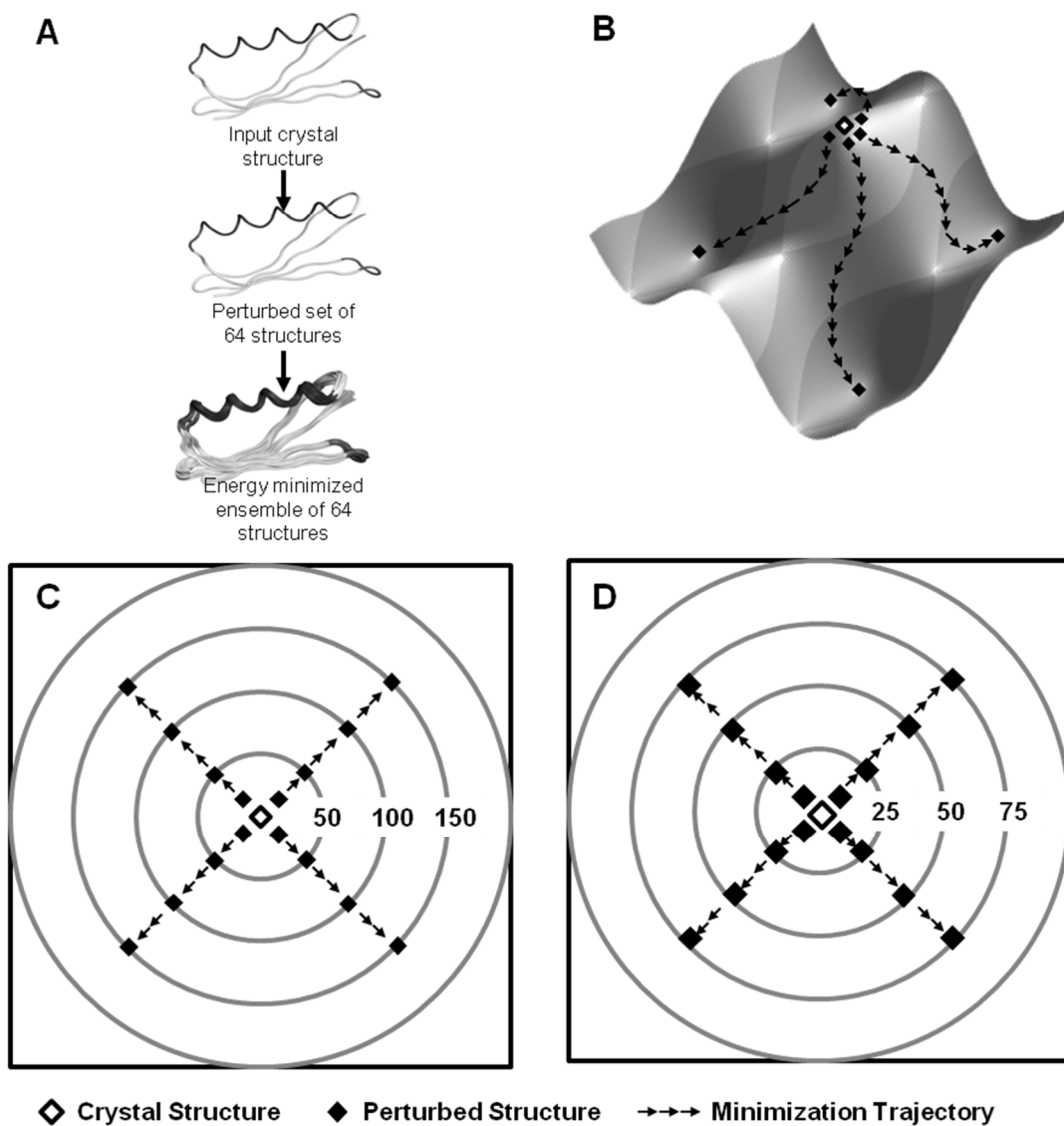


Figure 3.2. The coordinate perturbation and energy minimization method. PertMin ensembles are generated using a two-step procedure (A). Specifically, the ensembles originate from a single input structure that is repeatedly subjected to random coordinate perturbation to generate an ensemble of perturbed structures. The

coordinate perturbations applied should be small enough that the resulting set of perturbed structures are virtually indistinguishable from each other. The set of perturbed structures are then energy minimized to yield an ensemble of structures, each occupying a different local minima. In this way, PertMin exploits the known initial condition sensitive nature of energy minimization resulting from the introduction of small coordinate changes to the input protein structure (B). In PertMin, coordinate deviation from the crystal structure is directly related to the number of energy minimization iterations, represented by circles (C). For example, the starting structure (white diamond) is perturbed to yield 4 perturbed structures (black diamonds) which are energy minimized (arrows), causing them to diverge increasingly from the crystal structure as the number of minimization iterations increases. Ensemble diversity, which is the structure-to-structure divergence between members of the ensemble (represented by the circle arc) also increases with the number of minimization iterations (D).

While PertMin does not allow for a large area of protein conformational space to be explored, it enables the rapid and tunable generation of ensemble backbones having high coordinate similarity to their progenitor structure and low potential energy. Thus, application of PertMin ensembles in MSD is expected to result in improved prediction accuracy compared to SSD by reducing the number of false negatives and increasing the number of true positives. Ensemble properties, such as its root mean square (RMS) backbone coordinate deviation (RMSD) from the input structure (deviation) or backbone coordinate similarity between ensemble members (diversity) can be tuned by altering the PertMin protocol [Figure 3.3].

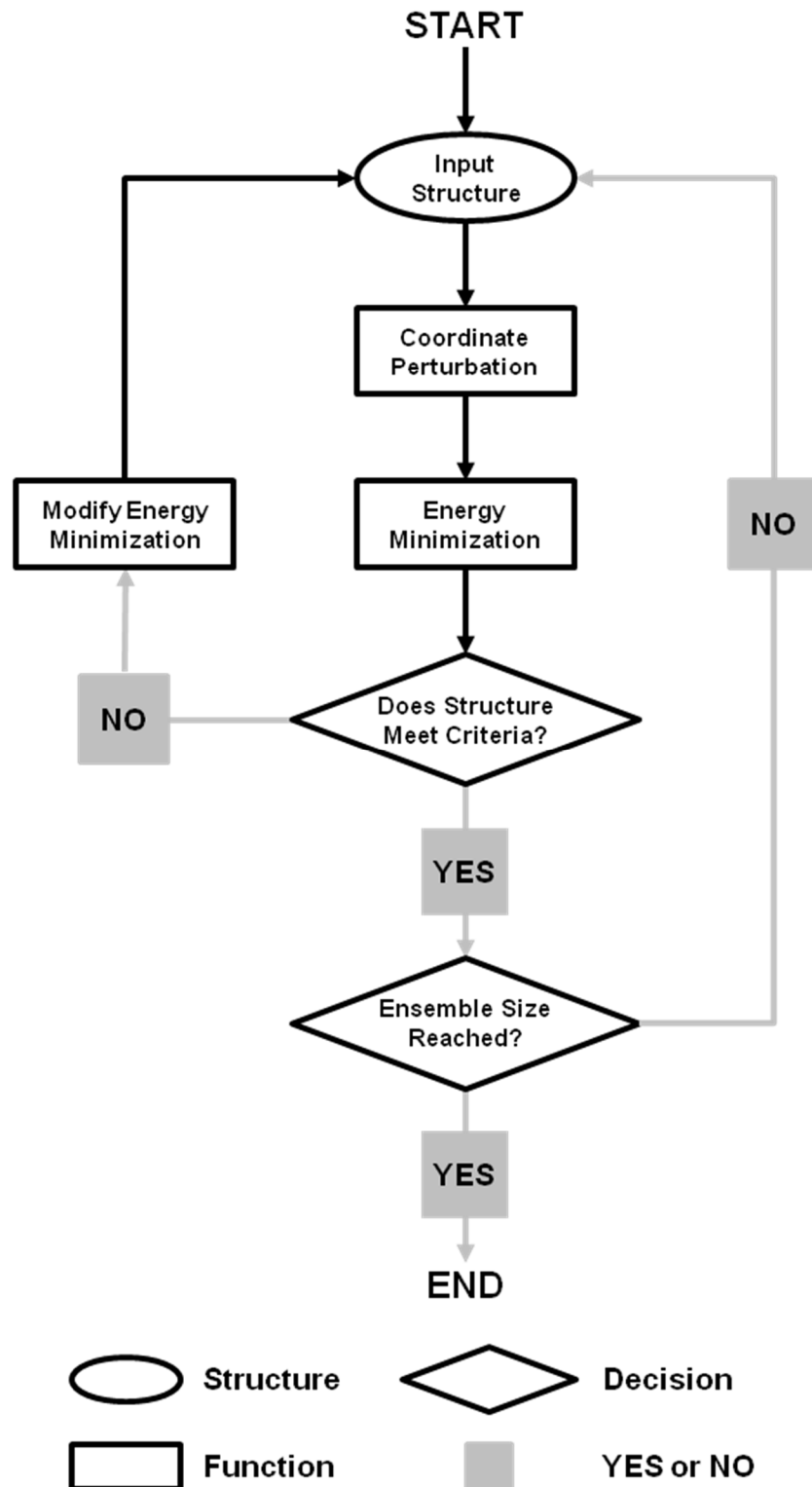


Figure 3.3. The coordinate perturbation and energy minimization algorithm.

3.3. Behaviour of the PertMin Algorithm

When generating an ensemble the choice of energy minimization algorithm can influence ensemble properties. For example, 30 member G β 1 ensembles produced by energy minimization under the AMBER force field [Wang *et al.*, 2000] with the truncated Newton algorithm produces an ensemble with higher deviation [Figure 3.4.A] and diversity [Figure 3.4.B] than the one produced with conjugate gradient minimization. This is because truncated Newton is a second order energy minimization algorithm that is more susceptible to initial condition sensitivity [Davey, 2011b] than conjugate gradient, which is first order.

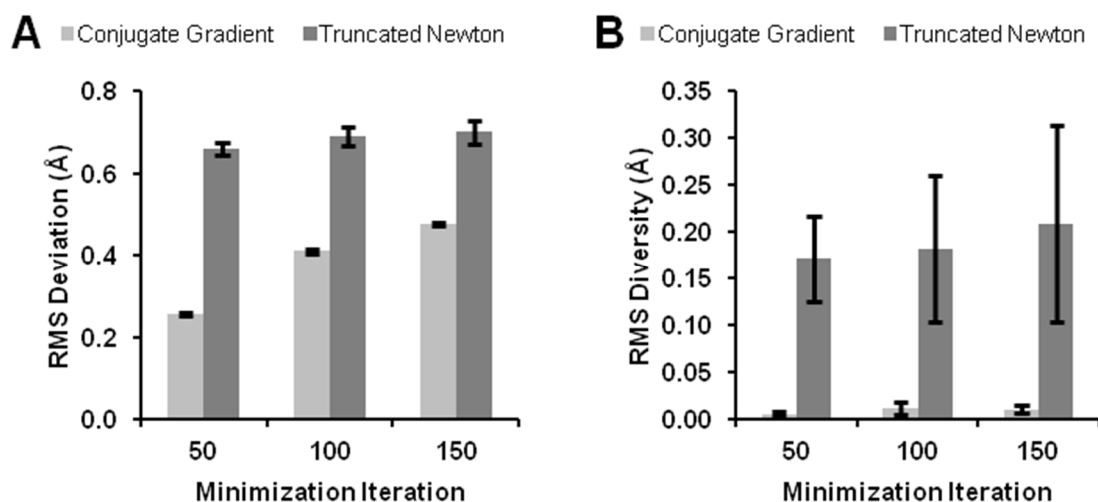


Figure 3.4. Tuning PertMin ensemble properties with minimization algorithm. 30-member G β 1 ensembles were prepared with variations on the following PertMin protocol: Cartesian coordinate perturbations of ± 0.001 Å along each axis followed by 50 iterations of either conjugate gradient or truncated Newton energy minimization in the absence of water solvent molecules. The effects of energy minimization algorithm and number of iterations on ensemble RMS coordinate deviation (A) and diversity (B) are reported.

In addition, average potential energy, as well as coordinate deviation and diversity increase with the number of minimization iterations, regardless of minimization algorithm, to a maximum value that is dependent on the location of energy minima on the protein potential energy surface.

Different types of perturbations (torsion or Cartesian) applied to the input structure coordinates will yield ensembles with similar average deviation [Figure 3.5.A] and diversity [Figure 3.5.B].

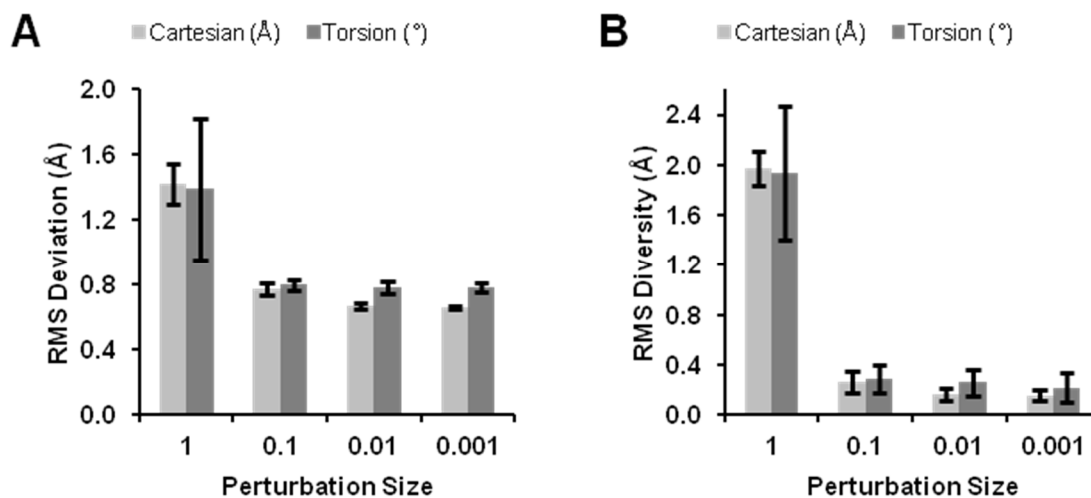


Figure 3.5. Tuning PertMin ensemble properties with perturbation size. 30-member Gβ1 ensembles were prepared with variations to the following PertMin protocol: Cartesian coordinate and bond torsion perturbations of various magnitudes were followed by 50 iterations of truncated Newton energy minimization in the absence of water solvent molecules. The effects of perturbation type and magnitude on ensemble RMS coordinate deviation (A) and diversity (B) are reported.

Perturbation magnitude does not significantly affect deviation or diversity when kept at values below or equal to 0.1 Å or degree. This is because small perturbations

result in similar perturbed structures occupying the same region of the potential energy surface, making accessible the same local minima [Davey, 2011c]. When the perturbation is sufficiently large (1 Å or degree), another region of the potential energy surface and a different set of local minima become accessible, resulting in ensembles with larger deviation and diversity.

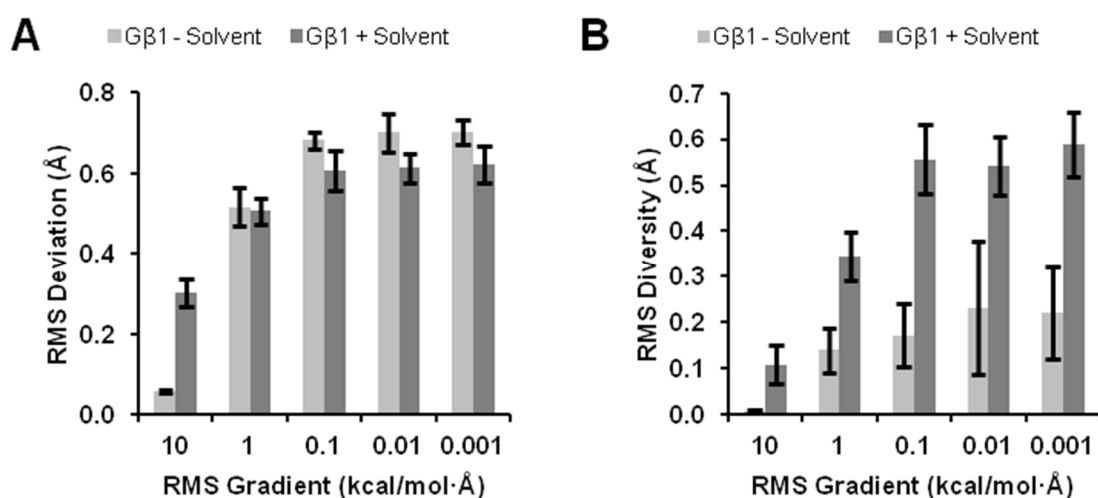


Figure 3.6. Tuning PertMin ensemble properties with system size. 30-member Gβ1 ensembles were prepared with variations to the following PertMin protocol: Cartesian coordinate perturbations of ± 0.001 Å along each axis followed by 50 iterations of truncated Newton energy minimization in the absence or presence of water solvent molecules. The effects of system size and minimization termination gradient on ensemble RMS coordinate deviation (A) and diversity (B) are reported. To increase system size, protein structures were solvated in a box of water molecules with a depth of 3 Å. In order to compare systems of different sizes, they should occupy regions on the potential energy surface located at similar distances to the nearest minimum. Therefore, energy minimizations were terminated at specific RMS gradients instead of at specific numbers of minimization iterations.

System size, i.e. the number of atoms subjected to energy minimization, will also affect the rate at which ensemble deviation [Figure 3.6.A] and diversity [Figure 3.6.B]

increase. When more atoms are included in the energy minimization calculation, for example by addition of solvent molecules, fewer number of iterations are required to produce the same amount of deviation and diversity [Davey, 2011a]. In all cases, whether altering the energy minimization protocol, perturbation method, or system size, a maximum deviation and diversity is reached. This is because PertMin generates structures at local minima, whose locations are fixed on the potential energy surface specific to the system.

3.4. Design Hypothesis and Training Set Sequences

To test the PertMin ensembles, we evaluated various ensembles with an aim to improve the accuracy of protein stability predictions made with MSD. Careful attention was paid to how calculation performance was affected by ensemble characteristics, such as coordinate similarity to crystal structure and potential energy. Importantly, we hypothesized those backbone ensembles representing a physically valid model of the desired protein fold, referred to on-target ensembles, should favour the prediction of sequences that have stability greater than the WT sequence [Figure 3.7.A]. Conversely, we hypothesized that ensembles representing non-physical models of the desired protein fold, which we call off-target ensembles, should favour the prediction of sequences that are either less stable than the WT or unfolded [Figure 3.7.B].

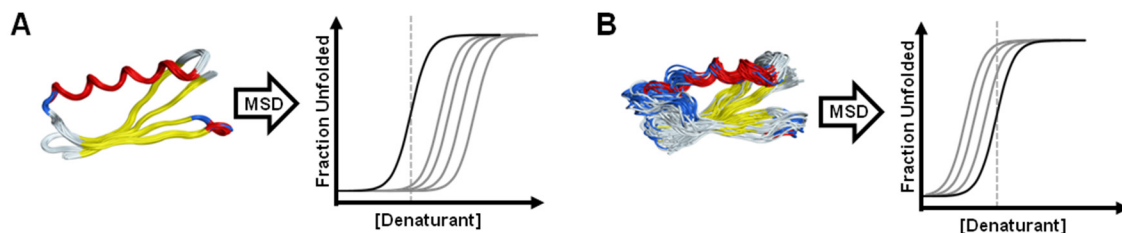


Figure 3.7. On- and off-target ensembles. An on-target ensemble (A) consisting of backbones similar to that of the crystal structure should result in the prediction by MSD of mutant sequences (grey) that are more stable than the WT sequence (black). An off-target ensembles consisting of backbones that do not represent the native fold of the protein of interest should result in the prediction of mutant sequences (grey) that are less stable than the WT sequence (black) by MSD. In both cases, which is the dashed grey line represents the C_m value of the WT, which is the concentration of denaturant at the midpoint of denaturation (C_m).

To test this hypothesis, various backbone ensembles intended to represent either on-target or off-target models of the folded G β 1 were prepared and used in MSD simulations to evaluate how accurately they could predict the known stability of 84 G β 1 mutant sequences that were previously reported [Allen *et al.*, 2010]. These 84 G β 1 sequences can be grouped into four categories reflecting their stability or folded behaviour [Figure 3.8]: (i) 24 sequences that have a stability equal to or greater than the WT sequence, (ii) 12 sequences that are folded but less stable than the WT sequence, (iii) 24 sequences that are unfolded, and (iv) 24 sequences whose unfolding did not fit a two-state model, presumably because they adopt an alternate non-native G β 1 fold [Allen *et al.*, 2010]. The specific sequence identity of the training set sequences is tabulated in table 3.1.

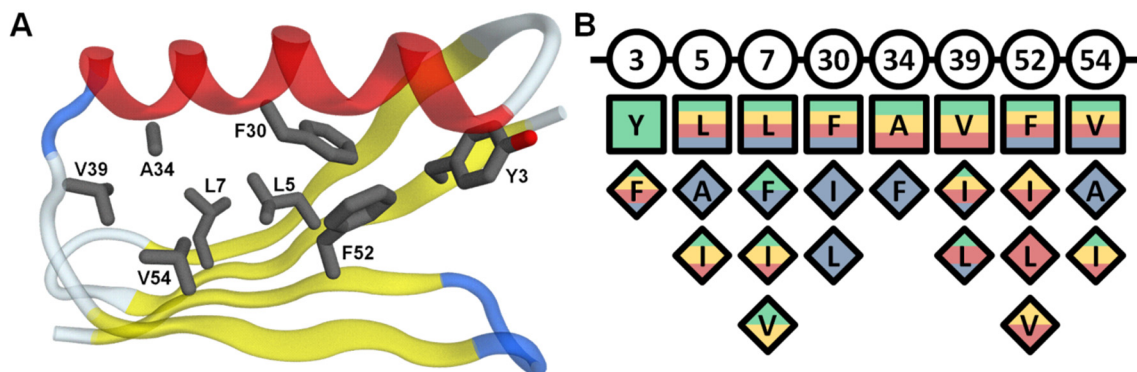


Figure 3.8. Gβ1 core residue design sequence space. The WT residues belonging to the sequence space of the Gβ1 design positions (A). Sequence permutations (B) are depicted by position (circle) showing WT (square) and mutant (diamond) residues. Sequence permutations are colored according to the folded behaviour of sequences which they belong to in the 84 member training set. Specifically: sequences more stable than the WT (green), sequences less stable than the WT (yellow), sequences that are unfolded (red), and sequences that are postulated to adopt alternative non-native Gβ1 folds (blue). The total sequence space searched during calculation consists of 5184 possible sequences.

Allen *et al.* demonstrated that MSD with a cMD ensemble resulted exclusively in the prediction of sequences displaying stability greater than or equal to the WT while MSD with an uMD ensemble resulted in the prediction of proteins that were presumed to adopt a non-native fold. MSD with a NMR structure ensemble and SSD with either the crystal structure or an average NMR structure resulted predominantly in the prediction of a combination of sequences that were unfolded or less stable than the WT. These results support our hypothesis that on-target ensembles, such as the cMD ensemble, should predict stable sequences which adopt the intended protein fold while MSD with off-target ensembles, such as the uMD ensemble, should result in the prediction of sequences that may be unfolded, have reduced stability, or adopt a non-native fold.

Stability Group	Sequence ID	Cm (M)	Design Positions												Origin Templates/Ensembles				
			3	5	7	20	26	30	34	39	43	52	54	SSD Xtal	SSD NMR-1	MSD NMR	MSD uMD	MSD cMD	
Stabilizing	WT	2.23	Y	L	L	A	A	F	A	V	W	F	V	✓	✓	✓	✓	✓	
	1	2.98	F	-	I	-	-	-	-	I	-	-	-	✓	✓	✓		✓	
	2	2.66	F	-	I	-	-	-	-	L	-	-	I					✓	
	3	2.65	F	-	F	-	-	-	-	L	-	-	I					✓	
	4	2.63	F	-	V	-	-	-	-	I	-	-	-	✓	✓			✓	
	5	2.67	F	-	I	-	-	-	-	-	-	-	-	✓	✓	✓		✓	
	6	2.62	F	-	F	-	-	-	-	I	-	-	-					✓	
	7	2.40	F	-	F	-	-	-	-	I	-	-	I					✓	
	8	2.85	F	-	I	-	-	-	-	I	-	-	I	✓				✓	
	9	2.43	F	-	I	-	-	-	-	L	-	-	-					✓	
	10	2.47	F	-	V	-	-	-	-	I	-	-	I	✓				✓	
	11	2.44	F	-	V	-	-	-	-	L	-	-	I					✓	
	12	2.40	F	-	-	-	-	-	-	-	-	-	-	✓	✓	✓		✓	
	13	2.30	F	-	F	-	-	-	-	L	-	-	-					✓	
	14	2.47	F	-	I	-	-	-	-	-	-	-	I	✓				✓	
	15	2.35	F	-	-	-	-	-	-	-	-	-	-	✓	✓	✓		✓	
	16	2.23	F	-	V	-	-	-	-	-	-	-	-	✓	✓			✓	
	17	2.13	F	-	V	-	-	-	-	L	-	-	-					✓	
	18	2.04	F	-	V	-	-	-	-	L	-	-	I					✓	
	19	2.06	F	-	V	-	-	-	-	-	-	-	I	✓				✓	
	20	1.98	F	-	-	-	-	-	-	-	I	-	-	✓				✓	
	21	2.01	F	-	-	-	-	-	-	-	L	-	-			✓		✓	
	22	1.97	F	-	F	-	-	-	-	-	-	-	-					✓	
	23	1.89	F	-	F	-	-	-	-	-	-	-	I					✓	
24	1.84	F	-	-	-	-	-	-	-	-	-	I	✓				✓		
Destabilizing	25	1.48	F	I	I	-	-	-	-	I	-	-	-	✓					
	26	1.48	F	I	I	-	-	-	-	I	-	-	I	✓					
	27	1.34	F	I	V	-	-	-	-	-	-	-	-	✓					
	28	1.16	F	I	-	-	-	-	-	I	-	-	-	✓					
	29	1.46	F	I	V	-	-	-	-	I	-	-	I	✓					
	30	0.95	F	I	-	-	-	-	-	-	-	-	-	✓					
	31	0.93	F	I	I	-	-	-	-	-	-	-	I	✓					
	32	0.81	F	I	V	-	-	-	-	-	-	-	-	✓					
	33		F	-	I	-	-	-	-	-	I	-	-		✓	✓			
	34		F	I	-	-	-	-	-	I	-	-	I	✓					
	35	0.27	F	I	-	-	-	-	-	-	-	-	I	✓					
	36		F	-	-	-	-	-	-	-	V	-	-		✓	✓			
Unfolded	37		F	I	V	-	-	-	-	-	-	I	✓						
	38		F	-	I	-	-	-	-	-	L	-	-		✓	✓			
	39		F	-	V	-	-	-	-	-	-	I	-		✓				
	40		F	-	-	-	-	-	-	-	L	-	-		✓	✓			
	41		F	I	I	-	-	-	-	-	-	-	-	✓					
	42		F	-	I	-	-	-	-	I	-	I	-		✓	✓			
	43		F	-	I	-	-	-	-	I	-	L	-		✓	✓			
	44		F	-	I	-	-	-	-	I	-	V	-		✓	✓			
	45		F	-	I	-	-	-	-	L	-	I	-			✓			
	46		F	-	I	-	-	-	-	L	-	L	-			✓			
	47		F	-	I	-	-	-	-	L	-	V	-			✓			
	48		F	-	I	-	-	-	-	-	-	V	-		✓	✓			
	49		F	-	-	-	-	-	-	I	-	I	-		✓	✓			
	50		F	-	-	-	-	-	-	I	-	L	-		✓	✓			
	51		F	-	-	-	-	-	-	I	-	V	-		✓	✓			
	52		F	-	-	-	-	-	-	L	-	I	-			✓			
	53		F	-	-	-	-	-	-	L	-	L	-			✓			
	54		F	-	-	-	-	-	-	L	-	V	-			✓			
	55		F	-	-	-	-	-	-	-	-	I	-		✓	✓			
	56		F	-	V	-	-	-	-	I	-	I	-		✓				
	57		F	-	V	-	-	-	-	-	-	L	-		✓				
	58		F	-	V	-	-	-	-	I	-	V	-		✓				
	59		F	-	V	-	-	-	-	-	-	L	-		✓				
	60		F	-	V	-	-	-	-	-	-	V	-		✓				
Non-native	61		F	A	F	-	-	-	F	I	-	-	A				✓		
	62		F	A	-	-	-	-	F	I	-	-	A				✓		
	63		F	A	F	-	-	-	I	F	I	-	A				✓		
	64		F	A	F	-	-	-	L	F	I	-	A				✓		
	65		F	A	-	-	-	-	I	F	I	-	A				✓		
	66		F	A	-	-	-	-	L	F	I	-	A				✓		
	67		F	A	F	-	-	-	F	L	-	-	A				✓		
	68		F	A	-	-	-	-	F	L	-	-	A				✓		
	69		F	A	F	-	-	-	I	F	L	-	A				✓		
	70		F	A	F	-	-	-	L	F	L	-	A				✓		
	71		F	A	-	-	-	-	I	F	L	-	A				✓		
	72		F	A	-	-	-	-	L	F	L	-	A				✓		
	73		F	A	F	-	-	-	-	F	I	-	-				✓		
	74		F	A	-	-	-	-	-	F	I	-	-				✓		
	75		F	A	F	-	-	-	-	I	F	I	-				✓		
	76		F	A	F	-	-	-	-	L	F	I	-				✓		
	77		F	A	-	-	-	-	-	I	F	I	-				✓		
	78		F	A	-	-	-	-	-	L	F	I	-				✓		
	79		F	A	F	-	-	-	-	F	L	-	-				✓		
	80		F	A	-	-	-	-	-	F	L	-	-				✓		
	81		F	A	F	-	-	-	-	I	F	L	-				✓		
	82		F	A	F	-	-	-	-	L	F	L	-				✓		
	83		F	A	-	-	-	-	-	I	F	L	-				✓		
	84		F	A	-	-	-	-	-	L	F	L	-				✓		

Figure 3.9. Gβ1 training set sequences. Amino acid residues found at each design position in the WT and mutant Gβ1 sequences (1-84) are listed. The 84 mutant

sequences are classified into one of four stability groups: 24 sequences displaying stability greater than or approximately equal to the WT (stabilizing: green), 12 sequences of lower stability than the WT (destabilizing: yellow), 24 sequences that do not fold (unfolded: red), and 24 sequences postulated to adopt an alternate non-native fold (non-native: blue). Sequence stabilities are reported as the C_m in a GndHCl denaturation experiment reported by Allen *et al.* along with check-marks indicating whether the sequences were obtained by SSD with the crystal structure (Xtal) or an average NMR structure (NMR-1), or by MSD with a NMR (NMR), uMD, or cMD ensemble [Allen *et al.*, 2010].

3.5. Creation and Characterization of G β 1 Ensembles

To ensure that the ensembles tested in this study represented either on- or off-target models of the G β 1 backbone, we employed different methods for their generation. Ensembles expected to represent off-target models include the NMR ensemble tested by Allen *et al.* as well as those created using high temperature MD simulations (500 and 1000 K) in an effort to force ensemble members to deviate significantly from the input crystal structure. On the other hand, ensembles expected to represent on-target models were created using lower temperature MD simulations (100 and 300 K) or Backrub motions, since such ensembles have been previously reported to allow for the correct prediction of protein properties by MSD [Allen *et al.*, 2010; Kapp *et al.*, 2012]. In addition, the PertMin ensemble generation methodology was employed to create additional candidate on-target ensembles.

Using the crystal structure of G β 1 (PDB code: 1PGA [Gallagher *et al.*, 1994]) as an input to these different ensemble generation methods, we created several native backbone ensembles. Four MD ensembles were generated from the energy minimized crystal structure using 100 ps unconstrained NVT simulations [Bond *et al.*, 1999;

Sturgeon and Laird, 2000] with explicit solvent under periodic boundary conditions, at 100, 300, 500, and 1000 K, and with sampling at every 1 ps during the production run to yield 64 member ensembles.

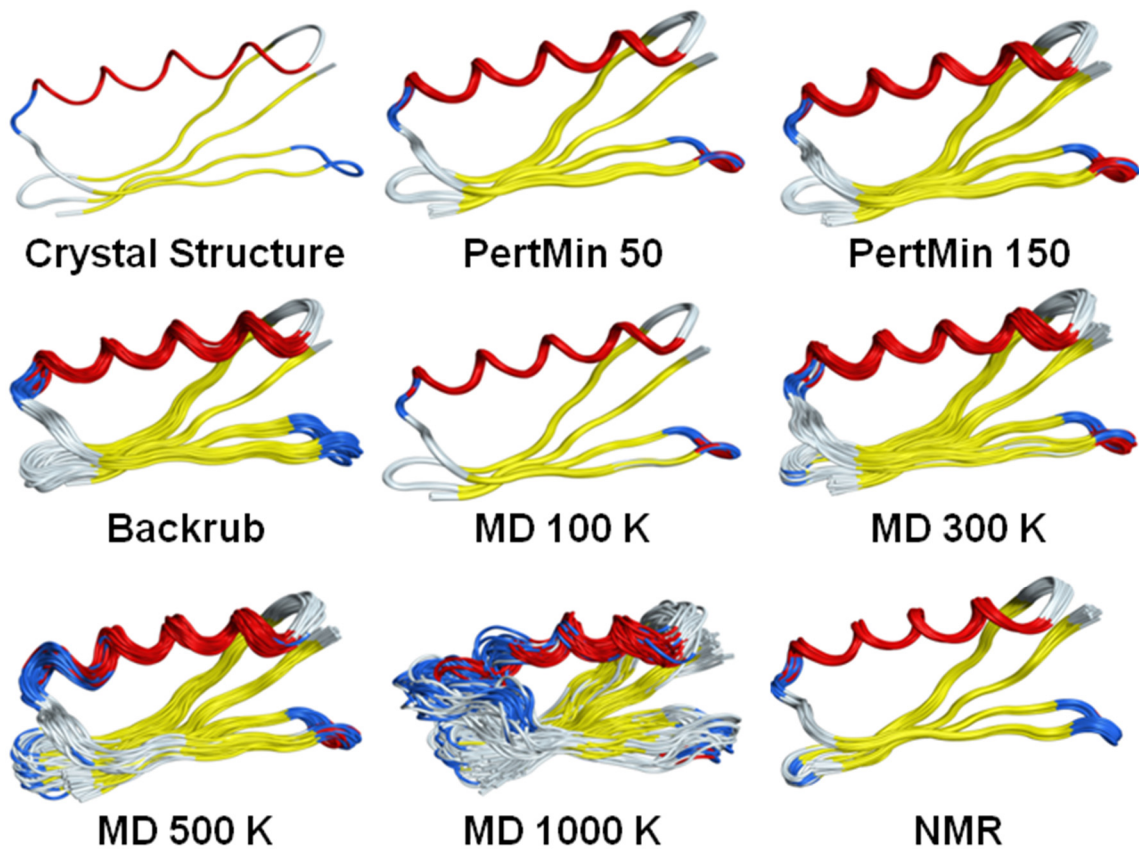


Figure 3.10. G β 1 backbone ensembles. The crystal structure of protein G domain β 1 (PDB code: 1PGA) along with superimposed members of the PertMin (50 and 150 iterations), Backrub, MD (MD 100 K, 300 K, 500 K, and 1000 K), and NMR ensembles are shown in cartoon representations. Backbones are coloured according to their secondary structural elements: α -helices, β -strands, turns, and undefined secondary structures are coloured red, yellow, blue, and white, respectively.

A Backrub ensemble containing 50 backbones was prepared from the energy minimized crystal structure using the RosettaBackrub [Lauck *et al.*, 2010]

implementation made available through the Kortemme lab server. Two PertMin ensembles were generated by randomly perturbing the coordinates of every atom of the non-minimized 1PGA structure to create perturbed structures that were subjected to either 50 or 150 iterations of truncated Newton (TN) energy minimization [Nash, 2000] in explicit solvent to yield two 64 member ensembles. In addition, a 60-member ensemble was prepared from solution NMR structures (PDB code: 1GB1 [Gronenborn *et al.*, 1991]), yielding a total of 494 unique backbones across the eight ensembles.

The eight ensembles generated as described above are shown in figure 3.10. For the ensembles generated using MD simulations, secondary structural elements are gradually lost as temperature is increased. As shown on figure 3.10, individual backbones from the MD 500 and 1000 K ensembles show a reduction in secondary structure similarity to the crystal structure, as expected given that they were generated using MD simulations run at unphysical temperatures resulting in partial unfolding of the backbones. Conversely, the NMR, Backrub, MD 100 K, MD 300 K, and both PertMin ensembles retained most of the G β 1 secondary structural elements.

The eight backbone ensembles also differ from one another according to their RMS deviation (RMSD) from the 1PGA crystal structure and their diversity, which is defined as the structural similarity between members from a same ensemble. As shown in figure 3.11, the PertMin 50, PertMin 150, and Backrub ensembles are the most similar to the crystal structure. In contrast, MD ensembles have a higher deviation from crystal structure, which increases with simulation temperature reaching up to 2.3 Å for the MD 1000 K ensemble. Interestingly, the MD 100 K ensemble, which has a 0.7 Å deviation

from crystal structure, has the smallest diversity of all the ensembles. These properties result from the fact that the MD 100 K trajectory has enough energy to escape the crystal structure potential energy well but lacks sufficient kinetic energy to explore a wider region on the potential energy surface.

		Average Backbone RMSD (Å)								
	Diversity	1PGA	PertMin 50	PertMin 150	Backrub	MD 100 K	MD 300 K	MD 500 K	MD 1000 K	NMR
PertMin 50	0.3	0.5		0.3	0.6	0.5	0.7	1.1	2.3	1.2
PertMin 150	0.4	0.5	0.3		0.7	0.5	0.7	1.1	2.3	1.2
Backrub	0.6	0.5	0.6	0.7		0.8	0.8	1.2	2.4	1.1
MD 100 K	0.2	0.7	0.5	0.5	0.8		0.7	1.2	2.4	1.3
MD 300 K	0.6	0.8	0.7	0.7	0.8	0.7		1.2	2.4	1.2
MD 500 K	0.9	1.1	1.1	1.1	1.2	1.2	1.2		2.2	1.5
MD 1000 K	2.2	2.3	2.3	2.3	2.4	2.4	2.4	2.2		2.4
NMR	0.4	1.0	1.2	1.2	1.1	1.3	1.2	1.5	2.4	

Figure 3.11. Structural comparison of Gβ1 backbone ensembles. Average backbone coordinate RMS deviation (RMSD, in Å) for each ensemble with respect to members within the same ensemble (diversity), the input crystal structure (1PGA), and members of other ensembles. Colours are assigned according to RMSD values: 0.2 – 0.5 Å (green), 0.6 – 0.9 Å (yellow), 1.0 – 1.9 Å (orange), and ≥ 2.0 Å (red). Ensemble names are as described in text.

The PertMin 50 and 150 ensembles also have low diversity due to the fact that energy minimization is not a method that allows for exploration of the potential energy surface, and only solves the location of the nearest local minimum. As was the case for the deviation from crystal structure, MD ensembles have diversities which increase with simulation temperature reaching 2.2 Å for the MD 1000 K ensemble. Interestingly, the NMR ensemble is unique in that it deviates significantly from the crystal structure while maintaining a low diversity. Finally, the Backrub ensemble is the only ensemble which has a diversity that is greater than its average deviation from crystal structure, a consequence of the ability of Backrub motions to mimic high resolution dynamics from solution NMR while maintaining structural similarity to the input crystal structure [Friedland and Kortemme, 2010].

Table 3.1. Potential energy and RMSG of backbone templates and ensembles

Input Structures	Potential Energy ^a (kcal·mol ⁻¹)	RMSG ^a (kcal·mol ⁻¹ ·Å ⁻¹)
1PGA	-195	20
Minimized Crystal	-796	2
PertMin 50	-680 ± 9	7 ± 1
PertMin 150	-678 ± 10	7 ± 1
Backrub	39 ± 51	320 ± 88
MD 100 K	-483 ± 14	19 ± 1
MD 300 K	-51 ± 25	31 ± 1
MD 500 K	326 ± 34	39 ± 1
MD 1000 K	1506 ± 89	55 ± 1
NMR	-231 ± 31	3 ± 1

^a calculated with the AMBER99 force field [Wang *et al.*, 2000] with a distance dependent dielectric in the absence of explicit solvent or a periodic boundary

Other properties that differentiate the eight ensembles are their average potential energy and RMS gradient (RMSG), which are reported on table 3.1. The calculation of potential energy and RMSG was carried out in MOE [MOE, 2011] using the AMBER99 force field [Wang *et al.*, 2000] with a distance dependent dielectric of 80. These calculations were conducted in the absence of explicit solvent or periodic boundary conditions in order to be consistent with the CPD environment present during sequence optimization. The average potential energy of an ensemble gives an approximation of its stability while the RMSG reflects the location of ensemble members on the potential energy surface relative to minimum (found at 0 kcal·mol⁻¹·Å⁻¹).

Taken together, these parameters are expected to be similar to or lower than those of the 1PGA crystal structure for on-target ensembles and significantly higher for off-target ensembles. The energy minimized crystal structure, as well as the PertMin 50, PertMin 150, MD 100 K, and NMR ensembles are the only structures with an average potential energy and RMSG that are lower than or equal to that of the 1PGA input crystal structure. These properties result from the fact that the minimized crystal structure, the NMR, and the PertMin ensembles are prepared via energy minimization and that the MD 100 K ensemble structures lack sufficient kinetic energy to deviate significantly from minimum.

As expected, the potential energy and RMSG of MD ensembles increase with the temperature of simulation because of the available kinetic energy that allows the exploration of additional regions on the potential energy surface. Interestingly, the Backrub ensemble, even though it was prepared from the minimized crystal structure,

has a high potential energy and the highest RMSG. These properties result from the fact that its structures are generated via rigid coordinate transformations [Ollikainen *et al.*, 2013] and not via exploration of or minimization on the potential energy surface.

3.6. Protein Stability Design Tasks

Based on the different properties of the ensembles described above, we postulated that the PertMin 50, PertMin 150, Backrub, MD 100 K, and MD 300 K ensembles should represent on-target models of the G β 1 fold while the NMR, MD 500, and 1000 K ensembles should represent off-target models. In order to test this hypothesis, we used each of the ensembles as inputs for MSD simulations and evaluated their ability at correctly predicting the known stability of WT G β 1 and its 84 mutants [Allen *et al.*, 2010]. To do so, the fitness values of these 85 G β 1 sequences were calculated using each ensemble to generate a ranked list of sequences reflecting their stability. For these calculations, fitness was computed as the Boltzmann weighted average across the individual scores of sequences on each backbone member of the ensemble. This averaging method was selected because it has previously been shown to allow for the successful prediction of stable sequences using MSD [Allen *et al.*, 2010]. In addition, the 85 sequences were also ranked by their SSD score for comparison, using the energy minimized crystal structure as the sole backbone template.

In order to validate the predictions at each task [Figure 3.12], the 84 protein G domain β 1 mutant sequences reported by Allen *et al.* [Allen *et al.*, 2010] were used. These mutant sequences include 24 that are as or more stable than the WT (black circle)

sequence (green circles), 12 that are less stable than the WT (yellow circles), 24 that are unfolded (red circles), and 24 that are presumed to adopt a non-native fold (blue circles). The WT sequence is represented by a white circle. (A) In sequence enrichment, an on-target ensemble is expected to result in a majority of sequences with stability greater than or equal to the WT in the top 25. Conversely, an off-target ensemble will result in a majority of sequences that are less stable than WT, unfolded, and that adopt a non-native fold. (B) In sequence binning, an ensemble that is on-target would ideally rank the 84 mutant sequences relative to the WT fitness value cut-off (dotted line) so as to eliminate any false positives or negatives, while an off-target ensemble would not be able to do so. (C) Correlation of *in silico* fitness with *in vitro* stability, represented by the C_m , is expected to be close to unity for an on-target ensemble, or close to 0 for an off-target ensemble. Next, the lists of ranked sequences obtained by MSD were used to evaluate the performance of each ensemble at three different CPD tasks of increasing difficulty [Figure 3.12]. The first CPD task is sequence enrichment [Figure 3.12.A], which is defined as the ability of CPD to identify stable protein sequences and score them favourably such that they represent a majority of the top-ranked sequences. Since the 85 G β 1 sequences tested herein contain 25 sequences with stability greater than or equal to that of the WT, it can be expected that MSD with an on-target ensemble should result in a list of sequences containing a majority, or exclusively, stable sequences [Figure 3.12.A, left] in the top 25. On the other hand, an off-target ensemble would result in a list whose 25 top ranked sequences would consist of a majority of sequences that are less stable than WT, unfolded, or adopt non-native folds [Figure 3.12.A, right].

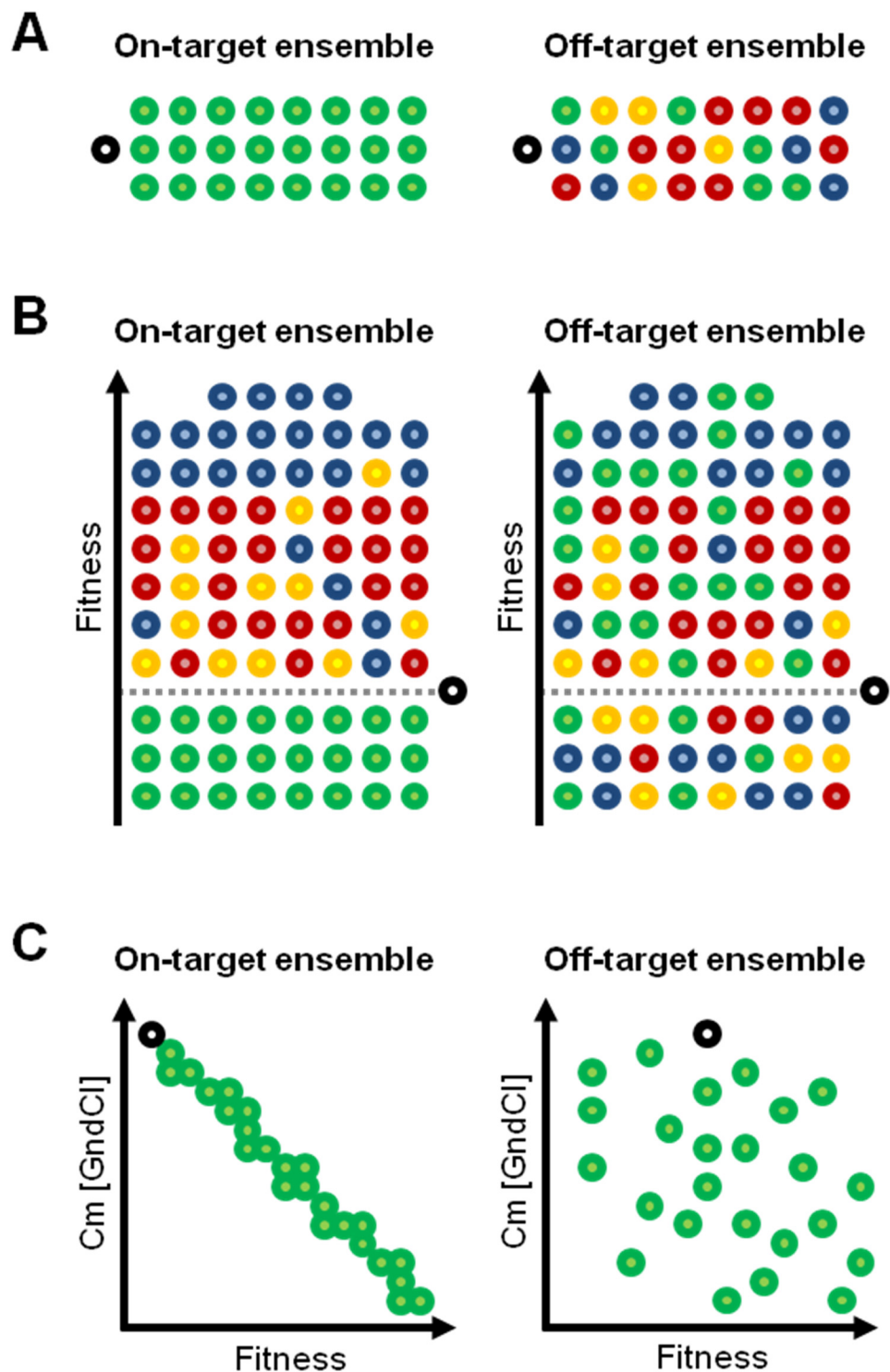


Figure 3.12. Protein stability design tasks. The three tasks used to test the accuracy of protein stability predictions made by MSD. These tasks include: (A) the ability to enrich predictions with stable sequences, (B) the ability to bin a specific set of sequences as stable or unstable, and (C) the ability to predict the stability of sequences.

The second CPD task involves the correct binning of sequences as either stable or unstable by comparing their fitness value relative to a cut-off. An ensemble that is an on-target model of the G β 1 fold would ideally rank the 85 sequences relative to the cut-off so as to eliminate any false positive or false negative sequences [Figure 3.12.B, left], while an off-target ensemble would not be able to do so [Figure 3.12.B, right].

The third task is the correlation of *in silico* fitness with *in vitro* stability. This task is the most difficult because it involves not only correctly accepting stable sequences and rejecting unstable ones but also correctly ordering the sequences according to their experimentally measured stability. For this task, an on-target ensemble should provide a correlation close to unity [Figure 3.12.C, left] while an off-target ensemble would be unable to correlate predicted with experimentally measured stability [Figure 3.12.C, right].

To evaluate these tasks single-state and multistate CPD were conducted using the FASTER algorithm [Allen and Mayo, 2006; Allen and Mayo, 2010] for sequence optimization. The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 [Dunbrack and Cohen, 1997] was used to provide side chain conformations to be threaded onto each fixed backbone template. Sequences were scored using a potential energy function consisting of a van der Waals term from the Dreiding II force field with atomic radii scaled by 0.9 [Mayo *et al.*, 1990], a direction specific hydrogen-bond term having a well depth of 8.0 kcal·mol⁻¹ [Dahiyat and Mayo, 1997], an electrostatic energy term modelled using Coulomb's law with a distance dependent dielectric of 40, and a surface area-based solvation penalty term [Lazaridis

and Karplus, 1999; Street and Mayo, 1998]. CPD was implemented using PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012]. Multistate sequence fitness was reported as a Boltzmann weighted average at 300 K from the individual scores for each backbone from the ensemble. The Boltzmann weighted average (E) for a given sequence is calculated by calculating the average energy of each state scale its respective Boltzmann weight (S_i) across an ensemble of n members. Thus, the Boltzmann weight represents the importance of the state with larger values indicating states that are more prevalently in the ensemble population. The Boltzmann weight for a given state is calculated as shown (Equation 3.1):

$$S = e^{(-E_i/kT)} \quad (\text{EQ 3.1})$$

where k is the Boltzmann constant ($1.987204118 \times 10^{-3}$ kcal·mol⁻¹·K⁻¹), T is the temperature in Kelvin, and E_i is the energy of the sequence on each specific state (i). The average of the Boltzmann weighted energies (E) represents the energy of the prevalently populated state in the ensemble (Equation 3.2).

$$E = \sum_{i=1}^n \frac{E_i S_i}{S_i} \quad (\text{EQ 3.2})$$

Evaluation of sequence fitness as the Boltzmann weighted average ensures that sequences that stabilize a majority of ensembles members are not penalized if they destabilize a few. Alternatively, the energy of a sequence on its most favorable scoring state can also be used as its fitness value [Howell *et al.*, 2014].

3.7. Enrichment of Sequence Predictions

To evaluate how each of the eight ensembles performed at sequence enrichment, the top 25 sequences resulting from MSD were analyzed to determine how many were as or more stable than WT, were less stable than WT, were unfolded, or adopted a non-native fold. For comparison, the same analysis was performed on the ranked list of sequences obtained by SSD with the minimized crystal structure.

Table 3.2. Sequence enrichment of SSD and MSD calculations

Input Structures	$\geq$ WT ^a	< WT ^b	Unfolded ^c	Non-native ^d
Minimized Crystal	15	6	4	0
PertMin 50	19	0	4	2
PertMin 150	17	3	5	0
Backrub	18	2	4	1
MD 100 K	18	0	7	0
MD 300 K	16	4	4	1
MD 500 K	11	5	7	2
MD 1000 K	11	2	11	1
NMR	4	1	20	0

^a Number of sequences in the top 25 that are as or more stable than WT G β 1

^b Number of sequences in the top 25 that are less stable than WT G β 1

^c Number of sequences in the top 25 that are unfolded

^d Number of sequences in the top 25 that are postulated to adopt a non-native fold

As can be seen on table 3.2, SSD using the minimized crystal structure resulted in a majority (60%) of the top 25 sequences being those that are known to be as or more stable than WT. This result demonstrates that SSD achieves what it is designed to do, which is to identify sequences that stabilize the input backbone template. Similarly, use of any ensemble in MSD, with the exception of the MD 500 K, MD 1000 K, and NMR ensembles, resulted in a majority of sequences that are as or more stable than WT in

the top 25. The NMR ensemble performed the worst with only 4 of the top 25 sequences being of equal or greater stability than the WT, a result in agreement with published data [Allen *et al.*, 2010].

Thus, sequence enrichment results suggest that the MD 500 K, MD 1000 K, and NMR ensembles are off-target. Conversely, the PertMin 50 ensemble was the best at performing sequence enrichment as it allowed for the greatest number of sequences with stability greater than or equal to the WT and the lowest number of unfolded sequences found in the top 25. This result confirms that it is an on-target model for G β 1. Results for the PertMin 150, Backrub, MD 100 K, and MD 300 K ensembles suggest that they are also on-target models of the G β 1 folded state. These results also suggest that while ensembles must maintain high coordinate similarity to the crystal structure to ensure that predictions are successfully enriched with desired sequences, some coordinate deviation is required to better accommodate structural differences arising from the combinatorial sequence space evaluated.

3.8. Binning of Sequence Predictions

Next, the performance of the eight ensembles at the second CPD task were evaluated [Figure 3.12.B]. This task is more difficult than sequence enrichment because it requires the assignment of a relevant cut-off value and the correct binning of sequences as being either stable or unstable. For our analysis, unstable sequences include all sequences that are undesirable such as those that are less stable than WT,

unfolded, or that are postulated to adopt a non-native fold. For this task, an ideal cut-off value would correctly eliminate all false positive and false negative sequence predictions. Since such a cut-off is design problem specific, we decided to use the computed fitness of the WT sequence as our cut-off value. This fitness value was selected because the WT sequence is known to be stable in the context of the G β 1 fold, and because the MSD simulation is expected to rank the WT sequence favourably on an on-target ensemble. Thus, any sequence ranked more favourably than the WT sequence would be expected to adopt the G β 1 fold and have stability greater than or equal to the WT. On the contrary, any sequence ranked less favourably than the WT sequence should be either unfolded, adopt a non-native fold, or be folded but have stability lower than that of the WT sequence. Following the assignment of the cut-off value, we evaluated the 85 sequences from the ranked lists obtained by SSD with the minimized crystal structure or by MSD with the eight ensembles and classified them into one of four categories [Figure 3.13]: (i) correct positives are sequences that are accepted *in silico* (i.e. that have a score/fitness value below the WT cut-off) and have an experimentally measured stability greater than or equal to the WT sequence, (ii) false negatives are sequences that are rejected *in silico* (i.e. that have a score/fitness value above the WT cut-off) but whose experimentally measured stability is greater than or equal to the WT sequence, (iii) false positives are sequences that are accepted *in silico* but that display a decreased stability compared to the WT sequence or are unfolded or adopt a non-native fold, and (iv) correct negatives are sequences that are rejected *in silico* and are less stable than WT, unfolded, or adopt a non-native fold.

Mutations from wild-type (WT)	#	Crystal	PertMin 50	PertMin 150	Backrub	MD 100 K	MD 300 K	MD 500 K	MD 1000 K	NMR
Y3F/L7I/V39I	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39L/V54I	2	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7F/V39L/V54I	3	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/L7V/V39I	4	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7I	5	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7F/V39I	6	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/L7F/V39I/V54I	7	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/L7I/V39I/V54I	8	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39L	9	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7V/V39I/V54I	10	✗	✓	✓	✓	✓	✗	✗	✗	✗
Y3F/L7V/V39L/V54I	11	✗	✓	✓	✓	✓	✗	✗	✗	✗
Y3F/V39I	12	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7F/V39L	13	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/L7I/V54I	14	✓	✓	✓	✓	✓	✗	✗	✗	✗
Y3F	15	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7V	16	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7V/V39L	17	✗	✓	✓	✓	✓	✓	✓	✓	✗
Y3F/V39L/V54I	18	✗	✓	✗	✓	✓	✓	✓	✓	✗
Y3F/L7V/V54I	19	✗	✓	✗	✓	✓	✗	✗	✗	✗
Y3F/V39I/V54I	20	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/V39L	21	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7F	22	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/L7F/V54I	23	✗	✗	✗	✗	✗	✗	✗	✗	✗
Y3F/V54I	24	✓	✓	✓	✓	✓	✓	✓	✓	✓
WT		✓	✓	✓	✓	✓	✓	✓	✓	✓
Cut-off										
Y3F/L5I/L7I/V39I	25	✓	✗	✗	✓	✗	✓	✓	✓	✓
Y3F/L5I/L7I/V39I/V54I	26	✗	✗	✗	✗	✗	✓	✓	✓	✗
Y3F/L5I/L7V/V39I	27	✗	✗	✗	✗	✗	✓	✓	✓	✓
Y3F/L5I/V39I	28	✓	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L5I/L7V/V39I/V54I	29	✗	✗	✗	✗	✗	✓	✓	✓	✗
Y3F/L5I	30	✓	✗	✗	✗	✗	✓	✓	✓	✓
Y3F/L5I/L7I/V54I	31	✓	✗	✗	✗	✗	✓	✓	✓	✓
Y3F/L5I/L7V	32	✗	✗	✗	✗	✗	✓	✓	✓	✓
Y3F/L7I/F52I	33	✗	✗	✗	✓	✗	✗	✗	✗	✓
Y3F/L5I/V39I/V54I	34	✗	✗	✗	✗	✗	✓	✓	✓	✗
Y3F/L5I/V54I	35	✗	✗	✗	✗	✗	✓	✓	✓	✗
Y3F/F52V	36	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5I/L7V/V54I	37	✗	✗	✗	✗	✗	✗	✓	✓	✗
Y3F/L7I/F52L	38	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7V/F52I	39	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/F52L	40	✗	✗	✗	✗	✓	✓	✓	✓	✓
Y3F/L5I/L7I	41	✓	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39I/F52I	42	✗	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39I/F52L	43	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39I/F52V	44	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39L/F52I	45	✗	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39L/F52L	46	✗	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/L7I/V39L/F52V	47	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L7I/F52V	48	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/V39I/F52I	49	✗	✗	✗	✗	✓	✓	✓	✓	✓
Y3F/V39I/F52L	50	✗	✓	✓	✓	✓	✓	✓	✓	✗
Y3F/V39I/F52V	51	✗	✗	✗	✗	✓	✓	✓	✓	✓
Y3F/V39L/F52I	52	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/V39L/F52L	53	✗	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/V39L/F52V	54	✗	✗	✗	✗	✗	✗	✓	✓	✗
Y3F/F52I	55	✗	✗	✗	✗	✓	✓	✓	✓	✓
Y3F/L7V/V39I/F52I	56	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L7V/V39I/F52L	57	✗	✗	✗	✓	✓	✓	✓	✓	✓
Y3F/L7V/V39I/F52V	58	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L7V/F52L	59	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L7V/F52V	60	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/A34F/V39I/V54A	61	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/A34F/V39I/V54A	62	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30I/A34F/V39I/V54A	63	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30L/A34F/V39I/V54A	64	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30I/A34F/V39I/V54A	65	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30L/A34F/V39I/V54A	66	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/A34F/V39L/V54A	67	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/A34F/V39L/V54A	68	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30I/A34F/V39L/V54A	69	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30L/A34F/V39L/V54A	70	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30I/A34F/V39L/V54A	71	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30L/A34F/V39L/V54A	72	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/A34F/V39I	73	✗	✗	✗	✗	✗	✓	✓	✓	✓
Y3F/L5A/A34F/V39I	74	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L5A/L7F/F30I/A34F/V39I	75	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30L/A34F/V39I	76	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30I/A34F/V39I	77	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30L/A34F/V39I	78	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/A34F/V39L	79	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/A34F/V39L	80	✗	✓	✓	✓	✓	✓	✓	✓	✓
Y3F/L5A/L7F/F30I/A34F/V39L	81	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/L7F/F30L/A34F/V39L	82	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30I/A34F/V39L	83	✗	✗	✗	✗	✗	✗	✓	✓	✓
Y3F/L5A/F30L/A34F/V39L	84	✗	✗	✗	✗	✗	✗	✓	✓	✗

Figure 3.13. Sequence binning of SSD and MSD calculations. Binning of the 85 Gβ1 sequences obtained by SSD with the crystal structure or MSD with the eight ensembles

with respect to a cut-off corresponding to the fitness of the WT sequence. Sequences are ordered according to their experimentally measured stabilities [Allen *et al.*, 2010] and grouped based on the following scheme: mutant sequences that are equally as or more stable than WT (green), the WT sequence (white), mutant sequences that are less stable than WT (yellow), unfolded mutant sequences (red), and mutant sequences postulated to adopt a non-native fold (blue). The maximum number of mutations to the WT sequence incorporated into the set of variant sequences is six. True positives, false negatives, false positives, and true negatives are indicated by green check-marks, red x-marks, red check-marks, or green x-marks, respectively.

Table 3.3 shows sequence binning results relative to the WT cut-off value for MSD with each of the eight ensembles as well as for SSD with the energy minimized crystal structure. As can be seen on table 3.3, SSD with the crystal structure results in the correct binning of 77% of the 85 G β 1 sequences. 16% of these 85 sequences are falsely rejected because of the known artefact of SSD resulting from the combined use of a single fixed backbone template and discrete rotamers.

However, using MSD with any of the ensembles, with the exception of the MD 500 K ensemble, reduces the number of false negative sequences. This result agrees with previous reports that demonstrated that MSD with native backbone ensembles decreases the number of false negative sequences compared to SSD [Allen *et al.*, 2010; Kapp *et al.*, 2012]. Most ensembles, with the exception of the PertMin 150 ensemble, achieve this at the expense of increasing the number of false positive sequences. This unintended consequence results from the fact that MSD finds the lowest possible fitness for all sequences regardless of their experimentally measured stability, due to the use of a Boltzmann weighted average that favours the few backbones that score the sequence favourably. Thus, unstable sequences are expected to be scored more favourably by MSD than SSD.

Table 3.3. Sequence binning of SSD and MSD calculations

Input Structures	WT Cut-off ^a (kcal·mol ⁻¹)	Success Rate ^b	Number of False Negatives	Number of False Positives
Minimized Crystal	-76	77%	14	5
PertMin 50	-74	86%	6	6
PertMin 150	-77	88%	9	1
Backrub	-69	84%	7	7
MD 100 K	-70	82%	6	9
MD 300 K	-67	65%	2	28
MD 500 K	-71	67%	14	14
MD 1000 K	-49	58%	10	26
NMR	-53	51%	11	31

^a The WT cut-off value shows the single-state score or the multistate fitness value for the WT sequence computed with the minimized crystal structure or each ensemble, respectively.

^b The percentage of correctly binned sequences relative to the WT cut-off.

Because the fitness values of the WT sequence obtained using the PertMin ensembles are the lowest, the number of false positive sequences is decreased compared to other ensembles, resulting in success rates approaching 90%, the highest for all ensembles. Conversely, because the MD 1000 K and NMR ensembles result in the least stringent fitness values for the WT sequence, the success rates for these ensembles are the lowest and the number of false positives are amongst the highest. Ensembles which were able to improve sequence binning over SSD include the PertMin 50, PertMin 150, Backrub, and MD 100 K ensembles, demonstrating that they are on-target. Again, the MD 500 K, MD 1000 K, and NMR ensembles performed poorly with a large number of false positives and negatives, demonstrating that these ensembles are

off-target. Although the MD 300 K ensemble improved sequence enrichment, it may not be on-target as it cannot perform sequence binning as successfully as the crystal structure. Indeed, the MD 300 K ensemble resulted in the second highest number of false positives and a success rate that is 12% lower than that obtained by SSD with the minimized crystal structure.

Table 3.4. Potential energy and RMSG of minimized ensembles

Ensembles	Potential Energy ^a (kcal·mol ⁻¹)	RMSG ^a (kcal·mol ⁻¹ ·Å)
Minimized Backrub	-97 ± 37	6 ± 1
Minimized MD 100 K	-573 ± 9	6 ± 1
Minimized MD 300 K	-304 ± 14	6 ± 1
Minimized MD 500 K	-234 ± 28	6 ± 1
Minimized MD 1000 K	-26 ± 52	6 ± 1

^a Calculated with the AMBER99 force field [Wang *et al.*, 2000] with a distance dependent dielectric in the absence of explicit solvent or a periodic boundary

Since the success rate of sequence binning is dependent on the value of the WT cut-off for each ensemble, we investigated whether the use of energy minimized ensembles in MSD would result in the assignment of more stringent cut-off values thereby improving binning results. To evaluate this, we energy minimized all of the ensembles that had not been previously minimized (i.e. the Backrub and MD ensembles) such that their RMSG values fell below that of the PertMin ensembles [Table 3.4] in order to ensure that all ensembles tested were at a similar distance from the nearest

local minimum. These newly minimized ensembles were then used as input templates for MSD and sequence binning results were evaluated [Table 3.5].

Table 3.5. Sequence binning results of MSD using energy minimized ensembles

Ensembles	WT Cut-off ^a (kcal·mol ⁻¹)	Success Rate ^b	Number of False Negatives	Number of False Positives
Minimized Backrub	-71	86%	9	3
Minimized MD 100 K	-68	76%	6	14
Minimized MD 300 K	-68	68%	7	20
Minimized MD 500 K	-76	80%	13	4
Minimized MD 1000 K	-61	78%	13	6

^a The WT cut-off value shows the multistate fitness value for the WT sequence computed with each minimized ensemble.

^b The percentage of correctly binned sequences relative to the WT cut-off.

For the minimized Backrub and MD ensembles, with the exception of the minimized MD 100 K ensemble, sequence binning success rates were improved relative to their non-minimized inputs. The most significant improvements were found for the minimized MD 500 and 1000 K ensembles, whose success rates were increased by 13 and 20 %, respectively. These improvements result from the use of more stringent WT cut-off values obtained following energy minimization that lead to large decreases in the number of false positives, in a similar fashion to results obtained by MSD using the PertMin pair of ensembles. On the other hand, the use of more stringent cut-off values for the minimized ensembles did not lead to reductions in the number of false negatives, with the exception of the minimized MD 500 K ensemble, whose number of false negatives decreased by one sequence. Interestingly, the minimized MD 100 K

ensemble is the only minimized ensemble for which the WT cut-off value increased following energy minimization, resulting in a lower success rate. The inferior performance of the MD 100 K ensemble following energy minimization results from a change in the distribution of individual single-state scores across ensemble members which narrowed the gap between the score of the most favorable state and those of its neighbours, leading to an inflated Boltzmann weighted average fitness value for the WT sequence. These results demonstrate that the use of energy minimization to prepare ensembles does not necessarily guarantee improved sequence binning results. Finally, energy minimization of off-target ensembles (MD 500 and 1000 K) did improve their sequence binning ability, resulting in success rates and numbers of false negatives and positives that are similar to those obtained by SSD with the minimized crystal structure. These results demonstrate that although energy minimization can improve the predictive ability of off-target ensembles, use of such ensembles in MSD does not lead to enhanced performance relative to SSD.

While the WT sequence fitness serves as both a practical and predictive cut-off, we investigated whether alternate cut-off values could be used. To do so, we calculated the true positive (fraction of true positives out of the positives) and false positive (fraction of false positives out of the negatives) rates for every possible cut-off value at 1 kcal·mol⁻¹ increments in the range of fitness values obtained with all eight ensembles (0 to -90 kcal·mol⁻¹), and plotted them to generate receiver operating characteristic (ROC) curves for each ensemble [Figure 3.14]. ROC curves allow for the evaluation of a binary classifier system as the cut-off is varied, helping to determine if there is an ideal cut-off

to be used. Our data demonstrates that all on-target ensembles (PertMin 50, Pertmin 150, Backrub, and MD 100 K) result in high true positive and low false positive ratios with the use of the WT cut-off.

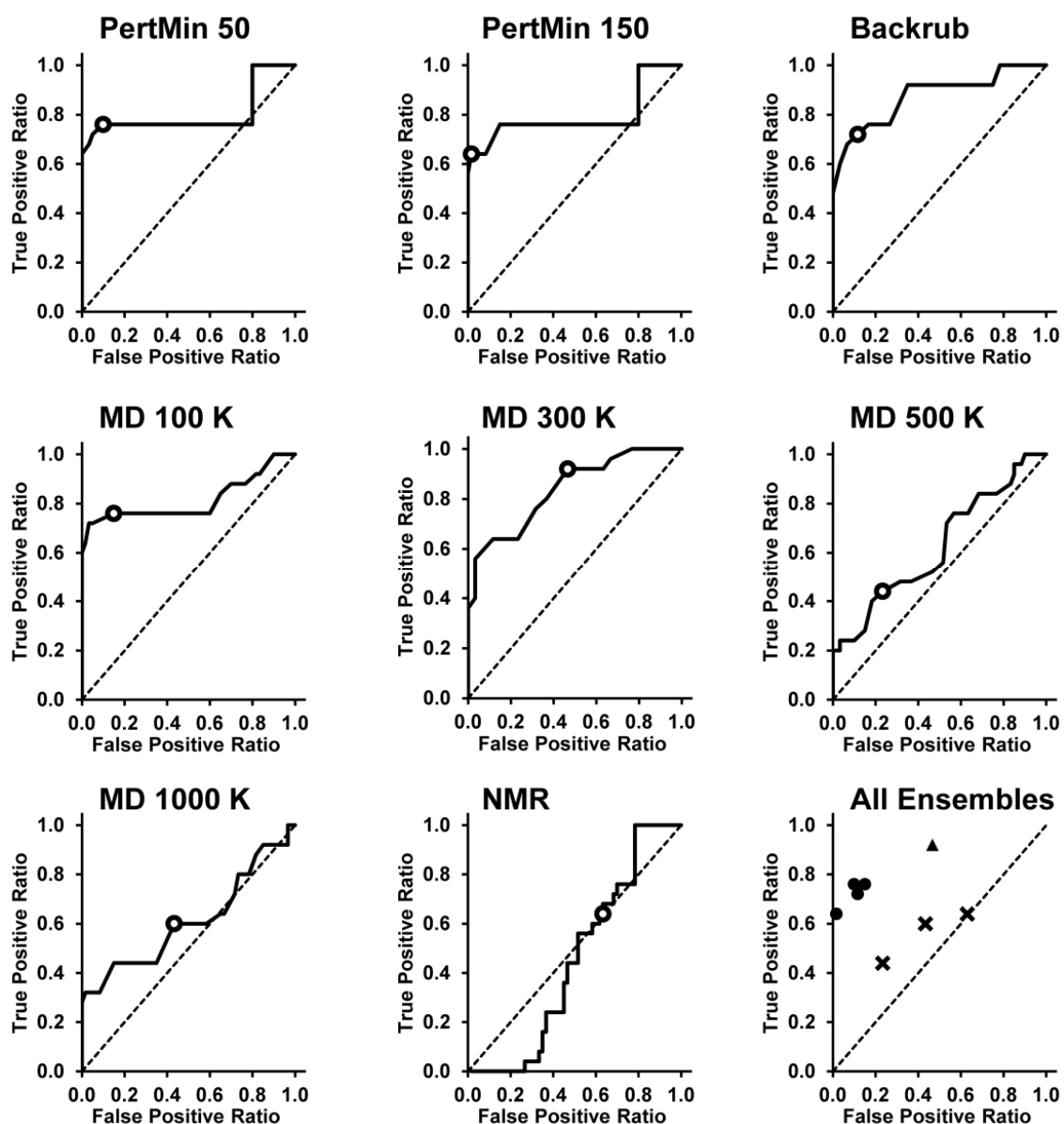


Figure 3.14. ROC curves for binning results from MSD with all ensembles. The true and false positive ratios obtained by using the WT fitness value as a cut-off are indicated by white circles. These values are also plotted on the all ensembles panel using circles, x-

marks, and a triangle to represent the on-target, off-target, and MD 300 K ensembles, respectively. The diagonal line represents random sequence binning.

Conversely, the off-target ensembles (MD 500 K, MD 1000 K, and NMR) have ratios that approach random sequence binning, represented by the diagonal line. Of note, the NMR ensemble is the only ensemble whose false positive ratio increases faster than the true positive ratio for the majority of cut-off values. ROC analysis also shows that for the PertMin ensembles, there is a broad range of cut-offs that do not result in improvements to the true positive ratio and instead only increase the false positive ratio. This is due to the fact that there are six true positive sequences bearing the L7F mutation (Figure 3.13, sequence 3, 6, 7, 13, 22, and 23) that do not score favourably on the PertMin ensembles. These sequences have average fitness values of -51 ± 2 and -53 ± 2 kcal·mol⁻¹ on the PertMin 50 and 150 ensembles, respectively, whereas their average fitness is lower (-64 ± 4 kcal·mol⁻¹) on the Backrub ensemble. Thus, increase to the true positive ratio is delayed for both the PertMin ensembles because the fitness values of these six sequences are clustered further from the WT cut-off [Table 3.5]. Interestingly, the MD 300 K ensemble is unique in that it includes a greater proportion of true positives with the use of a WT cut-off, however, this comes at the cost of a greater ratio of false positives, demonstrating why it is neither on- nor off-target. Finally, ROC analysis demonstrates that there exists no ideal cut-off value for sequence binning with any of the ensembles.

3.9. Correlation of Sequence Predictions

As a final assessment of the performance of the eight ensembles, we evaluated if they could correctly correlate *in silico* fitness values with *in vitro* stability results. As shown in figure 3.15, linear regressions between the Boltzmann weighted average fitness value predicted *in silico* and the C_m value reported for the 24 most stable mutant sequences of G β 1 [Allen *et al.*, 2010] demonstrate that MSD with none of the ensembles tested can provide a reliable correlation allowing for a predictive ranking of sequences by their stability.

A similar observation was reported by Allen *et al.* [Allen *et al.*, 2010], who showed that despite the predictive success of their cMD ensemble, the fitness of the folded state could not be used to rank sequences in a manner which reflects their experimentally measured stability. Of note, the correlation plots for the PertMin 50, PertMin 150, and MD 100 K ensembles appear to contain two clusters of data points. The cluster containing six data points with higher fitness values corresponds to the set of six true positive sequences containing the L7F mutation [Figure 3.13, sequence 3, 6, 7, 13, 22, and 23] which do not score favourably on these on-target ensembles.

To evaluate whether the lack of correlation between simulated and experimentally measured stability resulted from the use of an insufficient number of input states for MSD, we attempted to correlate the *in vitro* stability of the 24 most stable G β 1 mutant sequences with the Boltzmann weighted average fitness computed across a “mega ensemble” consisting of all 494 unique backbones derived from the eight ensembles tested in this study.

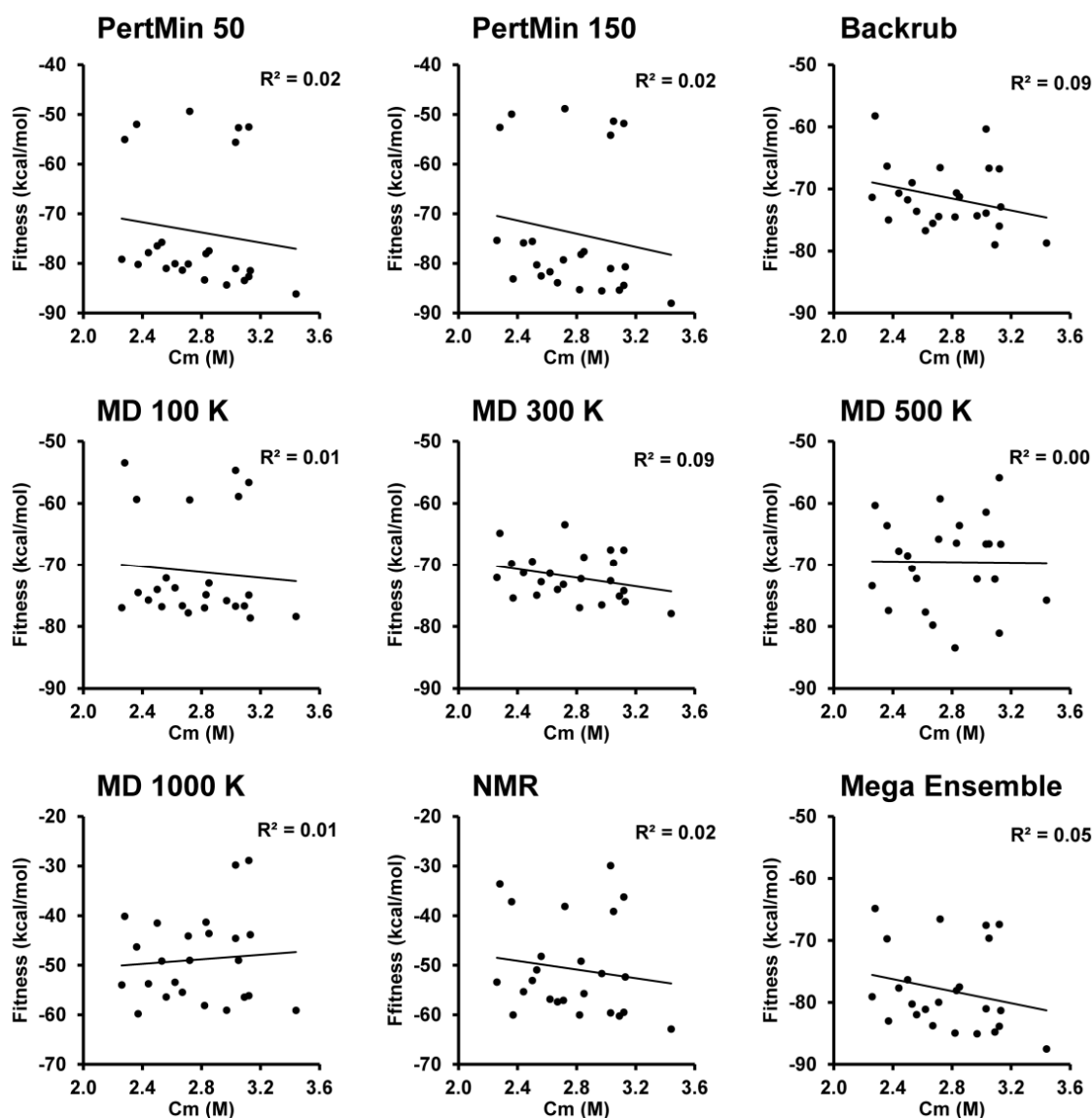


Figure 3.15. Correlation of MSD calculation results. Plots depict correlations between *in silico* fitness and experimentally measured stability. MSD fitness values for each ensemble are reported as a function of the C_m values for the 24 mutant sequences of protein G domain $\beta 1$ that are equally as or more stable than the WT sequence. The Mega Ensemble comprised all 494 unique ensemble member states from all eight ensembles.

The results, shown in Figure 3.15, demonstrate that no improvement to the correlation can be made by increasing the number of backbone templates used in a MSD simulation. This suggests that the inability of MSD to correlate *in silico* predictions

with *in vitro* stabilities is not the result of ensemble choice, or whether the ensemble chosen is on- or off-target, since the inclusion of additional states encompassing a wide range of properties failed to improve correlation. The observed lack of correlation may be due to inaccuracies in the energy function used to score the sequences [Kellogg *et al.*, 2011; Yin *et al.*, 2007a; Yin *et al.*, 2007b], to the choice of averaging function used to compute fitness, or to the fact that alternate competing states that the sequences may prefer to adopt were not considered.

3.10. The Effect of Ensemble Size on Multistate Design Predictions

The number of backbone templates included in an ensemble (i.e. ensemble size) can affect predictions made by MSD. For example, MSD performed using a 128-member 50 TN iteration PertMin ensemble results in the most favorable WT sequence fitness [Figure 3.16.A], the highest success rate [Figure 3.16.B], and the lowest number of false negatives [Figure 3.16.C]. By restricting the ensemble member space using structures from the original 128-member ensemble and repeating the calculation ($n = 30$) the affect of ensemble size on calculation outcome can be monitored.

As the ensemble size decreases, fitness of the WT sequence increases in energy, average success rate decreases, and the number of false negative sequences increases. Nevertheless, MSD using 64-member ensembles gives results similar to those obtained with the 128-member ensemble. While an ideal calculation would use an ensemble of sufficient size to model the design objective, MSD with a small 4-member ensemble is

still preferable to SSD with a single backbone template because it results in a higher success rate and fewer false negatives.

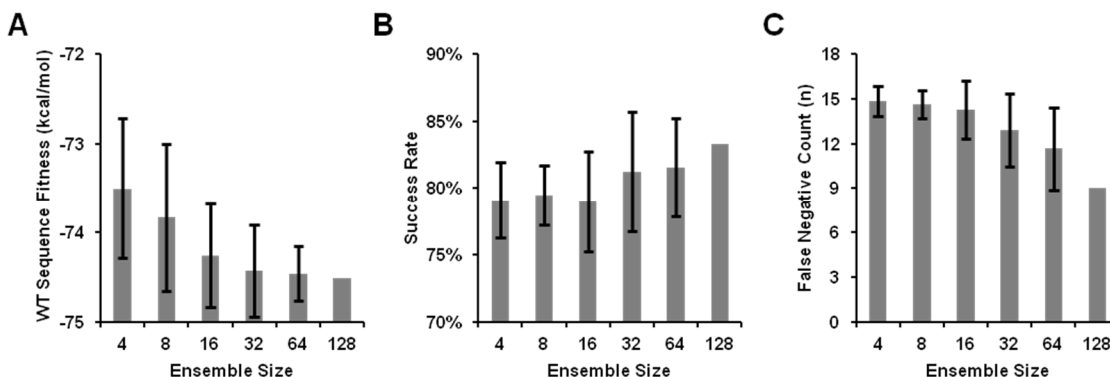


Figure 3.16. MSD with ensembles of various sizes. MSD was performed with a single 128-member G β 1 ensemble or with 30 ensembles containing random combinations of 4, 8, 16, 32, or 64 backbones extracted from the 128-member ensemble. The 128-member ensemble was prepared as described in section 3.2, with the exception that solvent water molecules were added to a depth of 3 Å. (A) The average WT sequence energy becomes more favorable as ensemble size increases. (B) The average success rate increases with ensemble size. (C) The average number of false negatives decreases with ensemble size. Error bars show the standard deviation of 30 independent MSD calculations.

3.11. Calculation Framework and Sequence Prediction Energetics

In SSD, sequences are ranked according to their score, which is the potential energy on a single backbone template. In MSA and MSD, sequences are ranked according to their fitness value, which is the Boltzmann weighted average energy across all backbone templates included in the PertMin ensemble. Because fitness is evaluated concertedly to sequence optimization in MSD but not in MSA, where fitness is instead

computed post-CPD [Figure 2.6], sequence ranking and fitness values obtained by these multistate approaches are not identical.

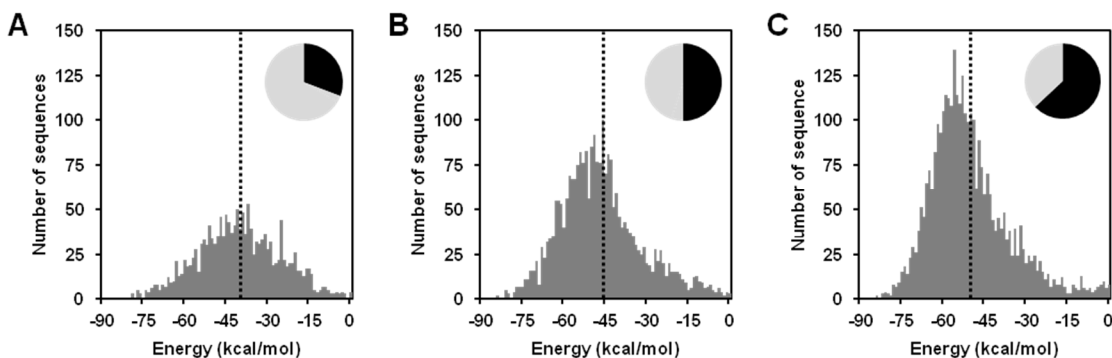


Figure 3.17. Calculation framework effect on predicted sequence energy. Distributions depict the number of sequences predicted by SSD (A), MSA (B), and MSD (C) with energy values grouped in incremental bins of 1 kcal·mol⁻¹. The average sequence energy is indicated by a dotted black line. The fraction of the pie charts in black (31%, 50%, and 63% for SSD, MSA, and MSD, respectively) corresponds to the percentage of the 5184 total possible sequences with predicted energy below 0 kcal·mol⁻¹.

Sequence energy distributions [Figure 3.17] show that more favorable energies are obtained for a larger number of sequences by the multistate approaches than by SSD, with average sequence energies of -40, -46, and -50 kcal·mol⁻¹ obtained by SSD, MSA, and MSD, respectively. The lower energies obtained by multistate approaches result from their ability to identify better backbone templates to score each sequence than the single template used in SSD [Howell *et al.*, 2014]. This is exemplified by the greater number of sequences that are scored with an energy lower than 0 kcal·mol⁻¹ by the multistate approaches (50% and 63% for MSA and MSD, respectively) than by SSD

(31%), highlighting how multistate approaches help to address the fixed backbone approximation.

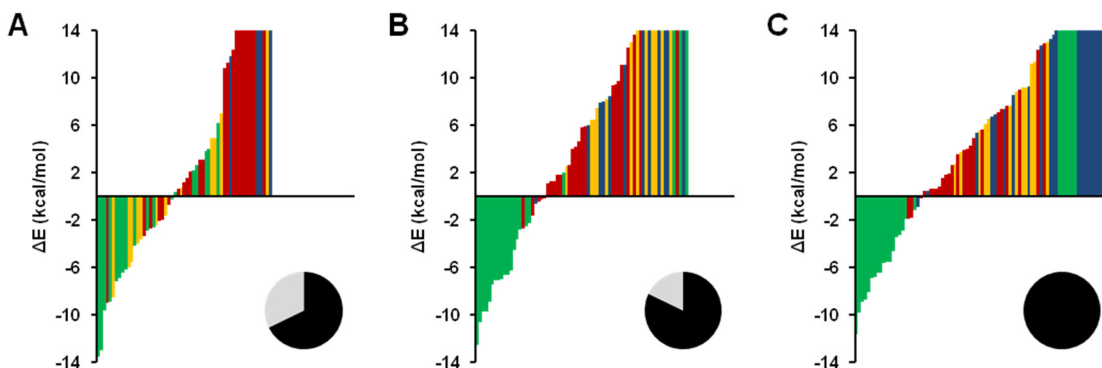


Figure 3.18. Calculation framework effect on binning predicted sequences. G β 1 training set sequences predicted by SSD (A), MSA (B), and MSD (C) calculations are binned according to their energy difference from the WT sequence. Sequences with lower energy than the WT ($\Delta E < 0$ kcal $\cdot$ mol $^{-1}$) are potential positive sequences while sequences with higher energy than the WT ($\Delta E > 0$ kcal $\cdot$ mol $^{-1}$) are potential negative sequences. Sequences are colored according to their experimental stability, with sequences having stability greater than or approximately equal to the WT in green (stabilizing), sequences having lower stability than the WT in yellow (destabilizing), sequences that do not fold in red (unfolded), and sequences postulated to adopt a non-native fold in blue (non-native). Positive ΔE values are capped at +14 kcal $\cdot$ mol $^{-1}$, even if the predicted energy difference is greater than this value. The fraction of the pie charts in black (68%, 82%, and 100% for SSD, MSA, and MSD, respectively) corresponds to the percentage of the 84 training set sequences with predicted energy below 0 kcal $\cdot$ mol $^{-1}$.

The prediction differences arising from choice of calculation framework can also be examined by performing the sequence binning procedure [Figure 3.18]. Sequence binning results shown on Figure 3.18 demonstrate that MSD is the only method that can score all 84 training set sequences below 0 kcal $\cdot$ mol $^{-1}$. In contrast, SSD and MSA score fewer of the training set sequences below 0 kcal $\cdot$ mol $^{-1}$ (68% and 82%, respectively). The

multistate methods correctly reject a higher number of the 60 true negatives [Figure 3.18, destabilizing, unfolded, and non-native sequences] and correctly accept a higher number of the true positives [Figure 3.18, stabilizing sequences], compared to SSD. As a result, fewer false negatives are predicted by MSA and MSD than by SSD, an expected result given that they help to address the fixed backbone approximation. Additionally, the success rates of multistate methods are greater than that of SSD, with the MSD success rate being the highest (88%). The number of false negative predictions made by MSD increases if the ensemble does not cover a sufficient structure space to score training set sequences, resulting in decreased average binning success rates [Figure 3.18]. These results further demonstrate the improved accuracy of predictions made by MSA and MSD calculations relative to SSD.

3.12. Calculation Framework and Sequence Prediction Diversity

The diversity of sequence space ranked favorably by CPD calculation is dependent on the framework employed (SSD, MSA, and MSD). To investigate the outcome of algorithm choice on sequence diversity, the top 100 sequences predicted by the various CPD methods was analyzed. To do so, the frequency of amino-acid residues found at each designed position as well as the number of identical sequences in the top 100 sequences predicted by each method was compared.

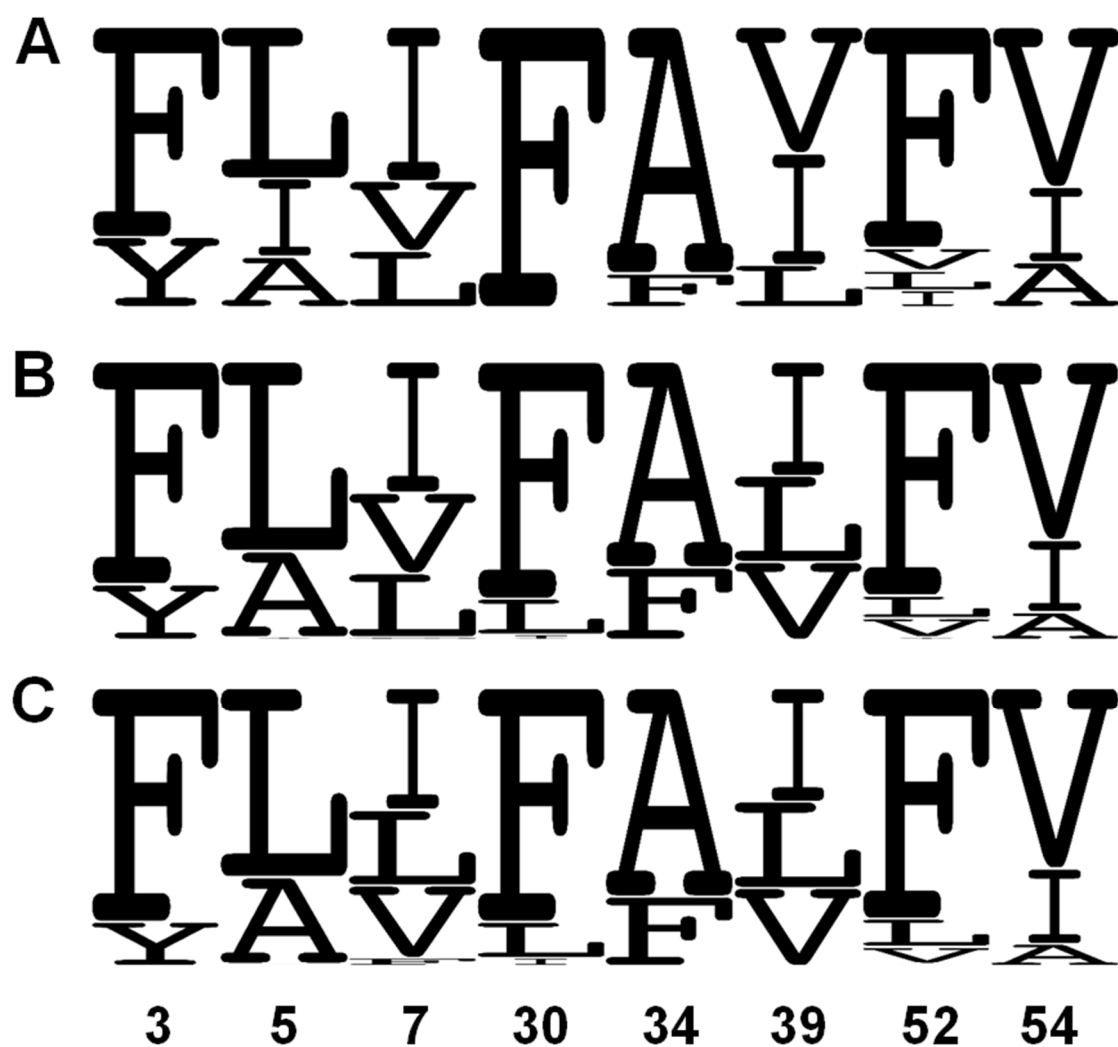


Figure 3.19. Calculation framework effect on predicted sequence diversity. Amino-acid diversity found at G β 1 designed positions in the top 100 ranked sequences predicted by various CPD methods. Sequence logos for SSD (A), MSA (B), and MSD (C), are shown with amino-acid substitution frequency proportional to letter height. Designed positions are indicated by numbers. Sequence logos were prepared using WebLogo 2.8.2 [Crooks *et al.*, 2004].

As shown on Figure 3.19, amino-acid diversity at each designed position of G β 1 is nearly identical in the top 100 sequences predicted by MSA and MSD. However,

sequence diversity obtained by SSD is significantly different, in particular at positions 5, 30, and 52. For example, many sequences predicted by SSD contain an isoleucine at positions 5 or 52, or do not include leucine or isoleucine substitutions at position 30, in contrast with sequences predicted by the multistate methods. The highly similar amino-acid diversity at each designed position obtained by the multistate methods suggests that their top 100 sequences are nearly-identical. To verify whether this is true, we compared the overlap in identical sequences contained in the top 100 sequences predicted by the various CPD methods. It was found that MSA and MSD share 89 of their top 100 ranked sequences, confirming that these methods predict nearly-identical top 100 sequences. In contrast, MSA and MSD share a much lower number of their top 100 sequences with SSD (54 or 51, respectively).

3.13. Properties of On-Target and Off-Target Ensembles

Our results demonstrate that the PertMin, Backrub, and MD 100 K ensembles are on-target models of G β 1 while the MD 500 K, MD 1000 K, and NMR ensembles represent off-target models. All on-target ensembles share a high level of structural similarity with each other and with the crystal structure, whereas off-target ensembles are unique and have low structural similarity to all other ensembles and to the crystal structure [Figure 3.11]. Based on these observations, we propose that the main characteristic that makes an ensemble on-target is its small deviation from crystal structure. This proposition is supported by the fact that all on-target ensembles have a backbone RMSD from crystal structure that is ≤ 0.7 Å. Notably, the MD 300 K ensemble,

which is neither on- nor off-target, deviates from the crystal structure by 0.8 Å, a RMSD value that may represent a threshold separating on- and off-target ensembles in the context of Gβ1. However, it is unlikely that this 0.8 Å RMSD value can be applied as a metric to predict *a priori* whether any ensemble will be on- or off-target in all design cases.

Nevertheless, when attempting to design sequences limited to a conservative search of the sequence space about the WT, for an ensemble prepared by any method to be on-target, it should deviate as little as possible from the input crystal structure. Failure to meet this condition will likely result in the generation of an off-target ensemble or in an ensemble that is neither on- nor off-target. Predictions made by MSD using such ensembles may increase the diversity of the enriched predicted sequence space but also risks the possibility that many of the sequences do not adopt the desired fold or target property. For example, although the NMR ensemble conserves the secondary structural elements of Gβ1, it deviates from the 1PGA crystal structure by 1.0 Å. This RMSD value is significantly higher than that of the on-target ensembles described herein [Figure 3.11]. The higher deviation from crystal structure of the NMR ensemble would explain why its use in MSD resulted in poor sequence enrichment [Table 3.2] and binning [Table 3.3]. These results are in agreement with those published by Allen *et al.* who reported that MSD with their NMR ensemble resulted predominantly in the prediction of Gβ1 sequences that are unfolded [Figure 3.9] [Allen *et al.*, 2010]. Our data also allows us to propose that low diversity is not an important characteristic of on-target ensembles. This proposition is supported by the fact that individual members of

the off-target NMR ensemble deviate from each other by 0.4 Å, a diversity value within the range of that of our on-target ensembles [Figure 3.11]. Finally, our results show that potential energy and RMSG are not important characteristics facilitating prediction as to whether an ensemble is on- or off-target, as evidenced by the fact that the Backrub ensemble is on-target even though it has the highest potential energy and RMSG. However, low potential energy is desirable as it improves sequence binning by providing a more stringent WT cut-off value that helps to reduce the number of false positives as observed through the use of the minimized ensembles [Table 3.5].

Regardless of whether the ensembles were on- or off-target, they all correctly rejected most of the Gβ1 sequences postulated to adopt a non-native fold [Figure 3.13, sequences 61-84]. However, these sequences were all incorrectly predicted to be stable by MSD with an uMD ensemble [Figure 3.9] that deviated from crystal structure by 0.84 Å [Allen *et al.*, 2010]. Our ensemble that is the most similar to the uMD ensemble is the MD 300 K ensemble, which has a near-identical RMSD from crystal structure (0.8 Å). Yet, the MD 300 K ensemble correctly rejected all but two of these 24 sequences. Although both ensembles were prepared using uMD simulations run at 300 K, the MD 300 K ensemble was prepared using explicit solvent under periodic boundary conditions while the uMD ensemble was prepared in vacuum. Thus, the simulation which produced the uMD ensemble was free to explore the potential energy surface outside the coordinate space of our MD 300 K ensemble which was restricted by the inclusion of explicit solvent.

In this way, conformational restrictions on the MD 300 K ensemble may have resulted in the rejection of sequences postulated to adopt a non-native fold. Interestingly, the MD 300 K ensemble is also the only ensemble that incorrectly accepted a majority (11/12) of the sequences known to be folded but less stable than WT [Figure 3.13, sequences 25-36]. As well, the MD 300 K ensemble correctly accepted 23 out of the 25 sequences that are equally as or more stable than WT [Figure 3.13, sequences 1-25] while also incorrectly accepting a majority of the unfolded (15/24) sequences [Figure 3.13, sequences 37-60]. The unique binning pattern of the MD 300 K ensemble illustrates why it is neither on- nor off-target, since it can be stabilized by sequences that are known to have stability greater than, equal to, or lower than the WT, as well as by many unfolded sequences. Thus, we propose that the MD 300 K ensemble may be a model of the molten globule state of G β 1 as it conserves native-like secondary structure content [Figure 3.10] but may have a protein interior that is loosely packed, allowing for destabilizing residues to be tolerated. Finally, the MD 1000 K and NMR ensembles, are the only ensembles that incorrectly accepted most (20/24 and 22/24, respectively) of the unfolded sequences [Figure 3.13, sequences 37-60]. Interestingly, while the MD 1000 K ensemble deviated the furthest from crystal structure, it was the NMR ensemble that performed the worst at sequence enrichment and binning. This observation is in agreement with results by Schneider *et al.* [Schneider et al., 2009] who demonstrated that NMR ensemble members are on average worse input templates for CPD than their crystal structure counterparts.

3.14. Improvements to Ensemble Based Multistate Design of Protein Stability

As a backbone ensemble generation method, PertMin allowed for the preparation of the on-target ensembles that performed the best in this MSD study. Indeed, PertMin 50 performed the best at sequence enrichment [Table 3.2] by providing the highest number of sequences displaying stability greater than or equal to WT and the lowest number of sequences that are less stable than WT or unfolded. Additionally, the PertMin 150 ensemble performed the best at sequence binning [Table 3.3], resulting in the highest success rate and the lowest number of false positives of all the ensembles tested. The success of the PertMin ensembles can be attributed to the fact that their members are generated by energy minimization, resulting in a high degree of similarity to the crystal structure and in low potential energy which are properties critical to decreasing the number of false negatives and false positives, respectively. This significant reduction in the number of false negatives and positives was also observed for MSD with the minimized Backrub and MD 100 K ensembles.

Since both the PertMin and the minimized Backrub ensembles were generated by coordinate perturbations that preserved similarity to the crystal structure followed by energy minimization, it is not surprising that their use in MSD provided similar results. This observation suggests that other perturbation-based strategies [Larson *et al.*, 2002], or the coupling of alternative conformational modelling strategies such as Backrub motions [Davis *et al.*, 2006] or Kinematic Loop Closure [Coutsias *et al.*, 2004] with PertMin, could be used to generate on-target ensembles provided that the

perturbations produce backbones similar to that of the crystal structure and that these perturbed structures are energy minimized, thereby ensuring a reduction in the number of false negatives and false positives, respectively. Compared to other methods, PertMin is advantageous because it is less computationally expensive than MD and because the protocol using coordinate perturbations is easier to implement than with Backrub. These advantages of PertMin, coupled with its ability to tune backbone characteristics without the need for artificial constraints, make it a practical alternative to other methods for the generation of on-target backbone ensembles for MSD.

In this study, we demonstrate that CPD can result in correct binning of protein sequences as stable or unstable with $\approx 90\%$ accuracy. To achieve this result, MSD with an on-target ensemble that deviates little from the crystal structure in combination with a cut-off value corresponding to the Boltzmann weighted average fitness of the WT was required. Application of these parameters in MSD may allow for similar results to be obtained when designing stable mutants of other proteins. Additionally, our results confirm that MSD with native backbone ensembles decreases the number of false negatives by improving the scoring of sequences, regardless of their stability, through the use of a collection of conformational states. On the other hand, this result is generally achieved in combination with an increased number of false positives as MSD also results in more favourable fitness values for unstable sequences, due to the use of a Boltzmann weighted average which diminishes score contributions to the fitness function from unfavourable backbones.

While none of the ensembles allowed for the correlation of predicted vs. experimentally measured stability when used in positive MSD, improvements to CPD force fields and energy functions [Kellogg *et al.*, 2011], use of backbone relaxation algorithms [Yin *et al.*, 2007a; Yin *et al.*, 2007b], and consideration of competing states (i.e. the aggregated or unfolded states) in negative design may result in improved correlation in the future.

3.15. Conclusions on the Study of Ensemble Generation Methodologies

In this chapter, a new backbone ensemble generation method termed PertMin was developed. PertMin is computationally inexpensive, easy to implement, and tunable to give desired ensemble characteristics. The algorithm allowed for the generation of native backbone ensembles that are on-target and that have low potential energy. Use of these ensembles as input structures in MSD resulted in improved accuracy of protein stability predictions compared to SSD with the crystal structure and MSD with other on-target ensembles. MSD with MD ensembles generated at high temperatures can be used to generate off-target ensembles that do not allow for the accurate prediction of stable protein sequences. Finally, in this chapter we demonstrated that MSD with an on-target ensemble closely resembling the crystal structure and with a cut-off value corresponding to the Boltzmann weighted average fitness of the WT sequence allowed for a $\approx 90\%$ correct binning of protein sequences as stable or unstable. This approach may represent a promising strategy to increase the success rate of CPD, which has traditionally relied on sequence enrichment requiring the testing of numerous

sequences in order to identify proteins exhibiting intended characteristics [Chica *et al.*, 2010; Röthlisberger *et al.*, 2008]. In the future, it will be interesting to see if MSD with backbone ensembles can also be used to more efficiently engineer proteins displaying desired properties other than improved stability, such as specific ligand binding or enhanced catalytic activity.

CHAPTER 4

Rotamer Bias Methodologies

CPD predictions are highly dependent on the structure of the input template used. However, it is unclear how small differences in template geometry translate to large differences in stability prediction accuracy. In this chapter, we explore how structural changes to the input template affect the outcome of stability predictions by CPD. To do this, alternate templates were prepared by *Rotamer Optimization followed by energy Minimization* (ROM) and used to recapitulate the stability of 84 protein G domain β 1 mutant sequences. In the ROM process, side chain rotamers for WT or mutant sequences are optimized on crystal or NMR structures prior to template minimization, resulting in alternate structures termed ROM templates. we show that use of ROM templates prepared from sequences known to be stable results predominantly in improved prediction accuracy compared to using the minimized crystal or NMR structures. Conversely, ROM templates prepared from sequences that are less stable than the WT reduce prediction accuracy by increasing the number of false positives. These observed changes in prediction outcomes are attributed to differences in side chain contacts made by rotamers in ROM templates. Finally, it's shown that ROM templates prepared from sequences that are unfolded or that adopt a non-native fold result in the selective enrichment of sequences that are also unfolded or that adopt a non-native fold, respectively. Our results demonstrate the existence of a rotamer bias

caused by the input template that can be harnessed to skew predictions towards sequences displaying desired characteristics.

Some of the language and figures in this chapter were adapted from a research article coauthored with Dr. Roberto A. Chica.

Davey JA and Chica RA (2015) Optimization of rotamers prior to template minimization improves stability predictions made by computational protein design. *Protein Science*, 24(4):545-560 [Davey and Chica, 2015].

4.1. Structure and Sequence Space Relationship

Traditionally, CPD calculations are performed using a SSD approach whereby sequences are evaluated in the context of a single fixed protein backbone template, which is typically a high-resolution crystal structure that may be energy-minimized to alleviate steric clashes present in the deposited coordinates. Despite numerous successes, the SSD methodology is prone to false negative predictions that result from the combined use of a single fixed backbone template and a discrete set of rigid rotamers [Davey and Chica, 2014]. This known artefact of SSD leads to the incorrect rejection of sequences that could have been otherwise tolerated using slightly different rotamer and backbone geometries during calculation [Choi *et al.*, 2009]. To address this artefact, a variety of computational strategies have been developed such as the use of softer repulsive potential energy terms [Dahiyat and Mayo, 1997; Grigoryan *et al.*, 2007; Kellogg *et al.*, 2011], iterative energy minimization [Borgo and Havranek, 2012; Wang *et al.*, 2005], continuous rotamer optimization [Gainza *et al.*, 2012], flexible backbone design [Murphy *et al.*, 2012; Ollikainen *et al.*, 2013; Smith and Kortemme, 2011], and MSD with native backbone ensembles [Davey and Chica, 2012]. Although these strategies reduce false negative predictions, the outcome of CPD calculations remains heavily dependent on the initial geometry of the input backbone template used [Allen *et al.*, 2010; Schneider *et al.*, 2009].

More recently, it was shown that the accuracy of protein stability predictions could be substantially improved by optimizing sequences on a set of WT and mutant protein crystal structure templates [Howell *et al.*, 2014]. This *backbone drift* approach

increased prediction accuracy by identifying the most favorable template to score each sequence allowing for a better correlation with its experimentally-measured stability. The examples described above clearly demonstrate that the choice of input template has a significant effect on the predictions made by SSD. However, it is still unclear how small differences in input template geometries translate to large differences in protein stability prediction accuracy.

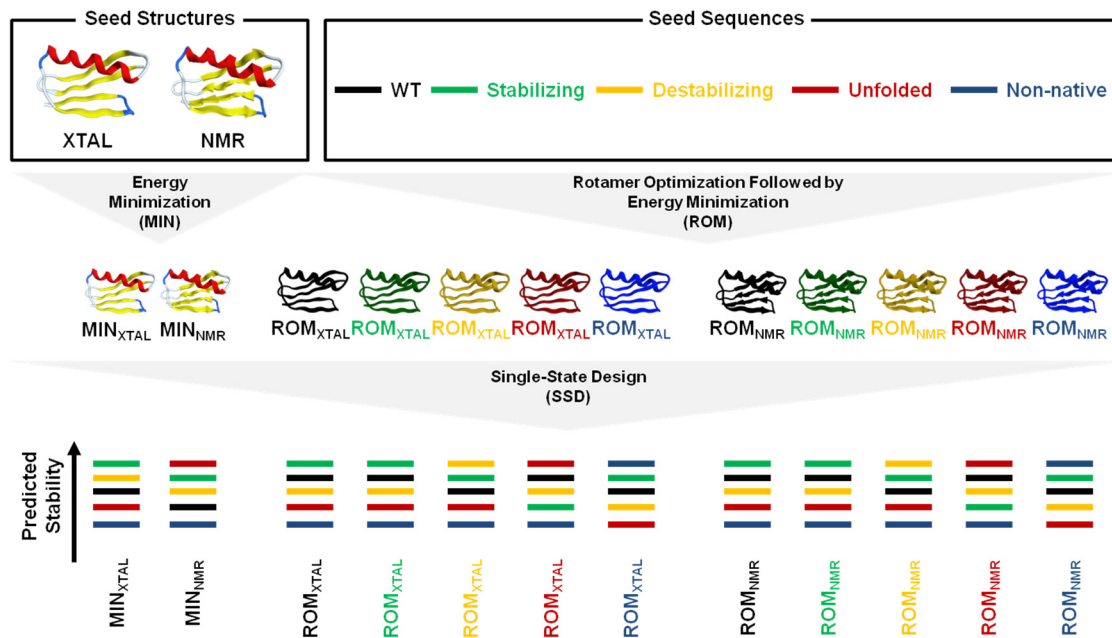


Figure 4.1. Rotamer optimization followed by energy minimization. Rotamer optimized and energy minimized (ROM) templates were prepared from stabilizing (green), destabilizing (yellow), unfolded (red), and non-native (blue) mutant G β 1 sequences as well as from the WT (black). To do so, rotamers for each seed sequence were optimized on the crystal (XTAL, PDB ID: 1PGA [Gallagher *et al.*, 1994]) or NMR (PDB ID: 2GB1 [Gronenborn *et al.*, 1991]) seed structures using SSD. The resulting templates were then energy minimized, completing the ROM procedure. Energy minimized crystal (MIN_{XTAL}) and NMR (MIN_{NMR}) templates were also prepared from the seed structures. MIN_{XTAL}, MIN_{NMR}, ROM_{XTAL}, and ROM_{NMR} templates were then used in SSD to predict the stability of each G β 1 sequence.

In this chapter, we explored how structural changes to the input template used in SSD affect the outcome of stability predictions. Because amino acid mutations can alter the protein backbone as well as the side chain conformation of neighbouring residues, we hypothesized that the use of templates prepared from sequences known to stabilize the protein fold would improve prediction accuracy because their structures would be more amenable to the identification of other stable sequences. To test this hypothesis, a method for template preparation called ROM for *Rotamer Optimization followed by energy Minimization* was developed. In this process, side chain rotamers for WT or mutant sequences, selected based on their relative stability or folding behaviour, are optimized by SSD on crystal or NMR structures and the resulting templates are subsequently energy-minimized, resulting in alternate structures that we call ROM templates. ROM templates were then used as inputs to SSD to predict the stability of mutant sequences [Figure 4.1]. Additionally, the use of ROM templates prepared from sequences known to be highly stable results predominantly in improved prediction accuracy compared to using the minimized crystal or NMR structures. Conversely, ROM templates prepared from sequences that are less stable than the WT reduce prediction accuracy by increasing the number of false positives. These observed changes in prediction outcomes are attributed to differences in side chain contacts made by rotamers in ROM templates. Finally, we have shown that ROM templates prepared from sequences known to be unfolded or postulated to adopt a non-native fold result in the selective enrichment of sequences that are also unfolded or that adopt a non-native fold, respectively. Our results demonstrate the existence of a rotamer bias caused by

the input template that can be harnessed to skew predictions towards sequences displaying desired characteristics by simply altering the template preparation procedure.

4.2. Sequence Binning with SSD Calculations using Rotamer Optimized and Energy Minimized Templates

In this study, the effect that rotamer identity and configuration in input templates have on the accuracy of protein stability predictions made by SSD were investigated. To do this, we prepared templates by optimizing rotamers of core residues on the crystal (PDB ID: 1PGA [Gallagher *et al.*, 1994]) and NMR (PDB ID: 2GB1 [Gronenborn *et al.*, 1991]) structures of G β 1 followed by energy minimization of each rotamer optimized template. This procedure, which we call ROM for *Rotamer Optimization followed by energy Minimization* [Figure 4.1], was used to generate different backbone templates derived from the WT or one of 84 G β 1 mutant sequences of known stability [Allen *et al.*, 2010]. These 84 mutant sequences, used in the previous chapter, are classified into one of four stability groups: (i) 24 sequences displaying stability greater than or approximately equal to the WT [Figure 4.2, stabilizing], (ii) 12 sequences of lower stability than the WT [Figure 4.2, destabilizing], (iii) 24 sequences that do not fold [Figure 4.2, unfolded], and (iv) 24 sequences postulated to adopt an alternate non-native fold [Figure 4.2, non-native] [Allen *et al.*, 2010]. Threading of the WT and 84 mutant sequences on the crystal and NMR structures followed by energy minimization

resulted in two sets of 85 ROM templates identified as ROM_{XTAL} and ROM_{NMR}, respectively [Figure 4.1].

Stability Group	Sequence ID	Cm (M)	3	5	7	20	26	30	34	39	43	52	54
Stabilizing	WT		Y	L	L	A	A	F	A	V	W	F	V
	1		F	-	I	-	-	-	-	I	-	-	-
	2		F	-	I	-	-	-	-	L	-	-	I
	3		F	-	F	-	-	-	-	L	-	-	I
	4		F	-	V	-	-	-	-	I	-	-	-
	5		F	-	I	-	-	-	-	-	-	-	-
	6		F	-	F	-	-	-	-	I	-	-	-
	7		F	-	F	-	-	-	-	I	-	-	I
	8		F	-	I	-	-	-	-	I	-	-	I
	9		F	-	I	-	-	-	-	L	-	-	-
	10		F	-	V	-	-	-	-	I	-	-	I
	11		F	-	V	-	-	-	-	L	-	-	I
	12		F	-	-	-	-	-	-	I	-	-	-
	13		F	-	F	-	-	-	-	L	-	-	-
	14		F	-	I	-	-	-	-	-	-	-	I
	15		F	-	-	-	-	-	-	-	-	-	-
	16		F	-	V	-	-	-	-	-	-	-	-
	17		F	-	V	-	-	-	-	L	-	-	-
	18		F	-	-	-	-	-	-	L	-	-	I
	19		F	-	V	-	-	-	-	-	-	-	I
	20		F	-	-	-	-	-	-	I	-	-	I
	21		F	-	-	-	-	-	-	L	-	-	-
	22		F	-	F	-	-	-	-	-	-	-	-
	23		F	-	F	-	-	-	-	-	-	-	I
	24		F	-	-	-	-	-	-	-	-	-	I
Destabilizing	25		F	I	I	-	-	-	-	I	-	-	-
	26		F	I	I	-	-	-	-	I	-	-	I
	27		F	I	V	-	-	-	-	I	-	-	-
	28		F	I	-	-	-	-	-	I	-	-	-
	29		F	I	V	-	-	-	-	I	-	-	I
	30		F	I	-	-	-	-	-	-	-	-	-
	31		F	I	I	-	-	-	-	-	-	-	I
	32		F	I	V	-	-	-	-	-	-	-	-
	33		F	-	I	-	-	-	-	-	-	I	-
	34		F	I	-	-	-	-	-	I	-	-	I
	35		F	I	-	-	-	-	-	-	-	-	I
	36		F	-	-	-	-	-	-	-	-	V	-
Unfolded	37		F	I	V	-	-	-	-	-	-	-	I
	38		F	-	I	-	-	-	-	-	-	L	-
	39		F	-	V	-	-	-	-	-	-	I	-
	40		F	-	-	-	-	-	-	-	-	L	-
	41		F	I	I	-	-	-	-	-	-	-	-
	42		F	-	I	-	-	-	-	I	-	I	-
	43		F	-	I	-	-	-	-	I	-	L	-
	44		F	-	I	-	-	-	-	I	-	V	-
	45		F	-	I	-	-	-	-	L	-	I	-
	46		F	-	I	-	-	-	-	L	-	L	-
	47		F	-	I	-	-	-	-	L	-	V	-
	48		F	-	I	-	-	-	-	-	-	V	-
Non-native	49		F	-	-	-	-	-	-	I	-	I	-
	50		F	-	-	-	-	-	-	I	-	L	-
	51		F	-	-	-	-	-	-	I	-	V	-
	52		F	-	-	-	-	-	-	L	-	I	-
	53		F	-	-	-	-	-	-	L	-	L	-
	54		F	-	-	-	-	-	-	L	-	V	-
	55		F	-	-	-	-	-	-	-	-	I	-
	56		F	-	V	-	-	-	-	I	-	I	-
	57		F	-	V	-	-	-	-	I	-	L	-
	58		F	-	V	-	-	-	-	I	-	V	-
	59		F	-	V	-	-	-	-	-	-	L	-
	60		F	-	V	-	-	-	-	-	-	V	-
	61		F	A	F	-	-	-	F	I	-	-	A
	62		F	A	-	-	-	-	F	I	-	-	A
	63		F	A	F	-	-	I	F	I	-	-	A
	64		F	A	F	-	-	L	F	I	-	-	A
	65		F	A	-	-	-	I	F	I	-	-	A
	66		F	A	-	-	-	L	F	I	-	-	A
	67		F	A	F	-	-	-	F	L	-	-	A
	68		F	A	-	-	-	-	F	L	-	-	A
	69		F	A	F	-	-	I	F	L	-	-	A
	70		F	A	F	-	-	L	F	L	-	-	A
	71		F	A	-	-	-	I	F	L	-	-	A
	72		F	A	-	-	-	L	F	L	-	-	A
	73		F	A	F	-	-	-	F	I	-	-	-
	74		F	A	-	-	-	-	F	I	-	-	-
	75		F	A	F	-	-	I	F	I	-	-	-
	76		F	A	F	-	-	L	F	I	-	-	-
	77		F	A	-	-	-	I	F	I	-	-	-
	78		F	A	-	-	-	L	F	I	-	-	-
	79		F	A	F	-	-	-	F	L	-	-	-
	80		F	A	-	-	-	-	F	L	-	-	-
	81		F	A	F	-	-	I	F	L	-	-	-
	82		F	A	F	-	-	L	F	L	-	-	-
	83		F	A	-	-	-	I	F	L	-	-	-
	84		F	A	-	-	-	L	F	L	-	-	-

Figure 4.2. Gβ1 benchmarking mutant sequences. Amino acid residues found at each design position in WT and mutant Gβ1 sequences (1-84) are listed. The 84 mutant

sequences are classified into one of four stability groups: 24 sequences displaying stability greater than or approximately equal to the WT (stabilizing, green), 12 sequences of lower stability than the WT (destabilizing, yellow), 24 sequences that do not fold (unfolded, red), and 24 sequences postulated to adopt an alternate non-native fold (non-native, blue).

Specifically, preparation of the crystal (PDB ID: 1PGA) [Gallagher *et al.*, 1994] and NMR-averaged (PDB ID: 2GB1) [Gronenborn *et al.*, 1991] structures of G β 1 were retrieved from the PDB. Water molecules were removed and hydrogen atoms were added to the protein using the Protonate3D [Labute, 2009] protocol available in the MOE software package [MOE, 2011]. Energy minimized crystal (MIN_{XTAL}) and NMR (MIN_{NMR}) templates were prepared by conjugate gradient minimization to a final gradient of 1 kcal·mol⁻¹·Å⁻¹ using the AMBER99 force field [Wang *et al.*, 2000] with the reaction-field model implemented in MOE [MOE, 2011]. ROM templates were prepared by optimization of side chain rotamers for WT and mutant G β 1 sequences using SSD as described below, followed by energy minimization as specified for the MIN_{XTAL} and MIN_{NMR} structures. ROM templates as well as minimized crystal and NMR structures [Figure 4.1, MIN_{XTAL} and MIN_{NMR}] were used as inputs for SSD in an attempt to recapitulate the known stability of the 84 G β 1 mutant sequences [Allen *et al.*, 2010]. Prediction accuracy of these calculations was evaluated by their ability to correctly bin mutant sequences as stable or unstable with respect to a cut-off value [Figure 3.12.B]. Since this value is design problem specific, the predicted energy of the WT sequence on each template was used as the cut-off because the WT sequence is known to be stable and ranked favorably in the context of the G β 1 fold [Davey and Chica, 2014]. In this sequence binning analysis, sequences of lower or higher scores relative to the WT were

predicted to be stable or unstable, respectively. For the purpose of our analysis, unstable sequences include the 60 undesirable mutant G β 1 sequences [Figure 4.2, destabilizing, unfolded, and non-native]. As shown in Table 1, binning analysis of SSD predictions using each ROM template, as well as the MIN_{XTAL} and MIN_{NMR} structures, classifies the 84 mutant G β 1 sequences into one of four categories: (i) true positives are stable sequences that are correctly scored below the WT, (ii) false negatives are stable sequences that are incorrectly scored above the WT, (iii) false positives are unstable sequences that are incorrectly scored below the WT, and (iv) true negatives are unstable sequences that are correctly scored above the WT. In addition, the success rate reports on the percentage of stable and unstable sequences that are correctly binned.

SSD was performed using the FASTER algorithm [Allen and Mayo, 2006] as implemented in PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012]. Side chain rotamers of core residues (positions 3, 5, 7, 20, 26, 30, 34, 39, 43, 52, and 54) were optimized on each fixed backbone template using the amino acid identities specified in Figure 4.2. The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 dihedral angles [Dunbrack and Cohen, 1997] was used to model side chain conformations. A four-term potential energy model consisting of a van der Waals term from the Dreiding II force field [Mayo *et al.*, 1990] with atomic radii scaled by 0.9, a direction sensitive hydrogen-bond term with well depth at 8.0 kcal·mol⁻¹ [Dahiyat and Mayo, 1997], an electrostatic energy term modelled using Coulomb's law with a distance dependent dielectric of 40, and a surface area-based solvation penalty term [Lazaridis and Karplus, 1999; Street and Mayo, 1998],

was used to score and rank sequences. Table 4.1 shows that SSD with the WT ROM_{XTAL} and WT ROM_{NMR} templates prepared using the WT sequence improves the scoring of the WT sequence resulting in a lower cut-off value compared to SSD with the MIN_{XTAL} and MIN_{NMR} structures, respectively. This result is to be expected given that core-residue rotamers were optimized prior to minimization in the WT ROM templates but not in the MIN_{XTAL} and MIN_{NMR} structures. The lower cut-off value increases stringency which improves the success rate through a reduction in the number of false positives. The increased stringency also has the consequence of decreasing the number of true positives in the case of the WT ROM_{NMR} template. The improved prediction accuracy obtained using WT ROM templates lead us to hypothesize that SSD with ROM templates prepared from alternate sequences known to stabilize the Gβ1 fold [Figure 4.2, stabilizing] might further improve the success rate by altering scoring of the WT and mutant Gβ1 sequences. As shown on Table 4.1, SSD with 18 of the 24 ROM_{XTAL} and 18 of the 24 ROM_{NMR} templates prepared from stabilizing sequences resulted in improved success rates relative to SSD with the WT ROM templates. SSD with all ROM templates prepared from stabilizing sequences also yielded a higher WT score than that obtained on the corresponding WT ROM template, a consequence of their structures no longer being perfectly complementary to the WT sequence. The higher WT score results in an equal or reduced number of false negatives and an increased number of false positives for SSD using a majority of ROM_{XTAL} (19/24) and ROM_{NMR} (22/24) templates prepared from stabilizing sequences.

Table 4.1. Sequence binning results for MIN and ROM prepared templates

	XTAL						NMR					
	Success rate (%)	True Positive	False Negative	False Positive	True Negative	Cutoff (kcal/mol)	Success rate (%)	True Positive	False Negative	False Positive	True Negative	Cutoff (kcal/mol)
MIN												
WT	79%	8	16	2	58	-71.5	71%	2	22	2	58	-63.6
ROM												
WT	81%	8	16	0	60	-78.2	73%	1	23	0	60	-80.9
ROM templates prepared from stabilizing sequences												
1	88%	16	8	2	58	-73.1	73%	9	15	8	52	-73.1
2	90%	18	6	2	58	-67.3	88%	18	6	4	56	-55.9
3	57%	19	5	31	29	-12.0	90%	22	2	6	54	-53.5
4	87%	18	6	5	55	-68.3	67%	9	15	13	47	-71.4
5	82%	10	14	1	59	-73.6	73%	7	17	6	54	-73.7
6	80%	15	9	8	52	-54.0	77%	12	12	7	53	-70.0
7	87%	19	5	6	54	-56.4	88%	19	5	5	55	-55.8
8	89%	18	6	3	57	-69.2	77%	14	10	9	51	-57.1
9	87%	14	10	1	59	-73.1	75%	9	15	6	54	-66.6
10	89%	18	6	3	57	-66.3	80%	14	10	7	53	-56.9
11	88%	17	7	3	57	-69.7	92%	18	6	1	59	-58.2
12	90%	16	8	0	60	-75.6	73%	7	17	6	54	-75.2
13	74%	23	1	21	39	-47.8	77%	8	16	3	57	-65.4
14	87%	15	9	2	58	-72.5	77%	14	10	9	51	-55.0
15	81%	8	16	0	60	-77.7	77%	5	19	0	60	-76.5
16	82%	10	14	1	59	-71.2	74%	7	17	5	55	-72.9
17	86%	14	10	2	58	-71.4	73%	9	15	8	52	-71.8
18	88%	18	6	4	56	-59.4	83%	18	6	8	52	-58.3
19	88%	16	8	2	58	-71.0	86%	14	10	2	58	-59.8
20	93%	18	6	0	60	-74.0	86%	15	9	3	57	-61.0
21	83%	10	14	0	60	-76.0	71%	7	17	7	53	-71.5
22	79%	9	15	3	57	-67.2	79%	6	18	0	60	-73.5
23	77%	10	14	5	55	-62.1	92%	20	4	3	57	-56.4
24	87%	13	11	0	60	-76.1	87%	15	9	2	58	-59.9
ROM templates prepared from destabilizing sequences												
25	73%	15	9	14	46	-68.2	67%	17	7	21	39	-51.1
26	67%	18	6	22	38	-48.5	52%	18	6	34	26	0.0 ^a
27	73%	18	6	17	43	-55.2	54%	18	6	33	27	-40.9
28	71%	14	10	14	46	-69.1	69%	14	10	16	44	-57.6
29	77%	18	6	13	47	-67.0	46%	18	6	39	21	-18.0
30	69%	9	15	11	49	-74.6	71%	15	9	15	45	-56.8
31	70%	13	11	14	46	-68.4	58%	13	11	24	36	-50.7
32	70%	11	13	12	48	-69.8	67%	9	15	13	47	-58.2
33	62%	9	15	17	43	-65.0	46%	3	21	24	36	0.0 ^a
34	76%	17	7	13	47	-71.4	76%	15	9	11	49	-57.7
35	73%	13	11	12	48	-73.1	75%	15	9	12	48	-57.2
36	64%	7	17	13	47	-66.7	57%	9	15	21	39	-31.5

^aCut-off values of 0 kcal·mol⁻¹ were assigned to ROM_{NMR} templates prepared from sequences 26 and 33 because SSD could not favorably score the WT sequence on these templates

Despite the increased number of false positives, the success rates of these calculations improved due to a disproportionate increase of true positives. This is best exemplified by SSD with the Sequence 3 ROM_{NMR} template, which resulted in 22 true positives and a success rate of 90%, a significant improvement in prediction accuracy compared to SSD with the minimized NMR structure, which resulted in only 2 true positives and a 71% success rate. This improved success rate results from the use of a template on which scoring of the WT sequence is worsened and scoring of stabilizing sequences is improved. To verify whether the use of templates on which the WT sequence is poorly scored is sufficient to improve binning success rate, we performed SSD using ROM templates prepared from the 12 destabilizing sequences [Figure 4.2].

As was observed for SSD with ROM templates prepared from stabilizing sequences, the WT cut-off value was increased relative to that obtained with the WT ROM templates, leading to a reduction of false negatives for most (11/12) or all of the ROM_{XTAL} and ROM_{NMR} templates, respectively. Despite a higher number of true positives for all but one of these calculations [Table 4.1, Sequence 36 ROM_{XTAL}], the success rate of SSD using ROM templates prepared from destabilizing sequences was decreased for all ROM_{XTAL} and 10 of the 12 ROM_{NMR} templates relative to results obtained for the WT ROM templates. This reduced prediction accuracy cannot be attributed to the stringency of the cut-off value, as the calculated WT scores fall mostly within the range of those obtained using ROM templates prepared from stabilizing sequences. Instead, the reduced success rates are due to the fact that unstable sequences are scored more favorably than the WT by SSD using ROM templates prepared from destabilizing

sequences. This scoring behaviour results in a large increase of false positives relative to SSD with the WT ROM templates. Clearly, predictions made by SSD are affected by the stability of the sequence used in ROM template preparation. This suggests the existence of a *rotamer bias*, which is a preference for certain sequences during calculation that is caused by the identity and configuration of rotamers found in the input template.

4.3. Sequence Enrichment with SSD Calculations using Rotamer Optimized and Energy Minimized Templates

To confirm the existence of rotamer bias, we examined the sequence enrichment profiles for SSD using all 170 ROM templates. Sequence enrichment is defined as the ability of SSD to identify stable protein sequences and score them favourably such that they represent a majority of the top-ranked sequences [Figure 3.12.A]. In this analysis, the top 24 sequences (excluding WT) were examined because the test set of 84 G β 1 mutants contains 24 sequences displaying stability greater than or approximately equal to the WT. Enrichment profiles [Figure 4.3] show the distribution of sequences found in the top 24 following SSD with all templates. Visual inspection of the enrichment profiles shows that SSD with a majority of ROM templates resulted in an enrichment of sequences from the same stability group as the sequence used to prepare the ROM template, supporting the existence of rotamer bias.

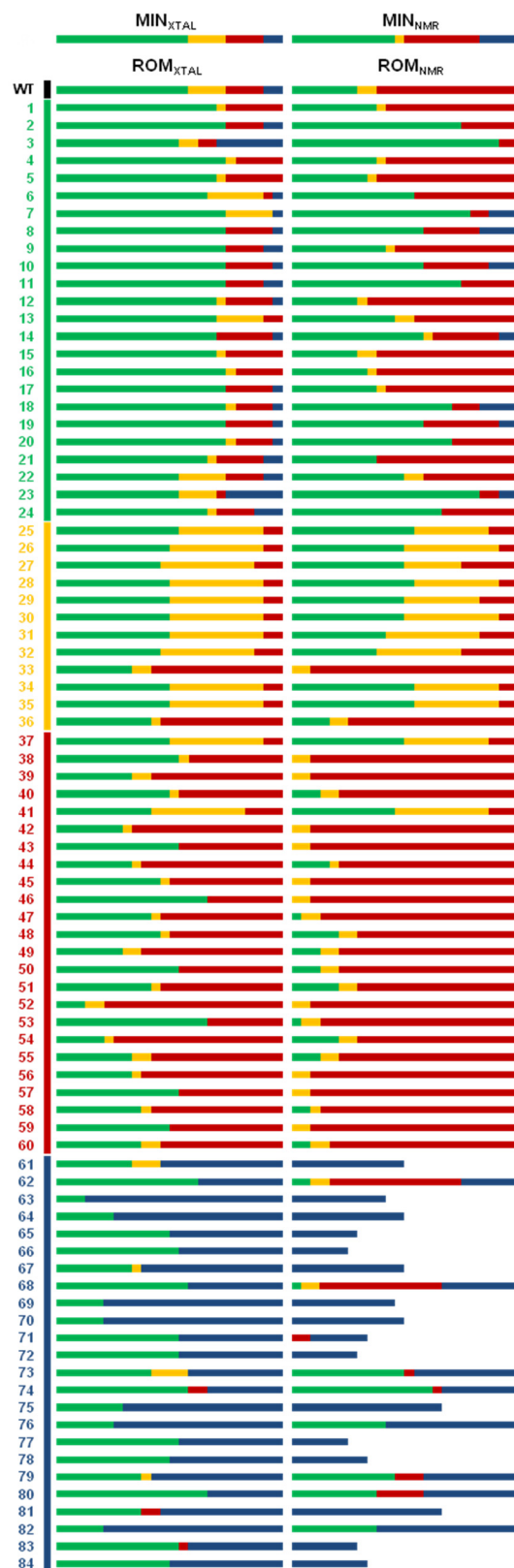


Figure 4.3. Enrichment profiles for SSD including ROM-minimized rotamers. The top 24 sequences (excluding WT) predicted by SSD using the energy minimized crystal (MIN_{XTAL})

and NMR (MIN_{NMR}) structures as well as rotamer optimized and energy minimized crystal (ROM_{XTAL}) and NMR (ROM_{NMR}) templates are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing, destabilizing, unfolded, and non-native sequences colored green, yellow, red, and blue, respectively. ROM templates were prepared from the WT (WT, black) and 84 mutant G β 1 sequences that are numbered and colored according to their stability group. Enrichment profiles of several ROM_{NMR} templates prepared from non-native sequences do not contain 24 sequences, and are shown as bars of reduced length.

Interestingly, ROM templates prepared from unfolded and non-native sequences bias SSD calculations in a way that results in enrichment of sequences from the same stability group even though it is unlikely that these mutants adopt the folded G β 1 structure in solution. Because sequence enrichment may result from the tendency of SSD to score more favorably rotamers optimized for compatibility with the template structure than rotamers that have not been optimized, we performed SSD calculations in which ROM-minimized rotamers were discarded and only unoptimized library rotamers were sampled.

As can be seen on Figure 4.4, similar enrichment profiles were obtained regardless of inclusion of ROM-minimized rotamers, with the exception of SSD calculations using ROM_{NMR} templates prepared from destabilizing sequences, which did not enrich for destabilizing sequences. These results demonstrate that sequence enrichment is not solely caused by preferential scoring of optimized rotamers.

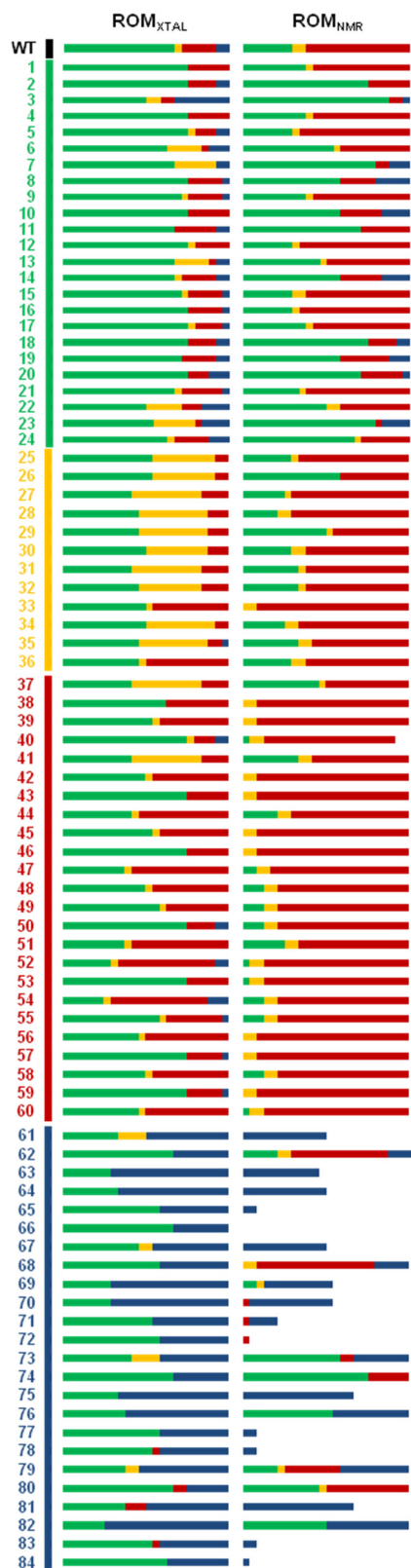


Figure 4.4. Enrichment profiles for SSD excluding ROM-minimized rotamers. The top 24 sequences (excluding WT) predicted by SSD using rotamer optimized and energy

minimized crystal (ROM_{XTAL}) and NMR (ROM_{NMR}) templates are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing, destabilizing, unfolded, and non-native sequences colored green, yellow, red, and blue, respectively. ROM templates were prepared from the WT (WT, black) and 84 mutant Gβ1 sequences that are numbered and colored according to their stability group. Enrichment profiles of several ROM_{NMR} templates prepared from non-native sequences do not contain 24 sequences, and are shown as bars of reduced length.

It is important to note that none of these SSD calculations resulted in an enrichment profile containing sequences from a single stability group, demonstrating that no template is ideal to score favorably all sequences for a given stability group. In addition, a majority of the enrichment profiles obtained from SSD calculations with ROM_{NMR} templates prepared using non-native sequences do not contain 24 sequences. This is due to the inability of SSD to score favorably (i.e. $< 0 \text{ kcal}\cdot\text{mol}^{-1}$) most sequences on these ROM_{NMR} templates, suggesting that their structures are incompatible with the sequences and rotamers used during calculation. Because SSD calculations including [Figure 4.3] or excluding [Figure 4.4] ROM-minimized rotamers yielded similar enrichment profiles, all subsequent calculations included these rotamers.

To evaluate whether observed differences in enrichment profiles between ROM templates from the four stability groups were significant, the average and standard deviation were calculated for the number of sequences from each stability group that were found in the top 24 across all ROM templates prepared from sequences of the same group. As shown on Table 4.2, SSD with ROM_{XTAL} templates prepared from sequences of each stability group results on average in a higher number of sequences from the same stability group in the top 24 ranked sequences compared to SSD with the

MIN_{XTAL} or with the WT ROM_{XTAL} template. On the other hand, SSD with ROM_{NMR} templates derived from stabilizing or unfolded sequences does not significantly increase the number of sequences from the same stability group found in the top 24 compared to the MIN_{NMR} or WT ROM_{NMR} template, respectively. This suggests that the crystal structure is more sensitive to rotamer bias than the NMR structure.

Table 4.2. Sequence enrichment results for MIN and ROM prepared templates

Template	Sequences in top 24			
	Stabilizing	Destabilizing	Unfolded	Non-native
MIN _{XTAL}	14	4	4	2
MIN _{NMR}	11	1	8	4
ROM_{XTAL} including ROM-minimized rotamers				
WT	14	4	4	2
Stabilizing	17 ± 2	2 ± 2	4 ± 2	2 ± 2
Destabilizing	11 ± 1	9 ± 3	4 ± 4	0
Unfolded	10 ± 3	2 ± 3	12 ± 4	0
Non-native	10 ± 4	0	0	13 ± 4
ROM_{NMR} including ROM-minimized rotamers				
WT	7	2	15	0
Stabilizing	13 ± 5	1 ± 1	10 ± 5	1 ± 1
Destabilizing	10 ± 4	8 ± 3	6 ± 7	0
Unfolded	3 ± 3	3 ± 2	19 ± 5	0
Non-native	3 ± 5	0	2 ± 4	10 ± 3
ROM_{XTAL} excluding ROM-minimized rotamers				
WT	16	1	5	2
Stabilizing	17 ± 2	2 ± 2	4 ± 2	2 ± 2
Destabilizing	11 ± 1	8 ± 3	4 ± 3	0
Unfolded	13 ± 4	1 ± 3	9 ± 4	1 ± 1
Non-native	11 ± 3	1 ± 1	0	12 ± 3
ROM_{NMR} excluding ROM-minimized rotamers				
WT	7	2	15	0
Stabilizing	13 ± 5	1 ± 1	9 ± 5	1 ± 2
Destabilizing	7 ± 3	1 ± 1	15 ± 3	0
Unfolded	2 ± 3	2 ± 0	20 ± 3	0
Non-native	3 ± 6	0	3 ± 5	7 ± 5

Finally, it is important to note that SSD with the MIN_{XTAL} template results in a majority of stabilizing sequences in the top 24 whereas SSD with the MIN_{NMR} template does not. This result is in agreement with Keating and coworkers [Schneider *et al.*, 2009] who showed that NMR structures are on average less useful templates for CPD than crystal structures. Nevertheless, our results show that applying the ROM procedure on NMR structures can lead to substantial improvements in stability prediction accuracy by increasing the number of true positives and decreasing the number of false negatives [Table 4.1].

4.4. Sequence and Structural Determinants of Rotamer Bias

The sequence binning and enrichment results described above demonstrate the existence of rotamer bias in CPD calculations and confirm observations made by others^{2,18,19} that the outcome of SSD is highly dependent on the input template used. However, our results do not explain which template property is responsible for the observed effects. Because fixed backbone methods tend to predict sequences with high identity to the WT sequence [Murphy *et al.*, 2012], we first investigated whether the observed rotamer bias could be attributed to similarity between sequences used to prepare ROM templates and other G β 1 sequences. Sequence similarity between pairs of sequences was computed as the BLOSUM62 [Henikoff and Henikoff, 1992] score for the 8 designed positions that varied in identity during calculation [Figure 4.2, residues 3, 5, 7, 30, 34, 39, 52, and 54]. In addition, we examined identity between the sequence used

to generate each ROM template and other G β 1 sequences by calculating the number of identical residues at the 8 designed positions between each sequence pair. The assessment of both sequence similarity and identity allowed us to perform a pair-wise comparison of the G β 1 sequences in a manner that is independent of the structural features of the templates.

Table 4.3. Similarity and identity of G β 1 benchmark sequences

Metric	WT	Stabilizing	Destabilizing	Unfolded	Non-native
BLOSUM62 Score					
Stabilizing	31 $\pm$ 2	34 $\pm$ 2	32 $\pm$ 3	28 $\pm$ 2	17 $\pm$ 4
Destabilizing	30 $\pm$ 2	32 $\pm$ 3	33 $\pm$ 4	28 $\pm$ 3	15 $\pm$ 8
Unfolded	26 $\pm$ 2	28 $\pm$ 2	28 $\pm$ 3	31 $\pm$ 3	11 $\pm$ 4
Non-native	14 $\pm$ 4	17 $\pm$ 4	15 $\pm$ 5	11 $\pm$ 4	31 $\pm$ 4
Sequence Identity					
Stabilizing	5 $\pm$ 1	6 $\pm$ 1	5 $\pm$ 1	5 $\pm$ 1	3 $\pm$ 1
Destabilizing	5 $\pm$ 1	5 $\pm$ 1	6 $\pm$ 1	5 $\pm$ 1	3 $\pm$ 1
Unfolded	5 $\pm$ 1	5 $\pm$ 1	5 $\pm$ 1	6 $\pm$ 1	2 $\pm$ 1
Non-native	2 $\pm$ 1	3 $\pm$ 1	3 $\pm$ 1	2 $\pm$ 1	6 $\pm$ 1

Similarity and identity relationships between sequences, reported as the average BLOSUM62 score or the average number of identical designed residues between sequences belonging to each stability group, are shown on Table 4.3. As expected, sequences from the same stability group have high similarity, represented by BLOSUM62 scores of approximately 30, and strong identity, represented by a high number of identical designed residues (6 $\pm$ 1 out of a possible 8). Interestingly, stabilizing, destabilizing, and unfolded sequences are also highly similar with each other and with the WT sequence, with average BLOSUM62 scores in the 26-32 range. This is not the case for non-native sequences which, although similar to each other, are considerably less similar to sequences from other stability groups and to the WT, with

BLOSUM62 scores lower than 20. Identical trends are obtained when comparing sequence identity.

A

	ROM _{XTAL}	XTAL	NMR	Stabilizing	Destabilizing	Unfolded	Non-native
Stabilizing		0.47 ± 0.02	1.16 ± 0.01	0.14 ± 0.04	0.14 ± 0.04	0.16 ± 0.04	0.20 ± 0.03
Destabilizing		0.47 ± 0.01	1.15 ± 0.01	0.14 ± 0.04	0.10 ± 0.03	0.13 ± 0.04	0.20 ± 0.04
Unfolded		0.48 ± 0.02	1.16 ± 0.01	0.16 ± 0.04	0.13 ± 0.04	0.12 ± 0.05	0.23 ± 0.03
Non-native		0.47 ± 0.02	1.16 ± 0.02	0.20 ± 0.03	0.20 ± 0.04	0.23 ± 0.03	0.17 ± 0.05

B

	ROM _{NMR}	XTAL	NMR	Stabilizing	Destabilizing	Unfolded	Non-native
Stabilizing		1.01 ± 0.02	0.74 ± 0.03	0.24 ± 0.08	0.26 ± 0.06	0.29 ± 0.08	0.31 ± 0.06
Destabilizing		1.00 ± 0.02	0.73 ± 0.02	0.26 ± 0.06	0.20 ± 0.10	0.30 ± 0.09	0.31 ± 0.06
Unfolded		0.96 ± 0.02	0.72 ± 0.03	0.29 ± 0.08	0.30 ± 0.09	0.22 ± 0.10	0.34 ± 0.06
Non-native		1.02 ± 0.03	0.72 ± 0.02	0.31 ± 0.06	0.31 ± 0.06	0.34 ± 0.06	0.34 ± 0.06

Figure 4.5. ROM prepared backbone template RMSD comparisons. Backbone RMSD (N-C_α-C=O) between pairs of ROM_{XTAL} (A) and ROM_{NMR} (B) templates are reported as the average and standard deviation for ROM templates grouped according to the stability of their seed sequence (stabilizing, destabilizing, unfolded, and non-native). Average

backbone RMSD values are colored blue, green, yellow, orange, or red if they fall within the 0.10-0.19 Å, 0.20-0.29 Å, 0.30-0.39 Å, 0.40-0.49 Å, or > 0.5 Å ranges, respectively. The backbones of ROM templates are also compared to that of the crystal (XTAL) and NMR seed structures.

These results demonstrate that, with the exception of non-native sequences, sequence similarity and identity alone cannot account for the observed differences in sequence binning and enrichment obtained by SSD using ROM templates, suggesting that rotamer bias arises from structural differences in input templates.

We next investigated how structural changes arising from the ROM procedure were responsible for the observed prediction differences. To do so, the backbone RMSD between pairs of ROM templates was calculated and the average for sequences belonging to each stability group was examined. As shown on Figure 4.5, ROM_{XTAL} and ROM_{NMR} templates are more structurally similar to their parent crystal or NMR structure, respectively. This result is expected given that the energy minimization step in the ROM procedure does not allow for exploration of the potential energy surface and only solves the location of the nearest local minimum. In addition, all ROM_{XTAL} or ROM_{NMR} templates deviate similarly from either the crystal or NMR structures regardless of which stability group they belong to. This result confirms that the rotamer optimization step of the ROM procedure does not lead to clashes in the protein core that would need to be alleviated by significant structural rearrangements during energy minimization.

The diversity of ROM templates, defined as the average pairwise RMSD between templates belonging to the same stability group, does not significantly vary between

stability groups and is lower than RMSD to the crystal or NMR structure from which they are derived. This result indicates that sequence differences in ROM templates belonging to the same stability group only result in subtle backbone structural changes highlighting the significance of small structural changes and their affect on the evaluation of sequence stabilities. Additionally, a pairwise comparison of stability groups shows that backbones belonging to ROM templates prepared from non-native sequences are the least similar to the other templates, consistent with observed sequence similarity and identity differences. However, backbone deviations are small across all stability groups, suggesting that they are not the main cause of the different enrichment profiles obtained by SSD. Interestingly, the average backbone RMSD values obtained for ROM_{NMR} templates [Figure 4.5.B] are higher than those of the ROM_{XTAL} templates [Figure 4.5.A], suggesting that rotamer configurations obtained for each sequence following the rotamer optimization step of the ROM procedure resulted in greater side chain orientation variability on the NMR structure than on the crystal structure.

To investigate side chain orientation variability, we examined the distribution of χ_1 dihedral angles (N-C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$) for rotamers found at each designed position in the templates used in this study [Figure 4.6]. In the case of the input crystal and NMR structures, energy minimization relaxes the χ_1 dihedrals without altering the general side chain conformation (e.g. *trans*, *gauche*⁺ and *gauche*⁻), as expected. However, minimization of the NMR structure resulted in a greater change to the χ_1 angle for all but two of the designed positions (39 and 54) compared to the crystal structure,

supporting the observation that NMR structures are less rotameric than crystal structures [Schneider *et al.*, 2009]. In the crystal and NMR structures, side chains at designed positions occupy one of three conformations, with residue 54 being in the *gauche*⁺ conformation, residues 5 and 39 being in the *trans* conformation, and residues 3, 7, 30, and 52 being in the *gauche*⁻ conformation.

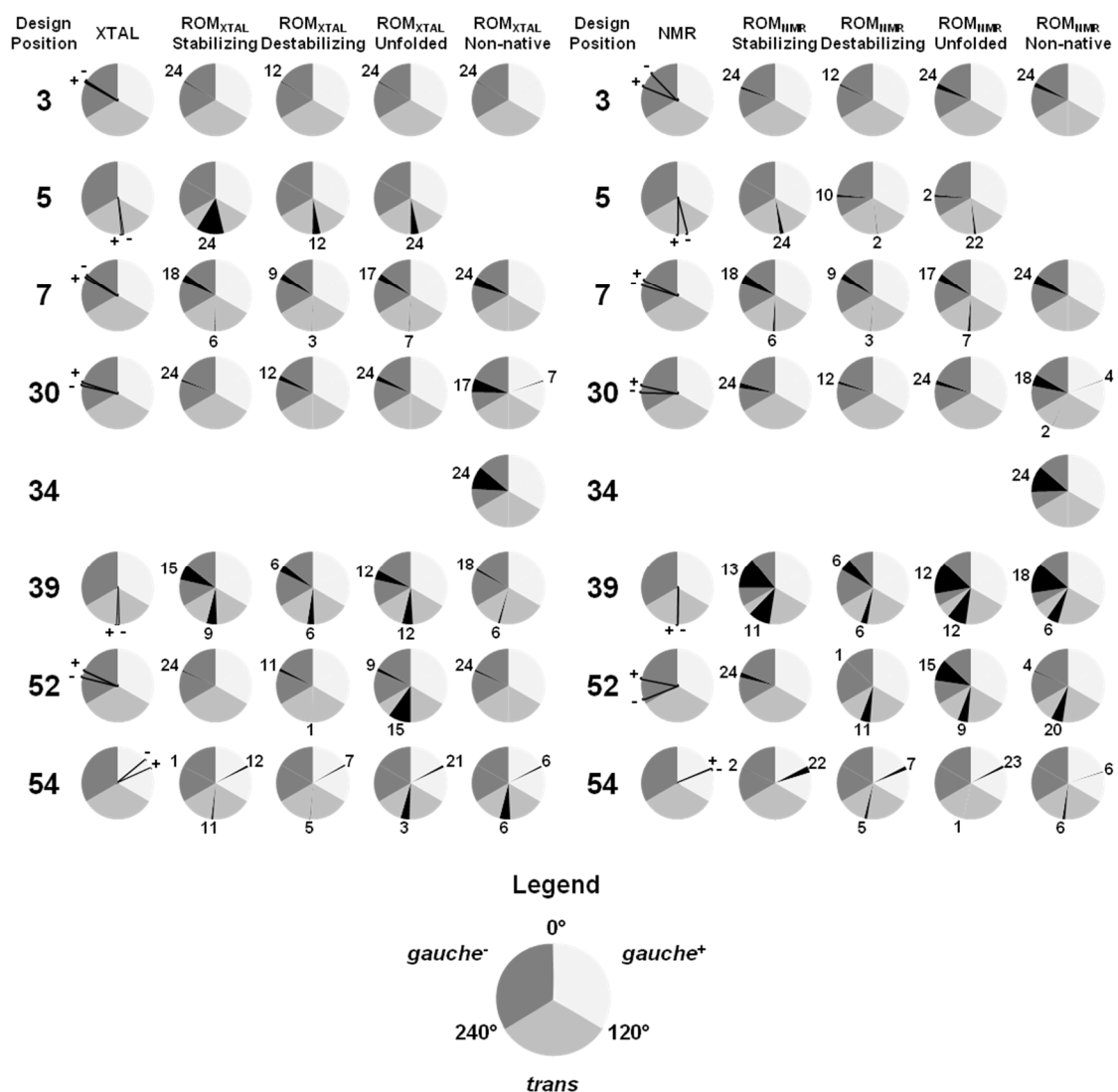


Figure 4.6. ROM prepared template side chain variability. χ_1 dihedral angles for rotamers found at each design position in the crystal (XTAL) and NMR structures prior to (-) or following energy minimization (+) are shown as black lines. χ_1 dihedral angles for rotamers found at each design position in ROM templates grouped by the stability of

their seed sequence are shown as sections of pie charts colored in black. These sections represent the range of values that the χ_1 dihedral can adopt in ROM templates. Dihedral values are binned into one of three general conformations: *gauche*⁺ (light grey), *trans* (medium grey), and *gauche*⁻ (dark grey). The number of ROM templates sharing the same general rotamer conformation at each design position is indicated next to each section.

These side chain conformations found at each designed position in the crystal and NMR structures are also found in most ROM templates, independently of stability group. However, additional side chain conformations not present in the crystal and NMR structures are also found in ROM templates. For example, both ROM_{XTAL} and ROM_{NMR} templates prepared with stabilizing sequences have two possible side chain conformations at designed position 7, with side chains of 18 templates occupying the *gauche*⁻ conformation present on the crystal and NMR structures and side chains of 6 templates occupying an alternate *trans* conformation. The *gauche*⁻ conformation is found on ROM templates containing either the WT leucine or mutant valine and isoleucine residues at position 7 whereas the *trans* conformation arises from the phenylalanine mutation. This example illustrates how amino acid identity at each designed position can lead to significant side chain conformational changes.

Interestingly, side chains of ROM_{NMR} templates have greater conformational variability than those of ROM_{XTAL} templates. For example, residues at designed position 5 in ROM_{NMR} templates prepared from destabilizing and unfolded sequences can adopt an additional conformation (*gauche*⁻) than equivalent residues in their ROM_{XTAL} counterparts. In addition, residues at designed positions 30 and 52 found in ROM_{NMR} templates prepared from non-native sequences can also adopt the *trans* conformation,

which is not found at these positions in ROM_{XTAL} templates. In the case of ROM_{XTAL} templates, only those prepared from stabilizing sequences contain an additional side chain conformation that is not present in the corresponding ROM_{NMR} templates [Figure 4.6, *trans* conformation at position 54]. The fact that ROM_{NMR} templates can accommodate a greater number of side chain conformations than ROM_{XTAL} templates may explain why they are less sensitive to rotamer bias, since their core cavity can tolerate a greater number of different side chain conformations. On the other hand, ROM_{XTAL} templates accommodate fewer side chain conformations but are more sensitive to rotamer bias, suggesting that small changes to rotamer configurations in these templates result in substantial changes to side chain contacts made by rotamers.

To investigate the nature of side chain contacts made by rotamers in ROM templates, we computed the van der Waals interaction energy between residues at designed positions and all other residues for each ROM template. The van der Waals energies were modelled using a Lennard-Jones 12-6 potential energy curve with AMBER99 parameters [Wang *et al.*, 2000]. The van der Waals energies were then averaged by stability group and by residue identity within each stability group, since van der Waals energies are highly dependent on the type and number of side chain atoms.

As shown on Figure 4.7, the contact map for ROM_{XTAL} templates demonstrates a clear dependence between average interaction energies and residue identity across templates from each stability group. For example, the interactions that leucine at designed position 7 makes with residues at positions 12 to 16 vary depending on the stability group of its corresponding ROM_{XTAL} template. Yet, leucine at designed position

7 has a similar χ_1 dihedral angle and *gauche* conformation in all ROM_{XTAL} templates [Figure 4.6], confirming that small changes in structure can result in substantial differences in side chain contacts. Thus, it is likely that the rotamer bias observed when performing SSD with ROM templates [Tables 4.1 and 4.2, and Figures 4.3 and 4.4] results from differences in side chain contacts.

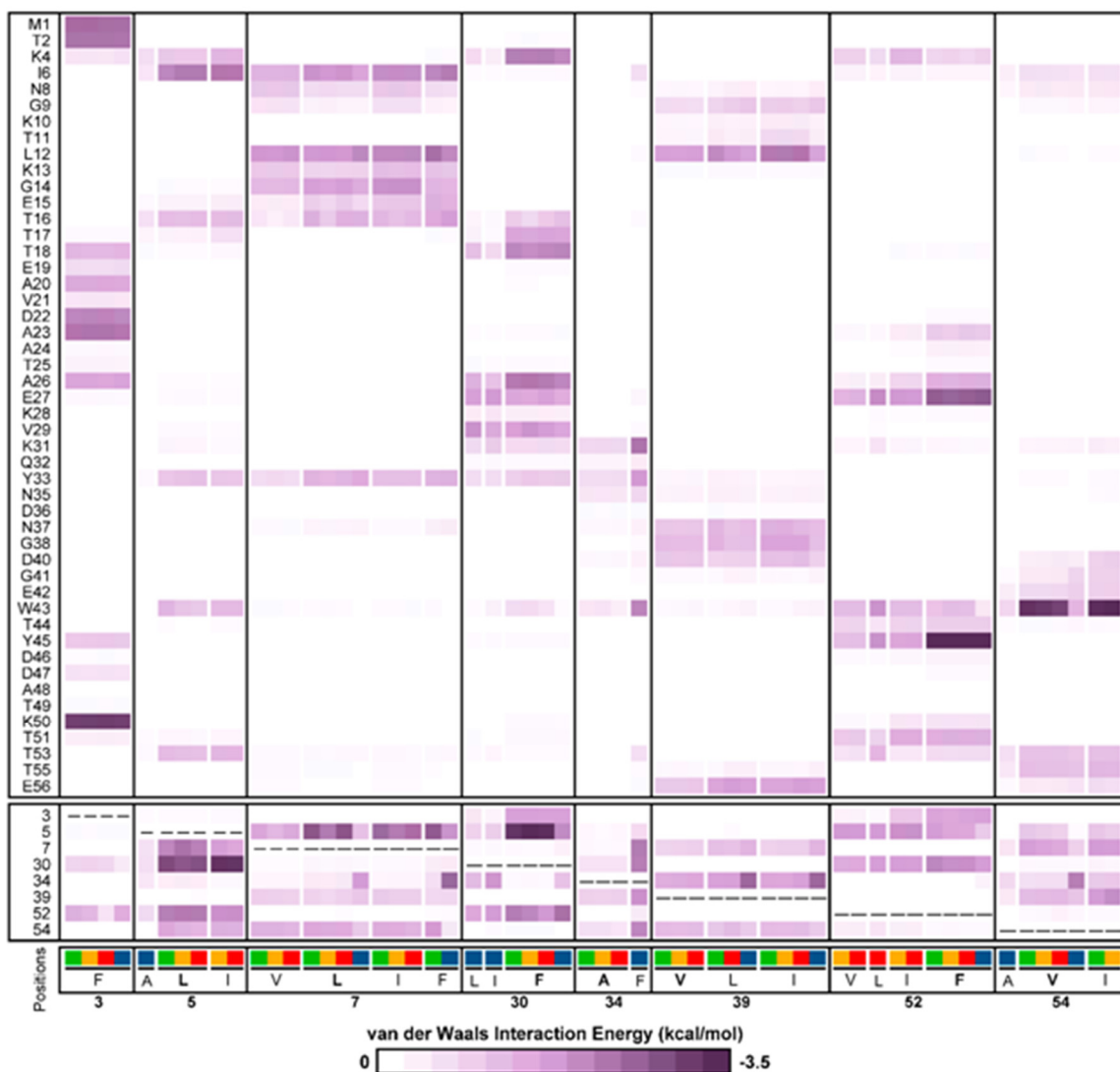


Figure 4.7. Energy contact map for ROM_{XTAL} templates. van der Waals interaction energies between residues at designed positions and all other residues in ROM_{XTAL} templates are averaged by stability group and by residue identity within each stability group. Residues found at each designed position in ROM_{XTAL} templates prepared from stabilizing, destabilizing, unfolded, and non-native sequences are colored green, yellow,

red, and blue, respectively, with the WT residue indicated in bold. Designed position residues are boxed separately from residues whose identity does not vary between templates. Interaction energies are colored according to their strength ranging from 0 kcal·mol⁻¹ (white, no interaction) to -3.5 kcal·mol⁻¹ (dark purple, strong favorable interaction).

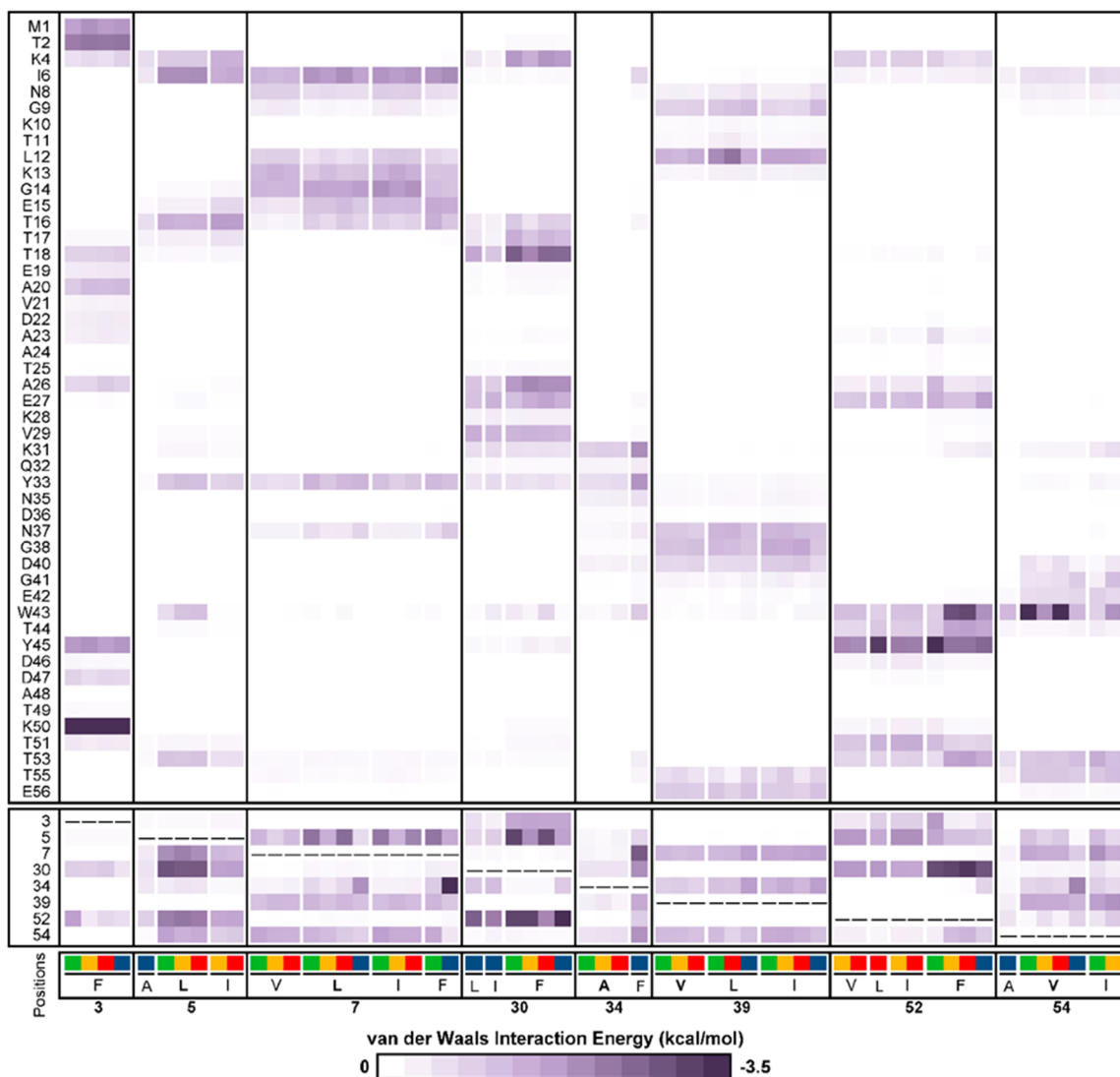


Figure 4.8. Energy contact map for ROM_{NMR} templates. van der Waals interaction energies between residues at designed positions and all other residues in ROM_{NMR} templates are averaged by stability group and by residue identity within each stability group. Residues found at each designed position in ROM_{NMR} templates prepared from stabilizing, destabilizing, unfolded, and non-native sequences are colored green, yellow, red, and blue, respectively, with the WT residue indicated in bold. Designed position residues are boxed separately from residues whose identity does not vary between templates. Interaction energies are colored according to their strength ranging from 0

kcal·mol⁻¹ (white, no interaction) to -3.5 kcal·mol⁻¹ (dark purple, strong favorable interaction).

While the distribution of side chain contacts is similar for both ROM_{XTAL} and ROM_{NMR} [Figure 4.8] templates, the strengths of interactions can vary significantly. For example, phenylalanine at designed position 52 interacts most strongly with tyrosine 45 in ROM_{XTAL} templates from all stability groups, but the same substitution interacts more strongly with tryptophan 43 for ROM_{NMR} templates prepared from destabilizing and unfolded sequences [Figure 4.8]. This result demonstrates that while SSD with ROM templates results in rotamer bias, the side chain contacts that give rise to this bias are different. Because it is difficult to predict *a priori* which side chain contacts are required to improve stability prediction accuracy by SSD, use of ROM templates prepared from stabilizing sequences in SSD represents a simple way to achieve this result *via* rotamer bias.

Our results described above suggest that the rotamer bias conferred through template preparation with the ROM procedure likely arises from variations in side chain contacts. To determine whether variations in side chain contacts result from alterations to the core cavity shape in each ROM template, two additional sets of templates were prepared using modified ROM procedures in which only side chain (scROM) or backbone (bbROM) atoms were relaxed by energy minimization, with all other atoms remaining fixed. If side chain contacts made by residues at designed positions cause the observed rotamer bias, it is expected that ROM templates prepared with their side chain atoms

fixed during minimization (i.e. bbROM templates) will not be able to confer a bias to SSD sequence enrichment because their core cavities will be identical.

Table 4.4. Sequence enrichment results for scROM and bbROM prepared templates

Template	Sequences in top 24			
	Stabilizing	Destabilizing	Unfolded	Non-native
scROM_{XTAL}				
WT	13	5	5	1
Stabilizing	13 ± 2	5 ± 1	5 ± 3	0
Destabilizing	10 ± 1	9 ± 2	5 ± 3	0
Unfolded	10 ± 1	4 ± 2	10 ± 3	0
Non-native	11 ± 3	4 ± 3	2 ± 2	7 ± 2
bbROM_{XTAL}				
WT	12	7	3	2
Stabilizing	11 ± 1	7 ± 1	6 ± 1	1 ± 0
Destabilizing	10 ± 1	7 ± 1	6 ± 1	1 ± 0
Unfolded	10 ± 1	6 ± 1	7 ± 1	1 ± 0
Non-native	11 ± 1	6 ± 1	5 ± 2	2 ± 1
scROM_{NMR}				
WT	14	2	8	0
Stabilizing	16 ± 3	1 ± 0	7 ± 3	0
Destabilizing	11 ± 5	8 ± 3	5 ± 6	0
Unfolded	3 ± 3	3 ± 2	17 ± 5	0
Non-native	1 ± 2	1 ± 1	7 ± 4	9 ± 1
bbROM_{NMR}				
WT	5	2	17	0
Stabilizing	5 ± 2	2 ± 1	17 ± 2	0
Destabilizing	11 ± 2	1 ± 1	12 ± 2	0
Unfolded	5 ± 3	2 ± 1	17 ± 2	0
Non-native	7 ± 4	0	10 ± 3	3 ± 3

As shown in Table 4.4 and Figure 4.9, SSD with bbROM templates did not result in significant biases in sequence enrichment profiles, regardless of the sequence used during template preparation. In contrast, SSD calculations using scROM_{XTAL} and scROM_{NMR} templates prepared with destabilizing, unfolded, and non-native sequences were able to skew enrichment profiles to favor sequences from their respective stability group. These results conclusively demonstrate that it is the shape of the core cavity

present in ROM templates that gives rise to biased SSD predictions, presumably through alterations to side chain contacts.

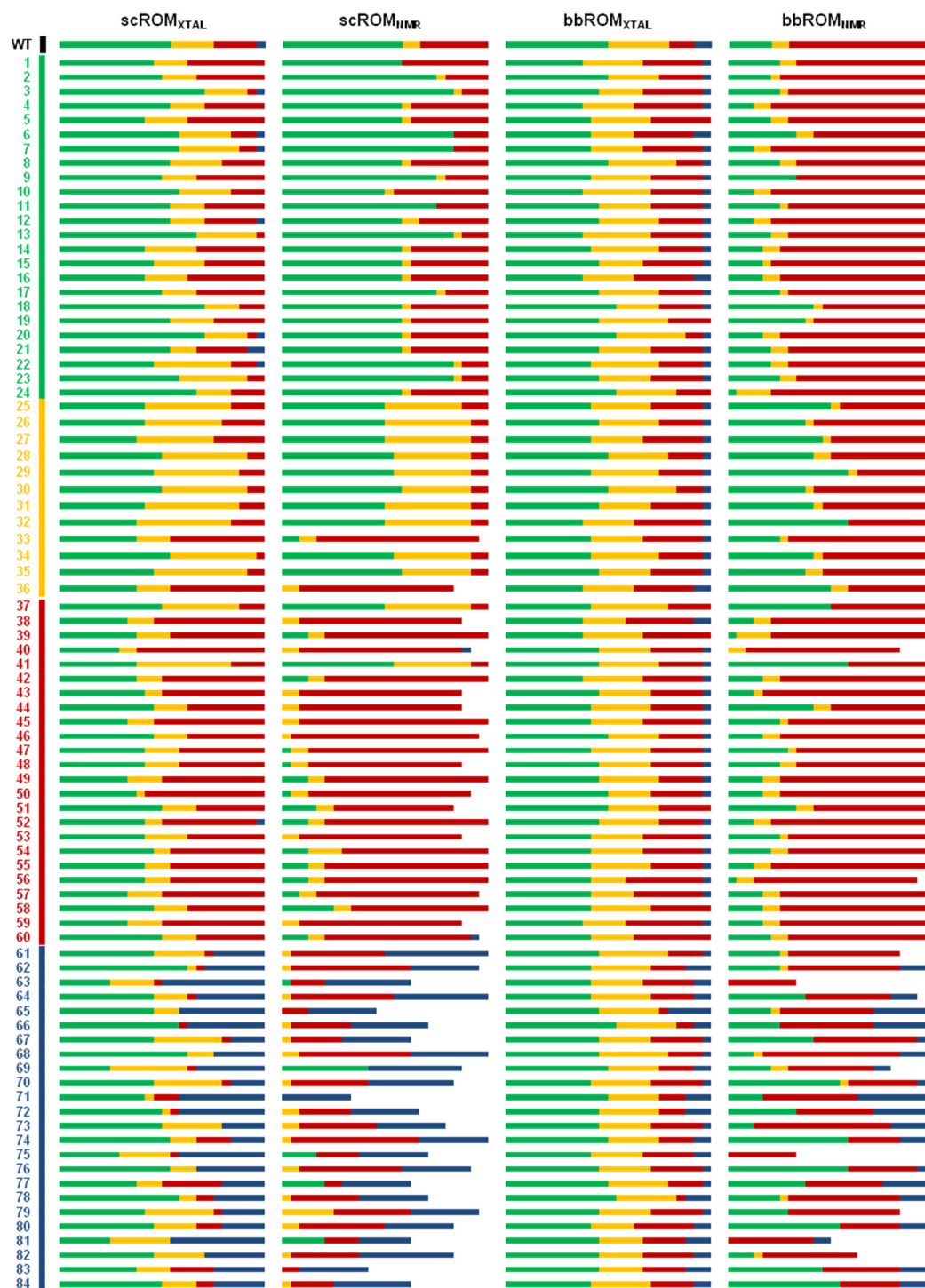


Figure 4.9. Sequence enrichment profiles with scROM and bbROM prepared templates. The top 24 sequences (excluding WT) predicted by SSD using rotamer optimized and

energy minimized crystal (ROM_{XTAL}) and NMR (ROM_{NMR}) templates are shown as bars. Only the backbone or side chain atoms were minimized in the case of bbROM or scROM templates, respectively, with all other atoms fixed during minimization. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing, destabilizing, unfolded, and non-native sequences colored green, yellow, red, and blue, respectively. ROM templates were prepared from the WT (WT, black) and 84 mutant Gβ1 sequences that are numbered and colored according to their stability group. Enrichment profiles of several ROM_{NMR} templates prepared from non-native sequences do not contain 24 sequences, and are shown as bars of reduced length.

4.5. Sequence Enrichment using Energetically Biased Single-state Design

The results described above demonstrate that rotamer bias arises from differences in side chain contacts present in the input template, which skew predictions during SSD. we therefore examined whether the rotamer bias effect could be reproduced by applying an energy bias favoring specific amino acid sequences or rotamer configurations as part of the scoring function (see Materials and Methods). Because there is no means of knowing in advance which bias weight to apply, we tested a range of weights [Figure 4.10 – Figure 4.13].

Calculations involving the application of a potential energy bias during SSD included an additional term to the scoring function. In the case of amino acid bias, each Gβ1 sequence was specified as a reference and sequence scores were adjusted according to their identity to the reference sequence. A range of amino acid bias weights were applied to each sequence (± 1000 , ± 10 , ± 4.2 , ± 2.8 , or ± 1.4 kcal·mol⁻¹), with biases applied to each sequence divided across the eight designed positions that vary in identity between Gβ1 sequences (positions 3, 5, 7, 30, 34, 39, 52, and 54). In the case of configuration bias, sequence scores were adjusted according to the rotamer

configuration and identity present at each designed position on a reference structure. No bias was applied to the alanine at position 34 because it has no rotamer. Because atomistic structures for each of the 84 mutant G β 1 sequences are not available, reference structures for configuration bias were generated by optimizing rotamers for each sequence on MIN_{XTAL} or MIN_{NMR} templates using SSD, as described above. A range of configuration bias weights were applied (± 1000 , ± 10 , ± 4.2 , ± 2.8 , and ± 1.4 kcal·mol⁻¹) to each designed position independently, in a similar fashion as was applied to amino acid bias weights. The configuration bias is applied to design position rotamers that share both identity and configuration with the rotamer at the same position in the reference structure. Inspection of enrichment profiles allowed us to determine optimal bias weights of 10 kcal·mol⁻¹ favoring rotamer configuration and 100 kcal·mol⁻¹ favoring amino acid sequence. It is important to note that the application of biases disfavoring either rotamer configuration or amino acid sequence were unable to alter enrichment profiles, and that the application of the strongest bias weight favoring amino acid sequence or rotamer configuration (-1000 kcal·mol⁻¹) yielded enrichment profiles containing fewer than 24 sequences. To determine whether the effect that these biases had on SSD enrichment profiles was significant, we calculated the average and standard deviation of the number of sequences from each stability group that were found in the top 24 for all templates prepared from sequences of the same group. Table 4.5 shows that the application of potential energy biases to SSD calculations can result in the selective enrichment of sequences for some but not all stability groups, unlike what was observed by SSD with ROM_{XTAL} templates.



Figure 4.10. Sequence enrichment profiles for amino acid biased MIN_{XTAL} SSD. The top 24 predicted sequences (excluding WT) are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing (1-24), destabilizing (25-36), unfolded (37-60), and non-native (61-84) sequences colored green, yellow, red, and blue, respectively. Various bias weights in $\text{kcal}\cdot\text{mol}^{-1}$ were applied to the scoring function to favor (negative bias weights) or disfavor (positive bias weights) the amino acid sequence of the WT (WT, black) or of one of the 84 $\text{G}\beta 1$ mutants. Enrichment profiles that do not contain 24 sequences are shown as bars of reduced length.



Figure 4.11. Sequence enrichment profiles for amino acid biased MIN_{NMR} SSD. The top 24 predicted sequences (excluding WT) are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing (1-24), destabilizing (25-36), unfolded (37-60), and non-native (61-84) sequences colored green, yellow, red, and blue, respectively. Various bias weights in $\text{kcal}\cdot\text{mol}^{-1}$ were applied to the scoring function to favor (negative bias weights) or disfavor (positive bias weights) the amino acid sequence of the WT (WT, black) or of one of the 84 $\text{G}\beta 1$ mutants. Enrichment profiles that do not contain 24 sequences are shown as bars of reduced length.

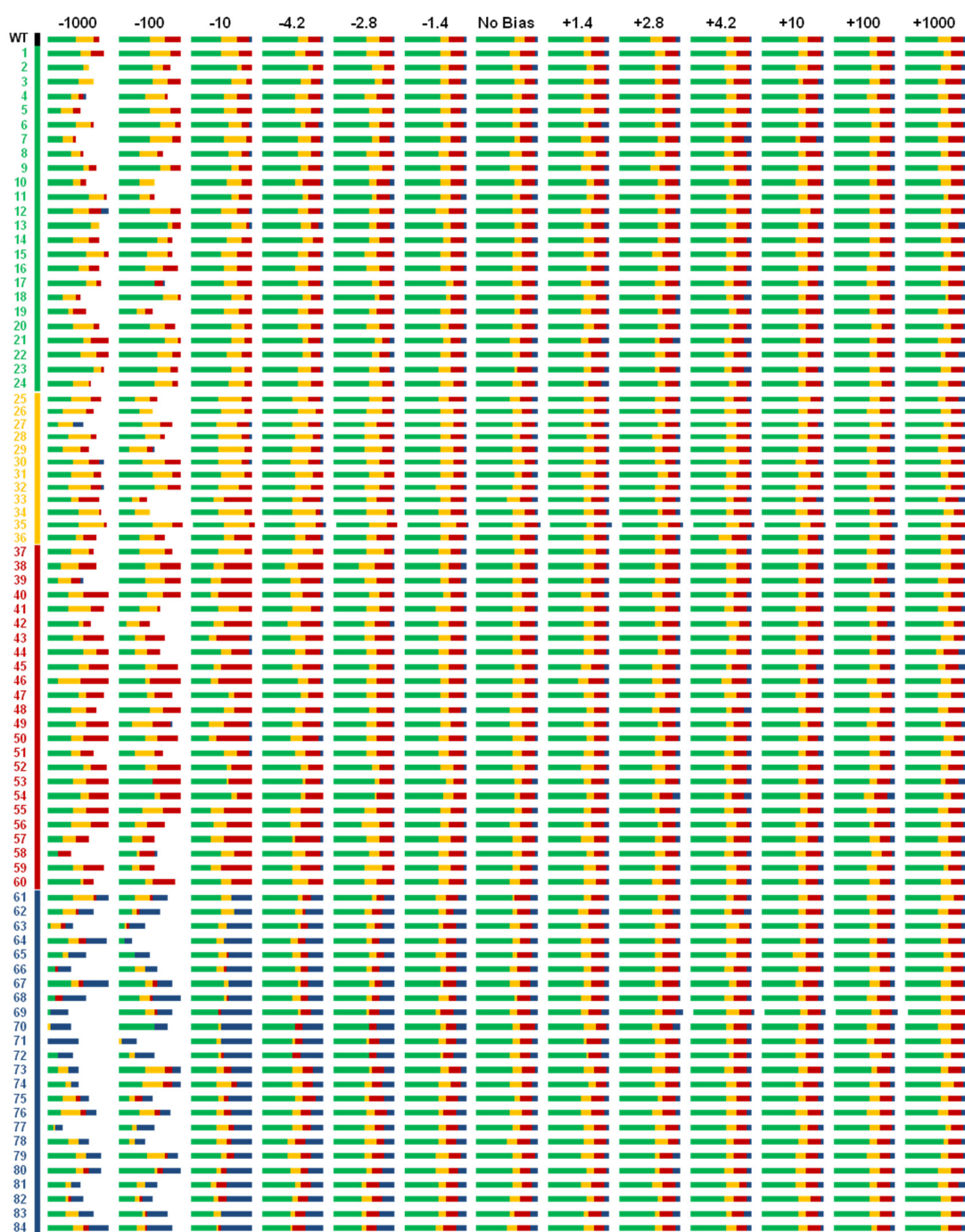


Figure 4.12. Sequence enrichment profiles for configuration biased MIN_{XTAL} SSD. The top 24 predicted sequences (excluding WT) are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing (1-24), destabilizing (25-36), unfolded (37-60), and non-native (61-84) sequences colored green, yellow, red, and blue, respectively. Various bias weights in kcal·mol⁻¹ were applied to the scoring function to favor (negative bias weights) or disfavor (positive bias weights) the rotamer configuration for the WT (WT, black) or one of the 84 Gβ1 mutants following rotamer optimization on the minimized crystal structure. Enrichment profiles that do not contain 24 sequences are shown as bars of reduced length.

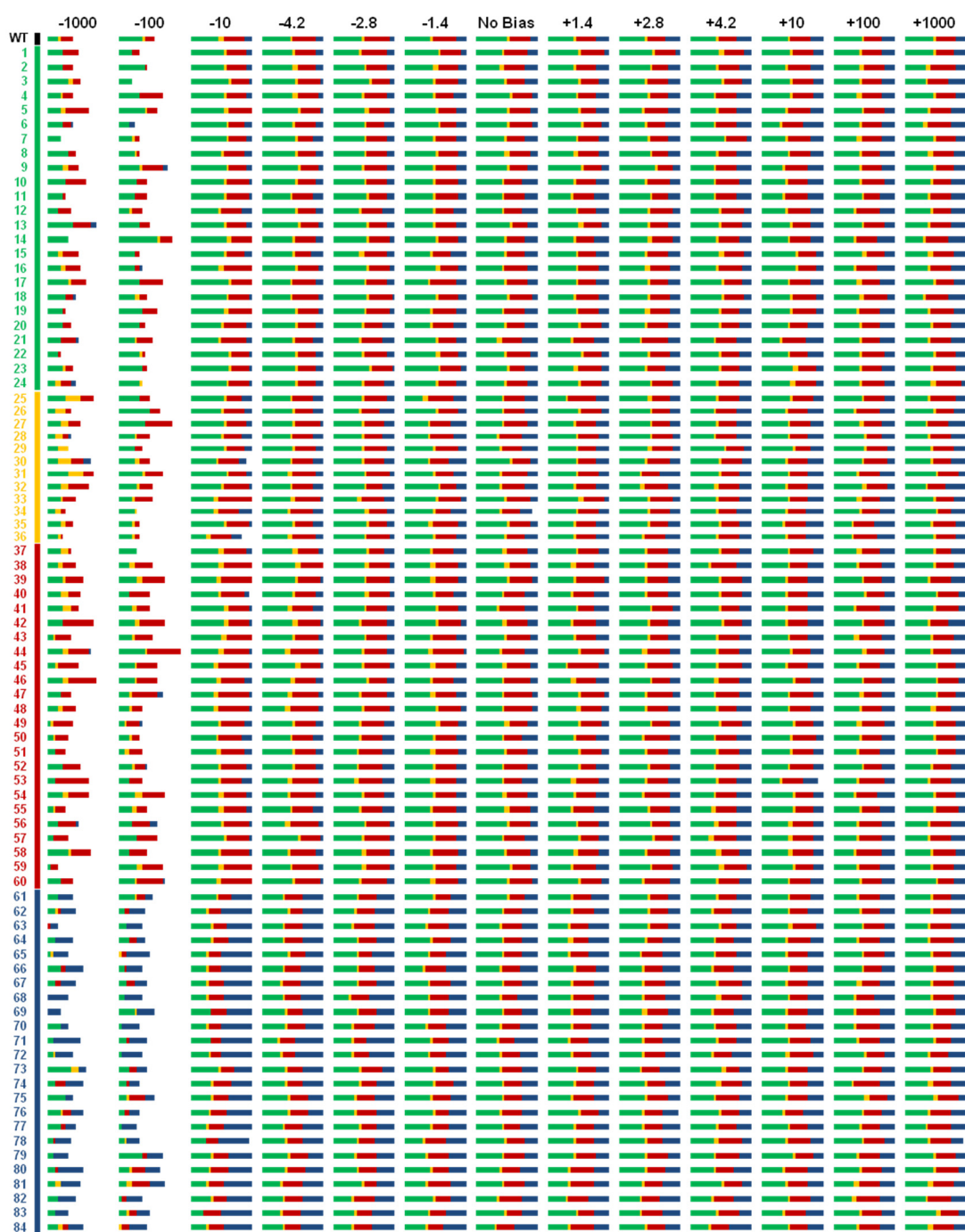


Figure 4.13. Sequence enrichment profiles for configuration biased MIN_{NMR} SSD. The top 24 predicted sequences (excluding WT) are shown as bars. Each bar is colored according to the proportion of sequences from each stability group found in the top 24, with stabilizing (1-24), destabilizing (25-36), unfolded (37-60), and non-native (61-84) sequences colored green, yellow, red, and blue, respectively. Various bias weights in $\text{kcal}\cdot\text{mol}^{-1}$ were applied to the scoring function to favor (negative bias weights) or disfavor (positive bias weights) the rotamer configuration for the WT (WT, black) or one of the 84 G β 1 mutants following rotamer optimization on the minimized NMR structure. Enrichment profiles that do not contain 24 sequences are shown as bars of reduced length.

Table 4.5. Sequence enrichment results for energetically biased SSD

Template	Sequences in top 24			
	Stabilizing	Destabilizing	Unfolded	Non-native
No Bias				
MIN _{XTAL}	14	4	4	2
MIN _{NMR}	11	1	8	4
MIN_{XTAL} Amino Acid Bias^a				
Stabilizing	16 ± 2	4 ± 1	4 ± 2	0
Destabilizing	12 ± 1	8 ± 2	4 ± 3	1 ± 1
Unfolded	11 ± 1	3 ± 2	10 ± 2	0
Non-native	10 ± 1	3 ± 2	1 ± 1	11 ± 2
MIN_{NMR} Amino Acid Bias^a				
Stabilizing	13 ± 2	1 ± 1	8 ± 1	1 ± 1
Destabilizing	11 ± 2	1 ± 1	8 ± 2	2 ± 1
Unfolded	10 ± 2	2 ± 1	10 ± 2	1 ± 1
Non-native	8 ± 2	1 ± 1	4 ± 2	11 ± 1
MIN_{XTAL} Configuration Bias^b				
Stabilizing	15 ± 2	5 ± 1	4 ± 1	0
Destabilizing	11 ± 1	8 ± 2	4 ± 3	0
Unfolded	10 ± 3	4 ± 2	9 ± 3	0
Non-native	11 ± 1	3 ± 1	1 ± 1	10 ± 2
MIN_{NMR} Configuration Bias^b				
Stabilizing	13 ± 1	1 ± 1	8 ± 1	2 ± 1
Destabilizing	11 ± 2	1 ± 1	8 ± 2	3 ± 2
Unfolded	11 ± 1	2 ± 1	10 ± 1	2 ± 1
Non-native	7 ± 1	1 ± 1	6 ± 1	10 ± 2

^aA 100 kcal·mol⁻¹ bias was applied in favor of amino acid identity at each design position

^bA 10 kcal·mol⁻¹ bias was applied in favor of rotamer configuration at each design position

Interestingly, amino acid bias does not lead to an enrichment of stabilizing sequences when using the MIN_{XTAL} or MIN_{NMR} structures, whereas rotamer configuration bias does so only for templates derived from the MIN_{NMR} structure. Furthermore, although application of an amino acid bias can improve binning success rates relative to

unbiased calculations [Table 4.6], only a minority of calculations biased towards stabilizing sequences on either the MIN_{XTAL} [Figure 4.2, Sequences 3, 13, and 23] or MIN_{NMR} structures [Figure 4.2, Sequences 1, 4, 8, 12, 13, 17, and 21] resulted in improved success rates relative to SSD using corresponding ROM_{XTAL} and ROM_{NMR} templates [Table 4.1].

A similar result is obtained by applying a rotamer configuration bias. These results demonstrate that the application of potential energy biases in the scoring function is not as efficient as the rotamer bias caused by the use of ROM templates in improving the accuracy of protein stability predictions by SSD.

Throughout this work, we have demonstrated that SSD calculations can be biased through a variety of methods, including amino acid sequence or rotamer configuration weights added to the scoring function as well as structural alterations to the template resulting from preparation procedures such as ROM, bbROM, and scROM. To identify which of these biasing methods is the most useful for skewing SSD predictions, we investigated how they would alter the position of the seed sequence (i.e. the input template sequence) in the ranked list of sequences predicted by SSD. The distributions shown in Figure 4.14 depict the frequency at which seed sequences occupy each position in the ranked list of 85 Gβ1 sequences obtained following biased SSD calculations.

Table 4.6. Sequence binning predictions of energetically biased SSD calculations

	MIN _{XTAL}						MIN _{NMR}					
	Success Rate (%)	True Positive	False Negative	False Positive	True Negative	Cut-off (kcal/mol)	Success Rate	True Positive	False Negative	False Positive	True Negative	Cut-off (kcal/mol)
No Bias												
WT	79	8	16	2	58	-71.5	71	2	22	2	58	-63.6
Amino Acid Bias												
WT	73	1	23	0	60	-171.5	73	1	23	0	60	-162.7
1	76	6	18	2	58	-142.1	75	3	21	0	60	-133.7
2	86	16	8	4	56	-117.4	81	8	16	0	60	-119.3
3	92	17	7	0	60	-125.6	85	11	13	0	60	-116.2
4	79	9	15	3	57	-140.5	75	3	21	0	60	-134.0
5	76	5	19	1	59	-156.1	73	1	23	0	60	-148.5
6	79	6	18	0	60	-142.0	76	4	20	0	60	-133.9
7	87	17	7	4	56	-115.1	82	9	15	0	60	-115.3
8	75	17	7	14	46	-90.4	80	8	16	1	59	-111.6
9	83	11	13	1	59	-140.7	75	3	21	0	60	-134.2
10	71	16	8	16	44	0 ^c	80	9	15	2	58	-105.5
11	89	18	6	3	57	-115.6	81	8	16	0	60	-118.6
12	77	5	19	0	60	-155.2	74	2	22	0	60	-147.9
13	82	9	15	0	60	-141.9	79	6	18	0	60	-134.0
14	82	14	10	5	55	-126.4	75	3	21	0	60	-132.3
15	73	1	23	0	60	-171.3	73	1	23	0	60	-162.8
16	76	4	20	0	60	-155.5	73	1	23	0	60	-148.5
17	82	9	15	0	60	-142.0	76	4	20	0	60	-134.2
18	81	8	16	0	60	-139.2	76	4	20	0	60	-132.7
19	81	9	15	1	59	-141.5	76	4	20	0	60	-131.4
20	81	9	15	1	59	-140.2	75	5	19	2	58	-122.2
21	75	3	21	0	60	-157.2	75	3	21	0	60	-148.5
22	75	3	21	0	60	-156.3	74	2	22	0	60	-148.4
23	79	6	18	0	60	-141.3	76	4	20	0	60	-131.7
24	75	3	21	0	60	-155.4	74	2	22	0	60	-146.7
Configuration Bias												
WT	73	1	23	0	60	-141.5	73	1	23	0	60	-133.7
1	77	9	15	4	56	-117.8	74	2	22	0	60	-113.7
2	88	14	10	0	60	-111.0	76	4	20	0	60	-97.8
3	86	16	8	4	56	-98.2	85	11	13	0	60	-97.1
4	79	10	14	4	56	-117.9	74	2	22	0	60	-113.7
5	76	5	19	1	59	-128.7	75	3	21	0	60	-116.8
6	81	13	11	5	55	-98.4	76	4	20	0	60	-107.1
7	74	16	8	14	46	-90.4	81	8	16	0	60	-96.8
8	80	12	12	5	55	-109.3	76	4	20	0	60	-93.0
9	80	7	17	0	60	-121.3	74	2	22	0	60	-113.7
10	75	14	10	11	49	-97.8	76	4	20	0	60	-93.8
11	75	16	8	13	47	-90.5	77	5	19	0	60	-96.5
12	76	4	20	0	60	-131.3	73	2	22	1	59	-123.7
13	86	12	12	0	60	-100.6	79	6	18	0	60	-106.6
14	81	11	13	3	57	-113.9	79	6	18	0	60	-97.9
15	73	1	23	0	60	-141.5	73	1	23	0	60	-133.7
16	77	6	18	1	59	-128.7	75	3	21	0	60	-116.7
17	81	8	16	0	60	-119.0	74	2	22	0	60	-113.7
18	89	15	9	0	60	-110.1	77	5	19	0	60	-102.7
19	74	11	13	9	51	-100.5	79	6	18	0	60	-96.9
20	81	9	15	1	59	-121.0	77	5	19	0	60	-101.6
21	77%	5	19	0	60	-118.8	75%	3	21	0	60	-121.6
22	76%	4	20	0	60	-110.6	74%	2	22	0	60	-116.6
23	79%	6	18	0	60	-110.5	76%	4	20	0	60	-106.8
24	75%	3	21	0	60	-131.3	75%	3	21	0	60	-112.4

^aA 100 kcal·mol⁻¹ bias was applied in favor of amino acid identity at each design position^bA 10 kcal·mol⁻¹ bias was applied in favor of rotamer configuration at each design position^cA 1cut-off value of 0 kcal·mol⁻¹ bias was assigned to sequence 10 because SSD with an amino acid bias could not favorably evaluate the WT sequence

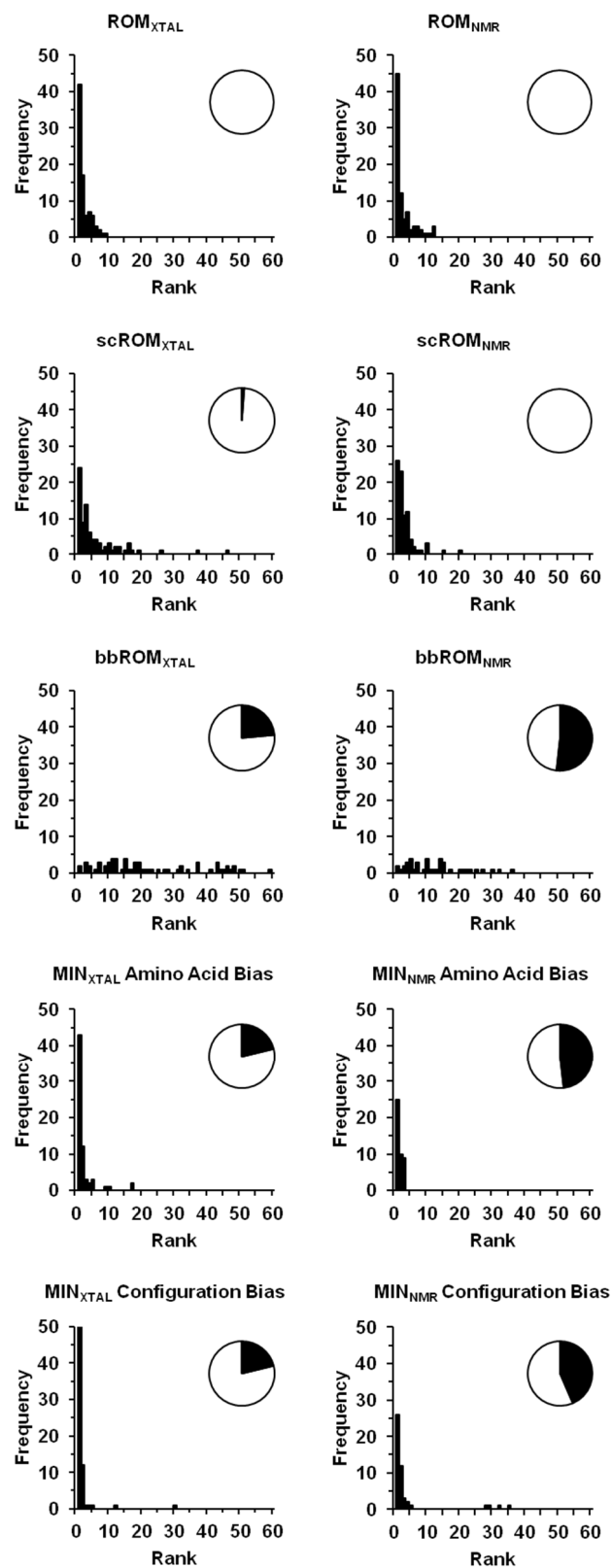


Figure 4.14. Seed sequence ranking by biased SSD calculations. Distributions depict the frequency at which seed sequences were ranked at each position in the ranked list of

sequences obtained following biased SSD calculations. Seed sequences (listed in Figure S1) are the sequences present in each input template used in SSD. The fraction of the pie chart in black corresponds to the percentage of the 85 G β 1 seed sequences that could not be scored favorably by biased SSD calculations. For the amino acid sequence and rotamer configuration biases, weights of -100 and -10 kcal·mol⁻¹ were applied to the SSD scores, respectively.

As expected, SSD using ROM_{XTAL} and ROM_{NMR} templates results in a narrow distribution of seed sequences occupying the top 5 positions in their respective rank ordered lists. Broader distributions are obtained for SSD calculations using scROM templates, however the majority of seed sequences still occupy the top 5 positions. In contrast, SSD calculations using bbROM templates could not improve the rank order for most seed sequences, nor could they favorably score a large number of the 85 G β 1 seed sequences (24% and 52% for bbROM_{XTAL} and bbROM_{NMR}, respectively). Finally, SSD using amino acid sequence or rotamer configuration biases also improved the position of seed sequences in the ranked list of scored sequences, giving narrow distributions with most seed sequences ranked in the top 5 positions. However, contrary to the results obtained using ROM and scROM templates, a large number of seed sequences (21% – 48%) could not be scored favorably on the MIN_{XTAL} and MIN_{NMR} templates even when using potential energy biases, suggesting that they are not suitable templates for these sequences. These results demonstrate that SSD with ROM templates performs better than other methods to bias calculations towards specific sequences.

4.6. Application of Rotamer Biased Templates to Multistate Analysis

The ROM procedure described here can be used to generate templates for SSD that lead to improved stability prediction accuracy. As well, this procedure can be used to generate templates that bias predictions towards undesirable sequences such as those that are unfolded or adopt a non-native fold, even in the absence of atomistic models for unfolded and non-native states of G β 1. This bias is likely the consequence of side chain contacts present in ROM templates that skew predictions towards sequences of similar folding behaviour. Because of this bias in SSD predictions, it may be possible to use ROM templates prepared from unfolded or non-native sequences as undesired competing states in negative design. In this approach, sequences that are scored more favorably on ROM templates prepared from unfolded or non-native sequences than on ROM templates prepared from stabilizing sequences would be eliminated from the ranked list of scored sequences, potentially leading to improved prediction accuracy by reducing the number of false positives. Another potential application of the ROM procedure would be as part of an iterative process combining rounds of CPD with experimental validation. In the first step of this iterative process, SSD with a MIN_{XTAL} or WT ROM_{XTAL} template would be used to generate a list of mutant sequences ranked by their predicted stability. Then, the stability of these predicted sequences would be assessed experimentally to identify stable mutants. Finally, stable mutant sequences would be used to prepare new ROM_{XTAL} templates that would serve as input backbones for subsequent rounds of SSD until no new sequences are predicted. In this way, the ROM procedure may be applied even in the absence of previously known stable mutant

sequences. Alternatively, this process may be initiated using prior stability information for mutant sequences obtained from databases such as ProTherm [Kumar *et al.*, 2006].

In addition to the applications stated above, ROM templates could also serve as structural models for backbone drift modelling. In backbone drift modelling [Howell *et al.*, 2014], a set of templates consisting of available crystal structures for a protein of interest is used to score mutant sequences in order to identify their most favorable score from each of the templates. Use of this score to represent each designed sequence results in substantial improvements to stability prediction accuracy [Howell *et al.*, 2014]. Backbone drift modelling thus makes the assumption that the correct rotamer configuration and score of any given sequence is that obtained from the backbone template on which that sequence is scored most favorably. To verify whether we could successfully perform backbone drift modelling with ROM templates, we identified the ROM template prepared from stabilizing sequences that results in the most favorable score for each G β 1 mutant sequence, and used those scores in sequence binning analysis. Backbone drift modelling with ROM_{XTAL} and ROM_{NMR} templates prepared from stabilizing sequences results in improved binning success rates [Table 4.7] compared to SSD using the MIN_{XTAL} or MIN_{NMR} structures, respectively [Table 4.1]. Backbone drift modelling was also performed using all G β 1 depositions available in the PDB, and this also resulted in improved success rates compared to SSD using MIN_{XTAL} and MIN_{NMR} input templates. However, backbone drift modelling using the ROM_{XTAL} templates resulted in the highest success rate (92%) due to the lowest number of false negatives and the highest number of true positives. The improved success rate of backbone drift

modelling with ROM_{XTAL} templates compared to backbone drift modelling with PDB structures demonstrates that structural models prepared *in silico* are as relevant for the scoring of mutant sequences as structures derived from experimental data. Thus, in the absence of crystal or NMR structures for mutant sequences, the ROM procedure may be used to generate high-quality input templates for CPD.

Table 4.7. Sequence binning backbone drift predictions with ROM templates

	Success Rate (%)	True Positive	False Negative	False Positive	True Negative	Cut-off (kcal·mol ⁻¹)
ROM _{XTAL} ^a	92	20	4	3	57	-74.8
ROM _{NMR} ^a	81	12	12	4	56	-76.5
PDB ^b	86	12	12	0	60	-78.4

^aROM templates were prepared using stabilizing sequences [Figure 4.2, Sequences 1-24]

^bSet includes 17 Gβ1 depositions available in the PDB: 1EM7 [Strop *et al.*, 2000], 1IGD [Derrick and Wigley, 1994], 1PGA [Gallagher *et al.*, 1994], 1PGB [Gallagher *et al.*, 1994], 2GB1 [Gronenborn *et al.*, 1991], 2IGD [Butterworth *et al.*, 1997], 2IGH [Lian *et al.*, 1992], 2J52 [Wilton *et al.*, 2008], 2J53 [Wilton *et al.*, 2008], 2KOP [Robustelli *et al.*, 2008], 2LGI [Wylie *et al.*, 2011], 2RMM (chain-A and chain-B) [Jee *et al.*, 2007], 2RPV [Saio *et al.*, 2009], 3GB1 [Kuszewski *et al.*, 1999], 3MP9 (chain-A and chain-B) [Tomlinson *et al.*, 2010]

ROM templates can also be used to access larger areas of sequence space during SSD. To illustrate this, we compiled all top 100 sequences predicted by SSD using ROM_{XTAL} templates from the same stability group and applied backbone drift modelling to find the best score for each unique sequence contained within these lists. All sequences were then re-ranked using the backbone drift modelling score and a new top 100 was compiled for each stability group. For the ROM_{XTAL} and MIN_{XTAL} templates

prepared from the WT sequence, no backbone drift modelling was applied to the top 100 sequences predicted by SSD. Analysis of the top 100 sequences [Figure 4.15] shows that SSD with the MIN_{XTAL} and WT ROM_{XTAL} templates results in the highest number of shared sequences (62 out of 100), an unsurprising result given that both templates were prepared from identical seed sequences. However, overlap of sequences predicted by backbone drift modelling using ROM_{XTAL} templates of different stability groups is lower, with a maximum of 54 shared sequences. This result shows that at least 46 different sequences can be found in the top 100 depending on the ROM templates used, demonstrating that incorporation of sequence information by the ROM procedure skews the predicted sequence space away from that of other templates.

Most impressively, backbone drift modelling across ROM_{XTAL} templates prepared from stabilizing sequences results in a top 100 containing all 24 stabilizing sequences, ten more than were predicted using the MIN_{XTAL} template, and only 30 sequences that are shared with the MIN_{XTAL} top 100. It may be possible that several of the remaining 70 sequences have a stabilizing effect on G β 1, but further experiments are needed to confirm this hypothesis. In addition, sequence motifs prepared from the top 100 sequences [Figure 4.15] show that ROM templates can be used to predict mutations that are not included in the 85 G β 1 seed sequences [Figure 4.2] from which they are derived. All of these results illustrate how SSD with ROM templates increases the diversity of predicted sequences, thereby allowing evaluation of novel combinations of mutations.

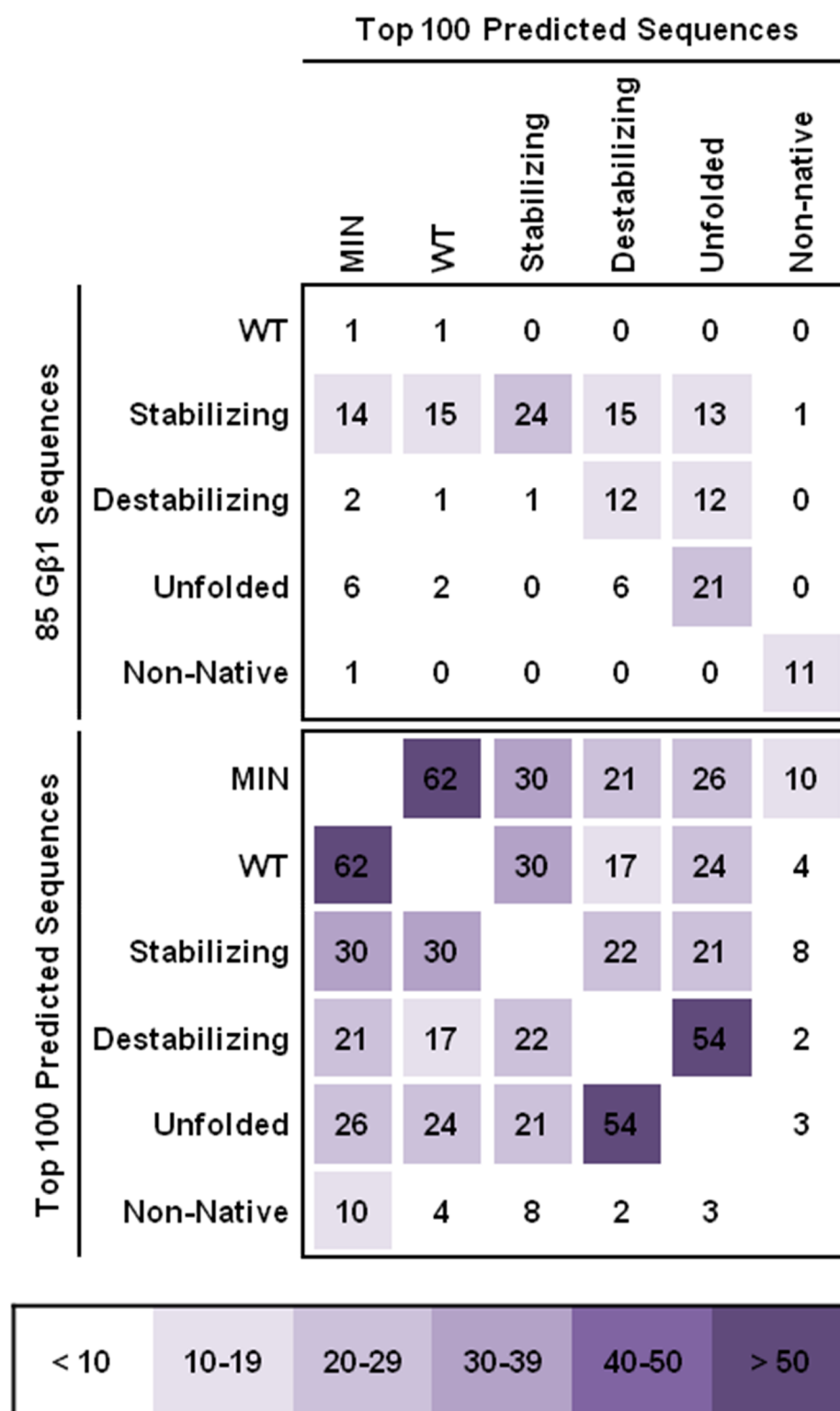


Figure 4.15. Sequence space of backbone drift modeling with ROM_{XTAL} templates. The top 100 sequences predicted by backbone drift modelling (see text for details) using ROM_{XTAL} templates from a same stability group were compared to those predicted for

other stability groups, and the number of shared sequences is indicated. Comparisons are also included for the MIN_{XTAL} (MIN) and WT ROM_{XTAL} templates. The top 100 sequences were obtained by designing core residues of Gβ1 (positions 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54) using hydrophobic amino acids (A, V, L, I, and F) as well as the WT Y amino acid at position 3. The number of shared sequences in the top 100 that are part of the 85 Gβ1 seed sequences [Figure 4.2] is also reported. Overlap in sequence space is colored according to the number of shared sequences.

The results obtained by backbone drift modelling illustrate the importance of using alternate structures to better score mutant sequences and increase the diversity of predicted sequences. Because amino acid mutations can alter the protein backbone as well as the side chain conformation of neighboring residues, the scoring by SSD of each mutant sequence on a template that is a more accurate model of its actual structure than the WT structure is preferable. However, since structures for all possible mutants of a protein of interest are not available, sets of templates prepared *in silico*, such as ROM templates, may provide more suitable structural models for scoring mutant sequences.

Methods incorporating backbone flexibility during calculation have also resulted in improved CPD predictions [Borgo and Havranek, 2012; Davey and Chica, 2014; Kapp *et al.*, 2012; Wang *et al.*, 2005] through the identification of more suitable backbones for the scoring of mutant sequences. Thus, we propose that many failed CPD predictions may result from scoring of sequences on incorrect templates for that sequence rather than by deficiencies in the energy function, in agreement with observations made by Gainza *et al.* [Gainza *et al.*, 2012].

4.7. Conclusions on the Study of Rotamer Bias Methodologies

This chapter demonstrates the existence of a bias in SSD that arises from variations in side chain contacts made by rotamers in the input template used. This rotamer bias causes calculations to score more favorably sequences of similar stability to the sequence used to generate the input template. Because of this, rotamer bias can be exploited to improve the accuracy of protein stability predictions through the use of ROM templates prepared from stable mutant sequences, leading to increased true positives and fewer false negatives. While other methods such as flexible backbone design [Babor *et al.*, 2011; Smith and Kortemme, 2008; Smith and Kortemme, 2010], MSD with native backbone ensembles [Allen *et al.*, 2010; Davey and Chica, 2014], continuous rotamer optimization [Gainza *et al.*, 2012], and on-the-fly energy minimization [Wang *et al.*, 2005] have been developed to improve the accuracy of predictions made by CPD, SSD with ROM templates prepared from stabilizing sequences represents a useful alternative that is easier to implement and less computationally expensive. Indeed, implementation of the ROM procedure requires only standard single-state CPD protocols and energy minimization. Although prior knowledge of stable mutant sequences is useful to prepare ROM templates that enable the improvement of stability prediction accuracy, application of the ROM procedure to the WT sequence can also lead to improved predictions by SSD. In the future, it will be interesting to see if ROM templates can be used as input templates in SSD for the prediction of protein sequences exhibiting desired properties other than improved stability, such as specific ligand binding or enhanced catalytic activity.

CHAPTER 5

Applied Positive Multistate Design:

Predicting the Substrate Multi-specificity of SMYD2 Methyltransferase

Characterization of lysine methylation has proven challenging despite its importance in biological processes such as gene transcription, protein turnover and cytoskeletal organization. In contrast to other key post-translational modifications (PTMs), current proteomics techniques have thus far been unsuccessful at characterizing methyl-lysine residues across the cellular landscape. To complement current biochemical characterization methods, we developed a multistate CPD procedure to probe the substrate specificity of the protein lysine methyltransferase (PKMT) SMYD2. Modelling of substrate-bound SMYD2 identified residues important for substrate recognition and predicted amino acids necessary for methylation. Peptide- and protein- based substrate libraries confirmed that SMYD2 activity is dictated by the motif LFM-K*-AFYMSHRK-LYK around the target lysine K*. Comprehensive motif based searches and mutational analysis further established four novel SMYD2 substrates. Our methodology paves the way to systematically predict and validate novel PTM sites while simultaneously pairing those with their associated enzyme.

Some of the language and figures in this chapter were adapted from a research article coauthored with Dr. Sylvain Lanouette, Dr. Roberto A. Chica, and Dr. Jean-François Couture.

Lanouette S, Davey JA, Elisma F, Ning Z, Figeys D, Chica RA, and Couture J-F (2015) Discovery of substrates for a SET domain lysine methyltransferase predicted by multistate computational protein design. *Structure*, 23(1):206-215 [Lanouette *et al.*, 2015].

I would like to formally acknowledge the contributions of these authors to the completion of this chapter. Specifically:

James A Davey: Calculation design, predictions, and comparisons with experiment

Dr. Sylvain Lanouette: Experimental design and results

Dr. Roberto A Chica: Calculation design

Dr. Jean-François Couture: Experimental design

5.1. Lysine Methylation in Biology

PTM of proteins adds a layer of complexity to the proteome that is critical to biological regulation. Although proteomic studies have enabled the comprehensive study of PTMs such as phosphorylation and acetylation, the characterization of many other types of PTMs remains a significant challenge [Choudhary *et al.*, 2009; Ptacek *et al.*, 2005]. Among these, lysine methylation - the transfer by PKMTs of one, two or three methyl groups to the α -amine of a lysine side chain [Schubert *et al.*, 2003] - is a prevalent PTM associated with critical cellular processes including cell cycle progression, chromosome segregation, and pathogen infection [Lanouette *et al.*, 2014]. The relatively low abundance and chemically inert nature of lysine methylation hinder its detection by current proteomics methods. Moreover, the expensive and tedious protocols currently used to study lysine methylation limit the high-throughput biochemical characterization of PKMT substrates [Kudithipudi *et al.*, 2012; Kudithipudi *et al.*, 2014; Rathert *et al.*, 2008].

SMYD2 is a PKMT that plays critical roles in muscle development and myofibril formation [Blais *et al.*, 2005; Donlin *et al.*, 2012; Voelkel *et al.*, 2013] as well as proper endodermal development during embryonic stem cell differentiation [Sese *et al.*, 2013]. It is misregulated in oesophageal squamous cell carcinomas [Komatsu *et al.*, 2009], bladder tumours [Cho *et al.*, 2012], leukemia stem cells [Sakamoto *et al.*, 2014; Zuber *et al.*, 2011] and doxorubicin-resistant breast cancer [Barros *et al.*, 2010]. The oncogenic phenotypes of SMYD2 depend on its methyltransferase activity [Komatsu *et al.*, 2009; Sakamoto *et al.*, 2014] which has also been shown to regulate gene transcription

(methylation of histone H3 [Abu-Farha *et al.*, 2008] and Estrogen Receptor α [Zhang *et al.*, 2013]), cell cycle progression (methylation of Retinoblastoma protein [Cho *et al.*, 2012; Saddic *et al.*, 2010]), apoptosis (methylation of p53 [Huang *et al.*, 2006]) and oxidative stress (methylation of Poly (ADP-ribose) polymerase 1 [Piao *et al.*, 2014]). The substrates currently known for SMYD2 are most likely only a subset of its full range of methylation targets.

The crystal structure of the SMYD2-p53 complex shows that substrate binding occurs in a narrow cleft between the catalytic SET domain and a C-terminal tetratricopeptide repeat domain [Ferguson *et al.*, 2011; Sirinupong *et al.*, 2011; Sirinupong *et al.*, 2010]. As the current list of SMYD2 substrates display little sequence or structural similarity in the topological arrangement of their protein fold, the mechanism by which this interface directs substrate specificity is currently unknown. A better understanding of the structural determinants of SMYD2 specificity would thereby allow for the identification of additional substrates, providing information on its underlying biological functions and disease associations.

We present here a CPD approach to define a substrate recognition motif for SMYD2. This approach is based on MSD, an emerging methodology in CPD that predicts stable protein sequences in the context of multiple backbones instead of a single fixed backbone template [Davey and Chica, 2012]. In collaboration with the Couture laboratory, the specificity of SMYD2 was evaluated using methyltransferase assays on both peptide and full-length protein substrates. The recognition motifs derived using both computational and experimental techniques were in remarkable agreement and

established that SMYD2 recognizes the LFM-K*-AFYMSHRK-LYK sequence. In combination with bioinformatics analyses and methyltransferase assays, we discovered four previously unknown substrates of SMYD2: SIN3B, SIX1, SIX2 and DDX15, demonstrating the utility of MSD as a discovery tool for the study of PTM enzymes.

5.2. Prediction of the SMYD2 Recognition Motif by Multistate Design

Several CPD protocols have been developed to design the specificity and stability of protein-protein interactions [Karanicolas and Kuhlman, 2009; Mandell and Kortemme, 2009]. These protocols are typically based on the SSD approach whereby protein sequences are optimized in the context of a single protein complex structure [Alvizo *et al.*, 2012; Grigoryan and Keating, 2006; Sammond *et al.*, 2007]. In this approach, discrete amino acid side chain rotamers are first threaded on the fixed polypeptide chain of each binding pair protein, followed by rotamer refinement to improve their packing interactions at the protein-protein interface. A list of ranked sequences is then returned based on their score value following sequence optimization. Additional procedures including protein docking [Huang *et al.*, 2007b; Jha *et al.*, 2010] and flexible backbone design [Smith and Kortemme, 2008] can also be incorporated in this process to improve predictions. Recently, MSD, a CPD methodology that utilizes conformational ensembles as inputs instead of a single backbone template, was developed as an alternate approach for the design of protein-protein interfaces [Humphris and Kortemme, 2008]. MSD improves the prediction of stable and functional protein sequences [Allen *et al.*, 2010; Babor *et al.*, 2011; Davey and Chica, 2014] through

improved packing interactions resulting from small variations in backbone geometry. Although MSD has been used to recapitulate known binding interactions [Smith and Kortemme, 2010] and to engineer new ones [Grigoryan *et al.*, 2009; Kapp *et al.*, 2012], it has not yet been applied towards the prediction and discovery of previously unknown substrates of PTM enzymes.

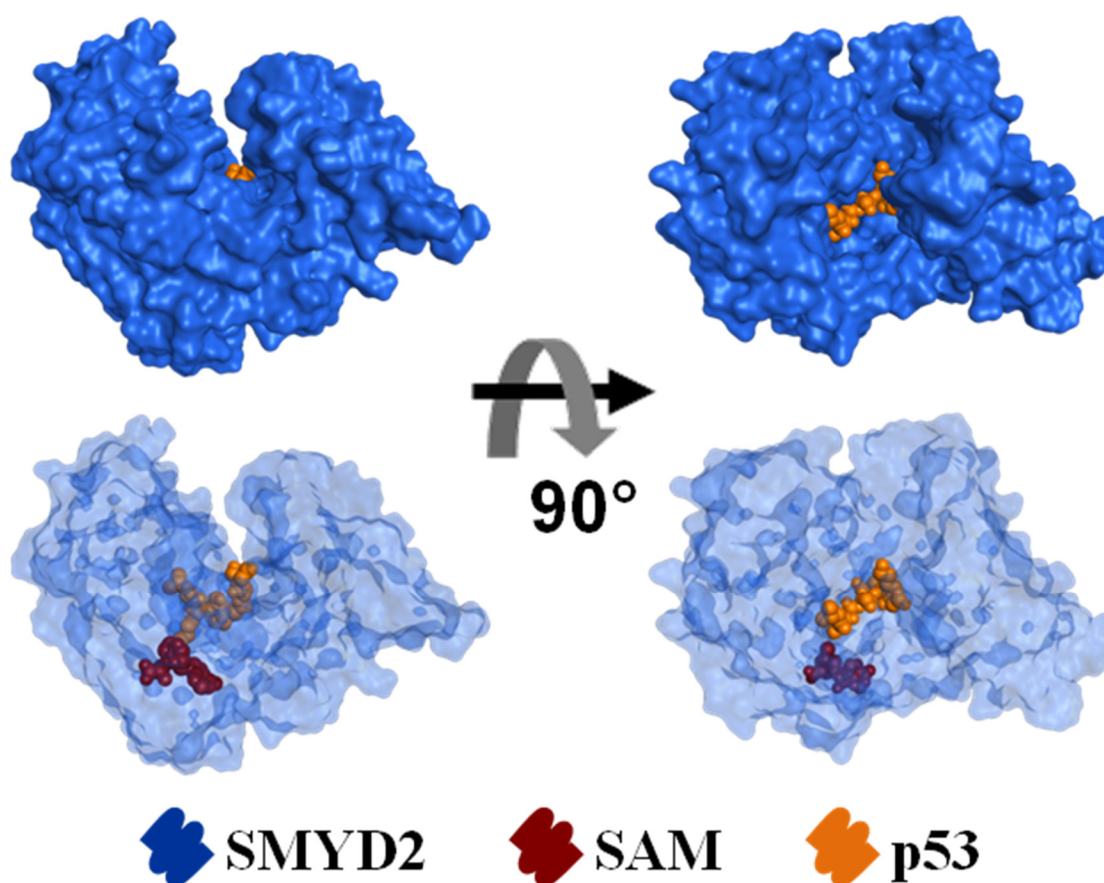


Figure 5.1. The p53-SMYD2 peptide-enzyme complex. Atomic coordinates from the 3S7F PDB crystallographic structure [Ferguson *et al.*, 2011] show the SMYD2 enzyme (blue) bound to its p53 peptide substrate (orange). Two orientations are depicted (left versus right) with transparent SMYD2 surfaces showing the SAM cofactor (red) contained inside SMYD2.

In this chapter, we developed a computational approach based on MSD to predict the substrate recognition space of SMYD2. We first generated *in silico* an ensemble of 180 unique backbone configurations of SMYD2 bound to a p53-derived peptide (PDB ID 3S7F) [Figure 5.1] to be used as our MSD input models. This ensemble was prepared using the coordinate perturbation and energy minimization (PertMin) algorithm that we previously developed [Davey and Chica, 2014]. PertMin involves the random perturbation of atomic coordinates followed by an energy minimization procedure that forces the calculation to adopt divergent descent trajectories resulting from the small coordinate changes to the input structure. Thus, PertMin results in an ensemble of similar protein structures found at alternate local minima that simulate conformational flexibility.

Specifically, atomic coordinates for the p53-bound structure of SMYD2 [Ferguson *et al.*, 2011] were retrieved from the PDB (PDB ID 3S7F). Hydrogens were added using the Protonate3D [Labute, 2009] utility found in the MOE software suite [MOE, 2012]. A 180-member PertMin ensemble [Davey and Chica, 2014] of the SMYD2-p53 complex was generated from the crystal structure with hydrogens added by randomly perturbing the coordinates of all atoms by ± 0.001 Å along each Cartesian coordinate axis. Because coordinate deviation from crystal structure increases proportionately with the number of minimization iterations, we energy minimized six sets of thirty perturbed structures using a truncated Newton minimization algorithm for 10, 25, 50, 100, 150 and 250 iterations, resulting in structures displaying a broad range of deviations from the crystal

structure. The energy minimizations were performed using the AMBER99 force field [Wang *et al.*, 2000] with a distance dependant dielectric of 80.

As shown in Figure 5.2, the resulting p53 peptide conformations in the PertMin ensemble adopt the same relative orientation in the SMYD2 binding cleft and occupy a tight conformational space. These conformations also preserve the interactions that the p53 sidechains [Figure 5.3.A] and backbone [Figure 5.3.B] make with the SMYD2 binding cleft. Inspection of the ensemble also shows that structural variation of the p53 peptide backbone is lower at positions -1, +1 and +2 relative to the K370 methylation site, suggesting that they dictate recognition through stronger interactions. We thus retained these positions for further design calculations.

Next, we used MSD to optimize each amino acid substitution with the exception of proline at positions -1, +1 and +2 of the p53 peptide in the context of the ensemble, evaluating a total of 57 substitutions. MSD of the SMYD2-p53 complex was conducted using the FASTER algorithm for sequence optimization [Allen and Mayo, 2006]. All amino acids with the exception of proline were introduced at p53 positions -1, +1 and +2 relative to the methylated lysine. Adjacent SMYD2 residues found in the binding cleft were allowed to sample alternate conformations during the design, but their identities were not modified (SMYD2 residues 19, 105, 108, 135, 179, 180, 181, 182, 184, 187, 191, 193, 196, 211, 215, 217, 238, 239, 240, 258, 344, 379 and 380, as well as the methylation target lysine belonging to the p53 peptide).

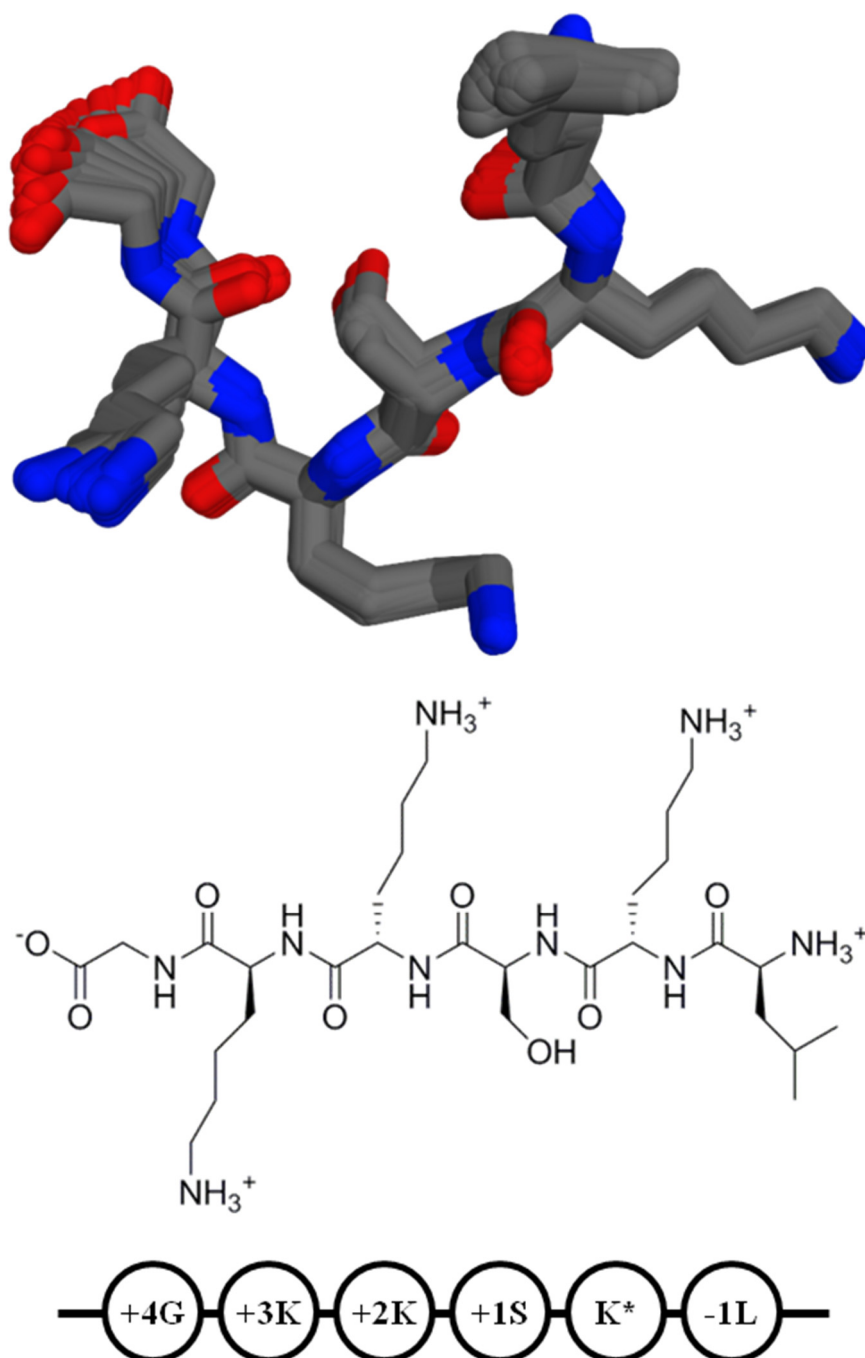


Figure 5.2. The PertMin ensemble of the p53 peptide. A superposition of ensemble members for the p53 peptide showing heavy atoms: carbon (grey), oxygen (red), and nitrogen (blue). In the crystallographic structure [Ferguson *et al.*, 2011] the p53 peptide is incomplete. The peptide is depicted C-terminus to N-terminus (left-to-right), with residues numbered according to their position relative to the target methylated lysine (K*).

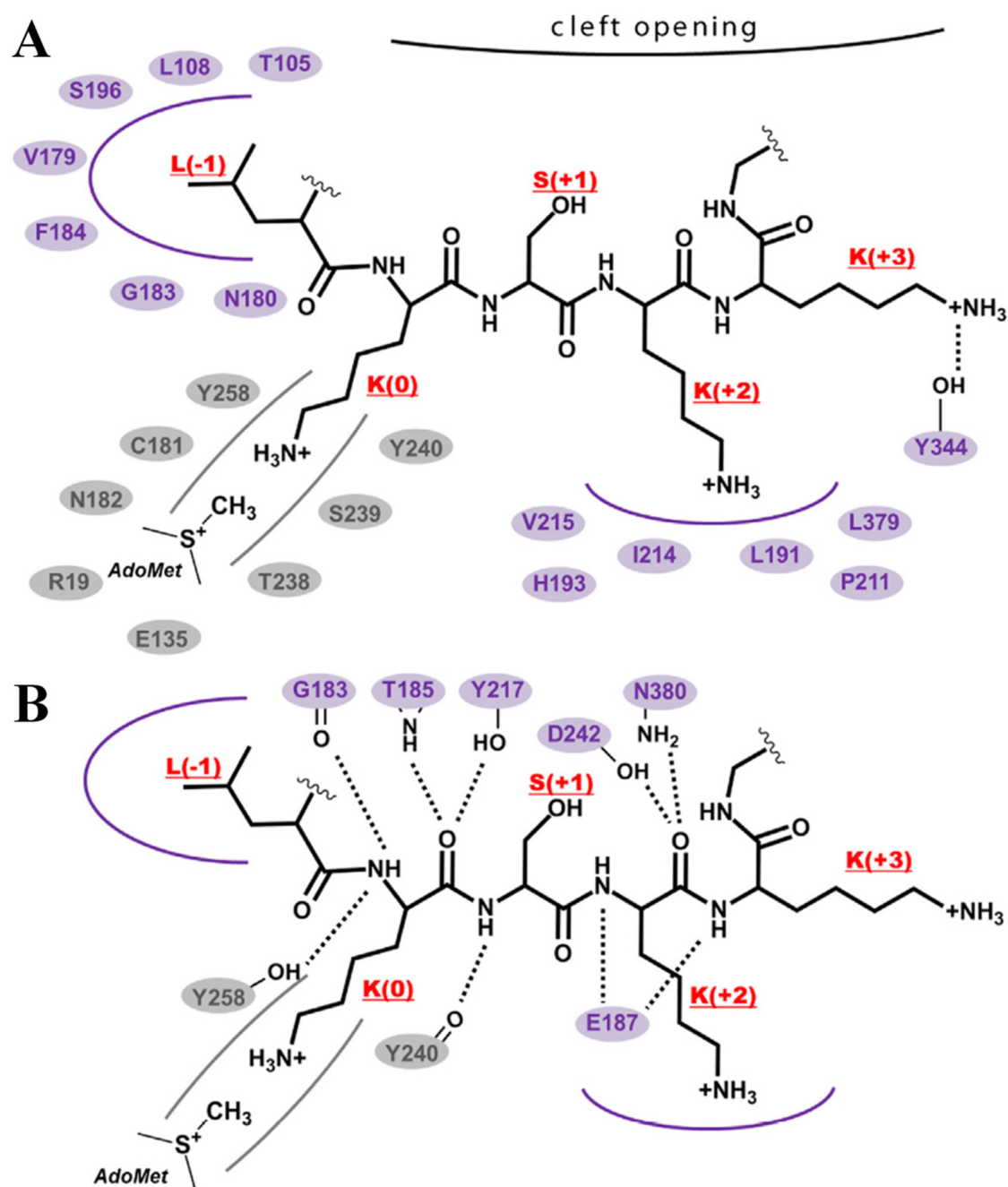


Figure 5.3. p53-SMYD2 peptide-enzyme interactions. The p53 peptide is illustrated as a stick model, with surrounding SMYD2 enzyme residues shown in ovals, and their interactions in dotted lines. SMYD2 residues belonging to the binding cleft are colored in purple and residues belonging to the methyl transfer channel are colored in grey. Interactions are shown between the SYMD2 enzyme with the p53 peptide sidechain (A) and backbone (B) atoms. Two hydrophobic pockets stabilize the -1L and +2K sidechains of the p53 peptide while no interactions stabilize the sidechain of residue +1S. This figure was provided courtesy of Sylvain Lanouette [Lanouette *et al.*, 2015].

The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around side chain χ_1 and χ_2 rotatable bonds was used [Dunbrack and Cohen, 1997]. The interaction energies of amino acid substitutions were scored using a four-term potential energy function that includes a van der Waals term from the Dreiding II force field where atomic radii were scaled by 0.9 [Mayo *et al.*, 1990], a hydrogen-bond term with a well depth of 8 kcal·mol⁻¹ and direction specific variables [Dahiyat and Mayo, 1997], a Coulomb electrostatic term with a distance dependent dielectric of 40 to approximate a protein-like environment, and a surface area based solvation penalty term [Lazaridis and Karplus, 1999; Street and Mayo, 1998]. MSD was implemented using PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012].

To predict the tolerated amino acid substitutions at each designed position, we calculated the fitness of each substitution by computing the Boltzmann weighted average of its interaction energy with the SMYD2-p53 complex in the context of each member of the ensemble. These fitness values reflect the stability of each mutation on the WT p53 peptide sequence, uncoupled from sequence perturbations at alternate positions along the peptide. As a result, the fitness value does not directly reflect binding affinity since the change in free energy (bound vs. unbound) for each peptide is not computed. Instead, the fitness value reports on the ability of each substitution to stabilize the SMYD2-p53 complex while preserving the binding mode represented in the ensemble.

Table 5.1. p53-SMYD2 interaction energy calculated by MSD

Substitution	Position Boltzmann Weighted Average Energy (kcal·mol ⁻¹)		
	-1	+1	+2
A	-62.39	-84.32	-65.32
C	-72.62	-86.53	-75.70
D	-64.14	-64.49	-52.54
E	-69.06	-72.43	-58.71
F	-92.88	-79.09	-61.56
G	-59.27	-73.00	-51.82
H	-79.14	-63.85	-72.19
I	-79.54	-16.60	-80.28
K	-91.28	-86.08	-84.64
L	-88.03	-75.05	-82.41
M	-90.02	-86.25	-82.63
N	-65.45	-65.05	-52.95
Q	-68.71	-75.06	-60.73
R	-71.98	-75.82	-36.08
S	-65.18	-81.27	-55.10
T	-75.20	-73.80	-61.75
V	-74.46	-51.40	-75.58
W	-78.52	-69.00	62.68
Y	-77.50	-76.63	140.53

As shown in Table 5.1, all substitutions with the exception of W and Y at position +2 result in negative and thus favourable fitness values, suggesting that they potentially stabilize the complex in the binding mode represented by the ensemble. To determine which residues to include in the predicted recognition motif, we clustered all substitutions at each designed position according to the similarity of their fitness values using the k-means cluster analysis algorithm [Kanungo *et al.*, 2002]. The k-means algorithm is an iterative method of clustering numerical values by partitioning them

according to their Euclidian distance from each cluster centroid, which in this case is the average fitness value of each substitution belonging to the cluster.

We partitioned each substitution with a negative fitness value into one of three clusters [Table 5.2], with cluster k1 containing the most favourable substitutions and cluster k3 containing the least favorable substitutions. A substitution was predicted to be part of the SMYD2 substrate recognition motif [Figure 5.4] if it is grouped in a cluster of equal or better fitness to the cluster containing the WT p53 amino acid ($L_{-1}/K_0/S_{+1}/K_{+2}$) at that position.

Table 5.2. k-Means cluster analysis of substitution interaction energy

Cluster (k) Substitutions ^a	
Position	Fitness ^b : k1 < k2 < k3
-1	k1: F
	k2: K M <u>L</u>
	k3: C I H R T W V Y A E D G N Q S
+1	k1: A C F K M <u>S</u>
	k2: E D G H L N Q R T W Y
	k3: I V
+2	k1: I <u>K</u> M L
	k2: A C H V
	k3: E D G F N Q S R T

^a The WT p53 residue at each position is underlined and in bold

^b Clusters ranked higher have more negative fitness values (kcal·mol⁻¹)

Thus, residues found in the k1 cluster of positions +1 and +2 as well as those from the k1 and k2 clusters of position -1 were included in the recognition motif [LFMK]₋₁-K*- [AFMSKC]₊₁-[KLIM]₊₂ (where K* is the methylated lysine) predicted by MSD. Interestingly, our k-means analysis showed that position +1 is more tolerant to substitution as more amino acids are included in the top cluster than that of positions -1 and +2. This result is in agreement with available crystal structures of SMYD2 in complex with Estrogen Receptor α and p53 which show that the side chain of the residue at position +1 is oriented outward of the binding cleft [Ferguson *et al.*, 2011; Jiang *et al.*, 2014], likely allowing for a broader range of substitutions by enabling a higher number of rotamers to be scored favourably at this position.



Figure 5.4. The SMYD2 recognition motif predicted by MSD. The motif is depicted N-terminal to C-terminal, left-to-right, for positions -1, 0, +1, and +2, relative to the target lysine (K*). Residues are colored according to their chemical property: hydrophobic (green), charged (blue), and polar (black). This figure was provided courtesy of Sylvain Lanouette [Lanouette *et al.*, 2015].

5.3. Prediction of the SMYD2 Recognition Motif by SPOT Peptide Array

To validate the SMYD2 recognition motif predicted by MSD, *Lanouette* and *Couture* produced three SPOT peptide arrays [Hilpert *et al.*, 2007; Winkler and Campbell, 2008] consisting of peptides with single mutant substitutions to the p53 sequence surrounding the SMYD2 methylation target lysine K370 (SSHLK₃₇₀SKKGQ). Three identical arrays were incubated with recombinant SMYD2 and tritiated SAM under conditions which reproduce SMYD2 specificity. To ensure that experimental array incubation conditions appropriately reflect the endogenous biological methylation activity of SMYD2, the specificity of SMYD2 against a selection of control peptides was evaluated under various buffer conditions [Figure 5.5].

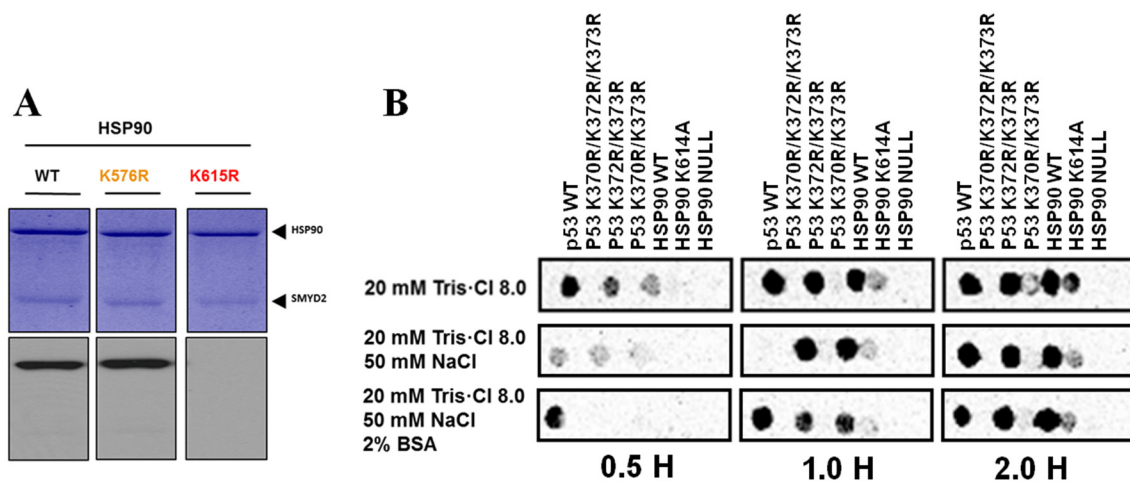


Figure 5.5. SPOT assay recapitulates the methylation activity of SMYD2. (A) Autoradiography (grey) of WT and mutant (K576R and K615R) HSP90 incubated with SMYD2 and tritiated SAM demonstrates that SMYD2 selectively methylates K615. SMYD2 and HSP90 gel mobility is identified by SDS-PAGE (blue). (B) Assay conditions and reaction incubation time length were optimized against p53 and HSP90 WT and mutant peptides. Autoradiography of the mutant K372R/K373R p53 peptide identified complete

methylation with respect to the WT peptide. Partial methylation activity was observed for the K614A mutation to the HSP90 peptide relative to its WT. Loss of methylation activity for the remaining peptides was observed as expected. Data provided courtesy of Dr. Sylvain Lanouette.

The protocol for preparation and assay of the SPOT peptide arrays was designed by *Lanouette* and *Couture*. Candidate substrate peptides were synthesized and immobilized (C-terminal) on a cellulose membrane support using the SPOT synthesis methodology with a 4,7,10-trioxa-1,13-tridecanediamine linker. Prior to incubation with SMYD2, the SPOT peptide arrays were primed thrice with 95% ethanol and three times with Tris buffered saline-Tween 0.1%. Following the priming procedure, the array was incubated over the course of an hour at 30 °C in the reaction buffer consisting of pH 8.0 50 nM Tris, 50 mM NaCl, 2% bovine serum albumin, 1 µg/ml recombinant SMYD2, 7.5 µM SAM, and 2.5 µM tritiated SAM (mCi/ml). Reaction incubation was halted by rinsing the membrane four times with solution of 8 M urea, 1% sodium dodecyl sulfate, and 0.5% β-mercaptoethanol. Termination of the reaction was concluded by three consecutive washings with solutions of 50% ethanol, 10% acetic acid, and 95% ethanol.

Any peptides being recognized and subsequently methylated by SMYD2 could be thus identified by autoradiography [Figure 5.6]. Methylation activity can be reported as a heat map by calculating the normalized averaged intensities for each peptide. The heat map shown in Figure 5.7 indicates that positions -1 to +2 surrounding K370 are the main determinants of SMYD2 activity. In contrast, positions -4 to -2 and +3 to +5 show

the highest tolerance to substitutions, validating our choice of positions designed by MSD.

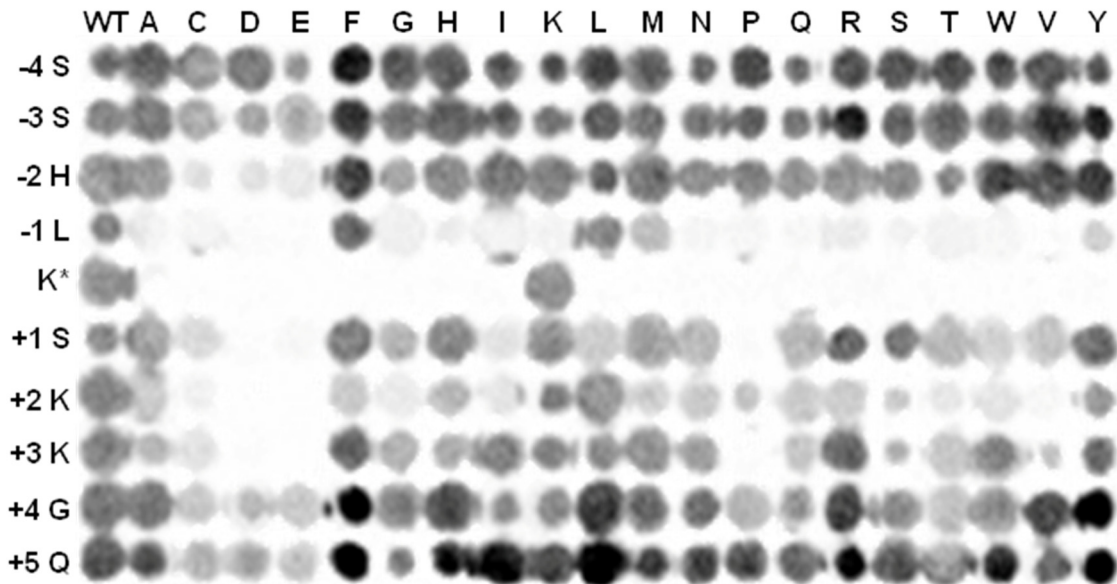


Figure 5.6. Autoradiography of an example SPOT peptide array. This example array contains autoradiography results for two-hundred WT and single-mutant peptides originating from p53 organized by amino acid substitution (column) and position (row) relative to the target lysine K*. Methylation activity is indicated by autoradiographic intensity, with more intense the spot signal (black) corresponding to increased methyl transfer. SPOT peptide array data provided courtesy Dr. Sylvain Lanouette.

To precisely define the SMYD2 substrate recognition motif, *Lanouette* and *Couture* calculated the relative methylation factor for each amino acid as the ratio of methylation for a given substitution relative to the difference of its methylation and the average methylation of all other substitutions at this position. This relative methylation factor thus reflects the preference of SMYD2 for a specific substitution at one position relative to all other substitutions.

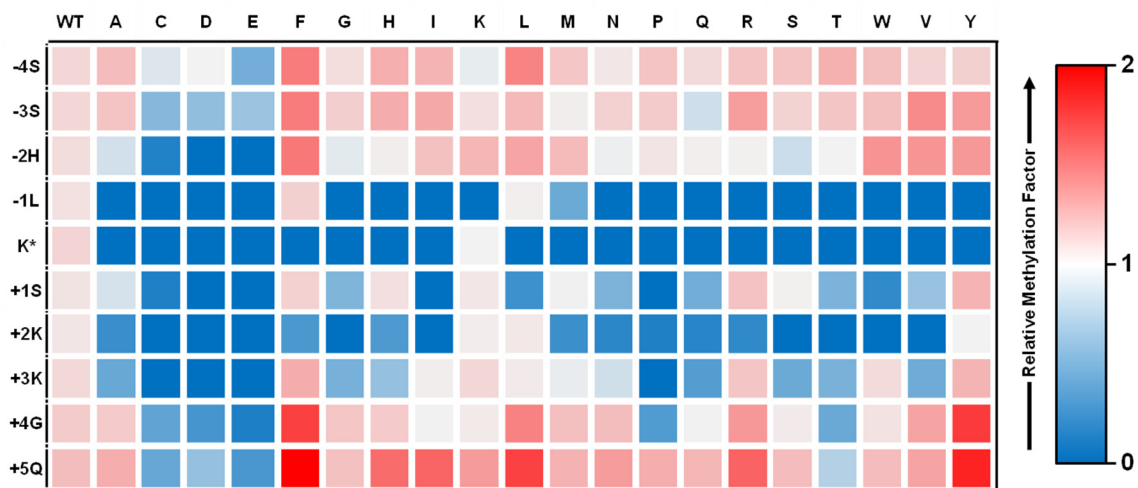


Figure 5.7. Heat map representation of autoradiographic SPOT array results. Average methylation intensities are reported for substitutions (column) to each position (row) of the p53 substrate peptide. The intensities are reported as relative methylation factors (RMF) with substitutions having no methylation activity (RMF = 0, blue), average activity (RMF = 1, white), and increased activity (RMF ≥ 2 , red).

Consistent with MSD predictions, we observed that SMYD2 exhibits a strong preference for L, F and M residues at position -1 [Figure 5.7]. In contrast, position +1 has a broad tolerance with 8 out of 20 substitutions allowing preferential methylation. This result is consistent with the solvent exposed orientation of side chains at position +1 found in the PertMin ensemble. Similar to residues found at position -1, +2 residues remain in a deep pocket within the SMYD2 peptide binding cleft [Figure 5.3]. Accordingly, SMYD2 shows narrow specificity for peptides containing K, L or Y substitutions at this position. Analysis of the relative methylation factors allows for construction of an experimentally defined recognition motif [Figure 5.8] and demonstrates that SMYD2 preferentially methylates the motif [LFM]₋₁-K*-[AFYMSHRK]₊₁-[LYK]₊₂.

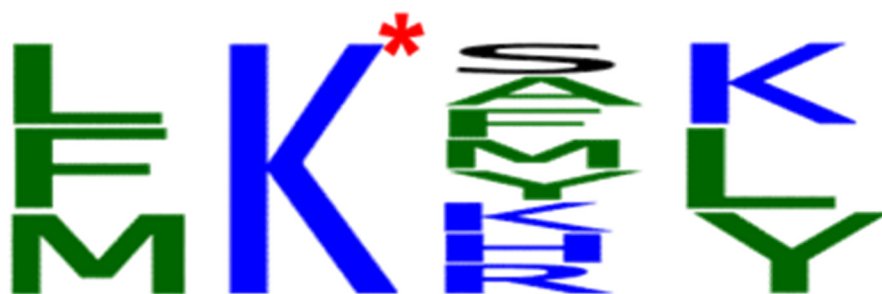


Figure 5.8. The SMYD2 recognition motif predict by SPOT peptide array. The motif is depicted N-terminal to C-terminal, left-to-right, for positions -1, 0, +1, and +2, relative to the target lysine (K*). Residues are colored according to their chemical property: hydrophobic (green), charged (blue), and polar (black). This figure was provided courtesy of Dr. Sylvain Lanouette [Lanouette *et al.*, 2015].

5.4. Calculation Comparison with Experiment

The recognition motif predicted using MSD [Figure 5.4] is in excellent agreement with results from the high-throughput methyltransferase assays [Figure 5.8]. Our *in silico* model accurately predicted that SMYD2 recognition is predominantly determined by amino acid identity at positions -1, +1 and +2 of the substrate peptide. The narrow specificity of SMYD2 for substitutions at position -1 observed in SPOT arrays is correctly identified by MSD, as position -1 includes only one false positive substitution (K) in the k1 and k2 clusters considered [Table 5.2]. SMYD2 methyltransferase activity shows a strong bias toward long side chain hydrophobic residues L, F and M at position -1, possibly due to the stabilization of these side chains in a hydrophobic pocket [Figure 5.3] [Ferguson *et al.*, 2011]. MSD also correctly predicted relaxed specificity at position +1 [Table 5.2], with cluster k1 recapitulating five of the eight substitutions observed to be

favourable *in vitro* [Figure 5.8]. Interestingly, Y at position +2 is one of the most unfavourable substitutions predicted by MSD [Table 5.1] whereas it is included in the recognition motif derived from the SPOT array experiments. This discrepancy may be due to the absence of an adequate p53 peptide backbone conformation in our PertMin ensemble required to favourably score this substitution, an issue that could be potentially alleviated by using ensembles produced from structures bearing peptide substrates in alternate binding orientations. Available crystal structures of SMYD2 complexes [Ferguson *et al.*, 2011; Jiang *et al.*, 2014] show that this residue is embedded in a deep hydrophobic pocket [Figure 5.3] that accommodates residues with long side chains, an observation consistent with the motif defined by MSD and the SPOT arrays. While our methyltransferase assays directly assessed the activity of SMYD2 toward peptide substrates, MSD assessed the interaction energy of substitutions in its binding cleft. The agreement of both methods suggests that association may be the main driver of SMYD2 specificity.

While there is remarkable agreement between recognition motifs predicted by both MSD and peptide arrays, there are four false negative (H/R/Y and Y at positions +1 and +2, respectively) and four false positive (K, C and I/M at positions -1, +1 and +2, respectively) substitutions predicted by MSD [Table 5.3]. To evaluate whether an alternate binning method could have resulted in fewer false positives and negatives, we used the fitness value of the WT at each designed position [Table 5.1] as a cut-off to include or exclude substitutions from the recognition motif predicted by MSD.

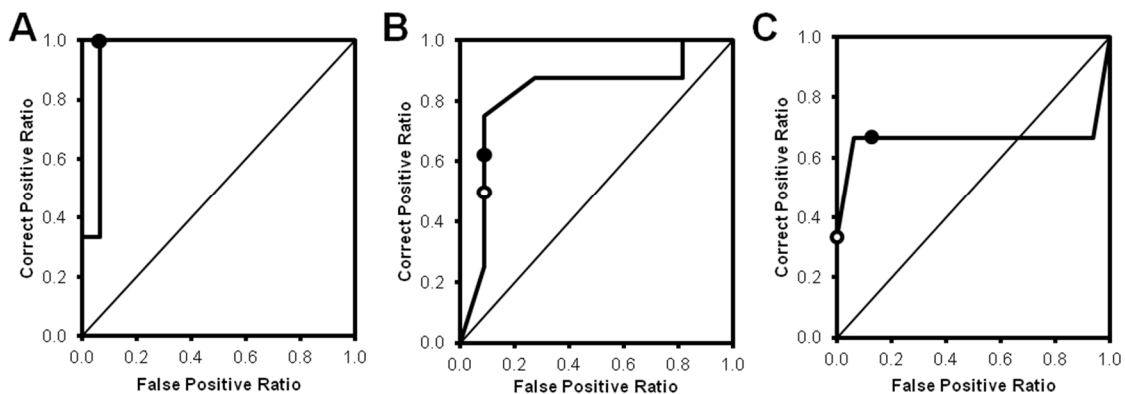


Figure 5.9. Sequence binning the MSD predicted recognition motif. ROC curves for all substitutions to positions -1 (A), +1 (B), and +2 (C) relative to the methylated lysine. The true and false positive ratios obtained by using the WT Boltzmann weighted average energy as a cut-off or by k-means cluster analysis are shown as white and black circles, respectively. For position -1, both binning methods yielded identical true positive and false positive ratios. The diagonal line represents random binning.

To assess whether MSD provided improved prediction accuracy compared to SSD, we performed SSD on a single fixed backbone using identical parameters as those used for MSD except that single-state scores were used to rank sequences instead of the Boltzmann weighted average fitness value. Figure 5.10 shows a comparison of the SSD scores and MSD fitness values computed for all substitutions except for W and Y at position +2, which were omitted because both their MSD fitness values and SSD scores are $> 0 \text{ kcal}\cdot\text{mol}^{-1}$. In all cases, regardless of substitution or position, MSD fitness values were more negative than SSD scores, indicating that substitutions are scored more favourably in the context of a backbone ensemble rather than a single structure. In the case of 10 substitutions (F/H/I/L/M/V/W/Y and F/I at positions +1 and +2, respectively), SSD scores fell above our accepted threshold of $0 \text{ kcal}\cdot\text{mol}^{-1}$. The high SSD scores for these 10 substitutions indicate unfavourable interactions likely resulting from steric

clashes. This result is consistent with a known artefact of SSD arising from the combined use of a fixed protein backbone template with rigid rotamers, which can lead to the rejection of favourable amino acid sequences that would have been accepted on a slightly different backbone geometry or with a slightly different rotamer configuration [Choi *et al.*, 2009].

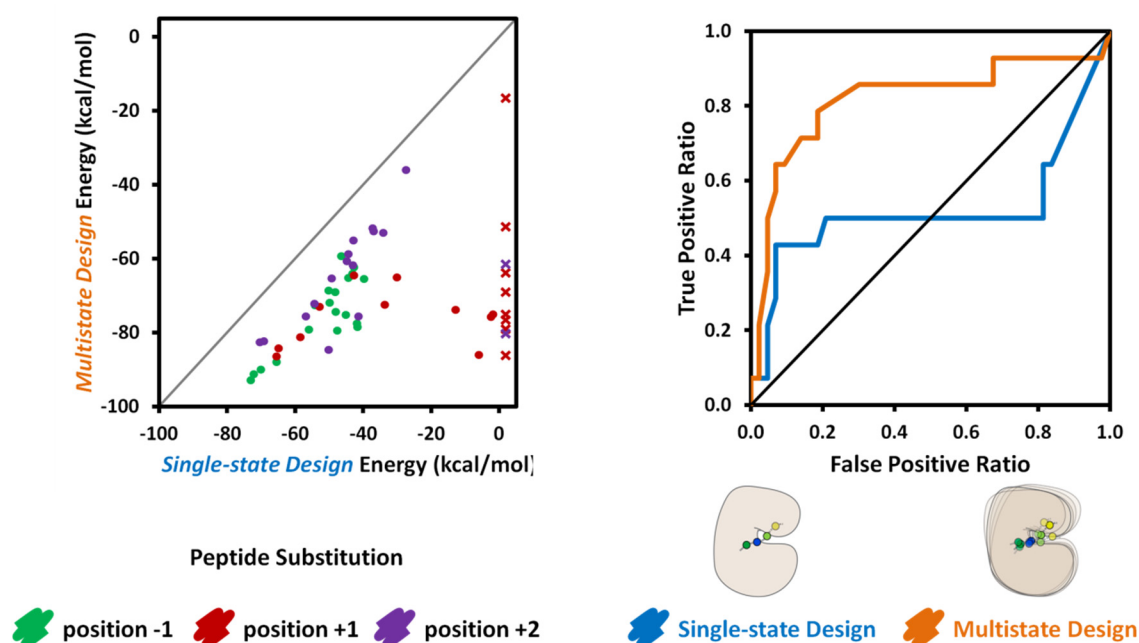


Figure 5.10. Comparison of motif predictions by SSD and MSD. (A) The MSD Boltzmann weighted average energy for each of 55 substitutions to the three design positions: -1 (green), +1 (red), and +2 (purple) of the p53 peptide are plotted as a function of SSD sequence energies. The 45 substitutions that resulted in negative (favorable) energy values are shown in circles. Evaluation of the 10 substitutions that resulted in negative (favorable) MSD fitness values and positive (unfavorable) SSD scores are shown as X-marks. To simplify the plot, these 10 sequences are given identical arbitrary positive SSD scores. (B) ROC analysis for all sequence binning of all motif substitutions predicted by SSD (blue) and MSD (orange). The diagonal line represents random binning.

To further compare the prediction accuracy obtained by SSD and MSD, we calculated the true positive (fraction of true positives out of the positives) and false

positive (fraction of false positives out of the negatives) ratios for every possible cut-off value at 1 kcal·mol⁻¹ increments in the range of score/fitness values obtained for all substitutions, and plotted them to generate ROC curves [Figure 5.10]. ROC curves show that a significant improvement in prediction accuracy is obtained by using MSD as evidenced by the larger area under the curve that indicates a higher probability of ranking a randomly chosen positive higher than a randomly chosen negative. The lower prediction accuracy of SSD results from its poor scoring of substitutions to position +1, which results in four additional false negatives (F, H, M, and Y). This incorrect rejection of true positives decreases the true positive ratio to ≈0.5, a value that cannot be improved over a broad range of cut-offs that instead only increase the false positive ratio. In contrast, predictions made by MSD result in a large increase to the true positive ratio that is accompanied by a small increase to the false positive ratio. It is likely that similar improvements to prediction accuracy could also be achieved through the use of alternate CPD methodologies which incorporate backbone flexibility during calculation [Murphy *et al.*, 2012; Smith and Kortemme, 2008]. Although predictions could not be improved by using SSD or an alternate binning method, MSD combined with k-means cluster analysis correctly binned substitutions as either being part or not of the recognition motif elucidated by SPOT array analysis with an overall accuracy of 86% (49/57), demonstrating the utility of this approach to the prediction of substrate recognition motifs for PKMTs.

5.5. Discovery of New SMYD2 Substrates

We next sought to determine whether the recognition motifs could be used to discover novel SMYD2 methylation targets. We probed a dataset of all reported genetic and physical SMYD2 protein interactors [Abu-Farha *et al.*, 2008; Abu-Farha *et al.*, 2011; Brown *et al.*, 2006; Cho *et al.*, 2012; Donlin *et al.*, 2012; Huang *et al.*, 2006; Huang *et al.*, 2007a; Saddic *et al.*, 2010; Voelkel *et al.*, 2013] using the ScanProsite motif search tool [de Castro *et al.*, 2006]. Among our dataset, 95 SMYD2 interactors include the peptide array motif [LFM]₋₁-K*-[AFYMSHRK]₊₁-[LYK]₊₂ and 135 include the MSD motif [LFMK]₋₁-K*-[AFMSKC]₊₁-[KLIM]₊₂ [Table 5.3]. As expected from their similarity, the recognition motif predicted by MSD identified a majority (77%) of the SMYD2 putative methylation targets identified using the motif determined by SPOT array.

To refine our list of possible substrates and increase the probability of identifying genuine substrates, SMYD2 relevant gene ontology terms, including expression patterns and biological activities, were cross-referenced against 95 probable SMYD2 interactors containing the recognition motif LFM-K*-AFYMSHRK-LKY [Figure 5.11]. This list includes a compilation of all reported physical and genetic interactors of SMYD2 [Abu-Farha *et al.*, 2008; Abu-Farha *et al.*, 2011; Brown *et al.*, 2006; Cho *et al.*, 2012; Donlin *et al.*, 2012; Huang *et al.*, 2006; Huang *et al.*, 2007a; Saddic *et al.*, 2010; Voelkel *et al.*, 2013].

Table 5.3. SMYD2 interactors containing the MSD and SPOT recognition motifs

MSD & SPOT		MSD		SPOT	
SWISSPROT	UNIPROT	SWISSPROT	UNIPROT	SWISSPROT	UNIPROT
1433T	P27348	ABCCB	Q96J66	ADRB2	P07550
AHI1	Q8N157	AKA11	Q9UKA4	ARX	Q96QS3
ACAD9	Q9H845	APC	P25054	BMP5	P22003
ADA30	Q9UKF2	ARSI	Q5FYB1	CC110	Q8TBZ0
APX	Q96QS3	AT1A4	Q13733	CCD25	Q86WR0
ASPM	Q8IZT6	B4GN2	Q8NHY0	CCL7	P80098
ATD2B	Q9ULI0	CLAP2	O75122	CGT	Q16880
BANK1	Q8NDB2	CN166	Q9Y224	COF1	P23528
BPTF	Q12830	CO8A	P07357	DDX1	Q92499
BTF3	P20290	COBL	O75128	DDX5	P17844
CQO80	Q9BSJ5	CP3A4	P08684	DMD	P11532
CROCC	Q5TZA2	CP46A	Q9Y6A2	E2F1	Q01094
CTDP1	Q9Y5B0	DDX6	P26196	MK11	Q01094
CTR2	P52569	DLEC1	Q9Y238	NCTR2	O95944
DDX55	Q8NHQ9	DLGP1	O14490	PTPS	Q03393
DESP	P15924	DOT1L	Q8TEK3	SHRM3	Q8TF72
DHX15	O43143	DYH5	Q8TE73	SIX4	Q9UIU6
EVPL	Q92817	E2F5	Q15329	STIP1	P31948
F111A	Q96PZ2	E41L3	Q9Y2J2	UN45B	Q8IWX7
FANCM	Q81YD8	E41LA	Q9HCS5	UNC5A	Q6ZN44
FAT2	Q9NYQ8	FER	P16591	YSK4	Q56UN5
FRPD3	Q5JV73	GLT13	Q81UC8		
FRYL	O94915	HECW1	Q76N89		
GNL3	Q9BVP2	HS90A	P07900		
GRP78	P11021	HS90B	P08238		
GTR1	P11166	IF2B1	Q9NZI8		
HNRPL	P14866	IF2B3	O00425		
HNRPU	Q00839	IPO8	O15397		
IQCG	Q9H095	K1210	Q9ULL0		
JAK2	O60674	K1F11	P52732		
K220	Q01546	LONF2	Q11529		
KNTC1	P50748	LSP1	P33241		
LIPA4	O75335	MACOI	Q8N5G2		
MANEA	Q5SRI9	MAMC2	Q7Z304		
MYO5A	Q9Y411	MCM5	P33992		
MYO7A	Q13402	MCM7	P33993		
NFL	P07196	MYH15	Q9Y2K3		
NUAK2	Q9H093	NAL12	P59046		
OCRL	Q01968	PAPOG	Q9BWT3		
OST48	P39656	PKP4	Q99569		
PARP1	P09874	PLXB2	O15031		
PCM1	Q15154	PTBP1	P26599		
PDIP3	Q9BY77	PRPRK	Q15262		
PI4KA	P42356	RBM34	P42696		
PLD1	Q13393	RHG28	Q9P2N2		
PRKDC	P78527	RL27	P61353		
REXON	Q96IC2	RPC6	Q9H1D9		
RIF1	Q5UIP0	RPP30	P78346		
RIOK1	Q9BR52	RTF1	Q92541		
SCML2	Q9UQR0	RU2B	P08579		
SF3B1	O75533	S10A2	P29034		
SIN3A	Q96ST3	SHAN3	Q9BYB0		
SIN3B	O75182	SMC4	Q9NTJ3		
SIX1	Q15475	SMRCD	Q9H4L7		
SIX2	Q9NPC8	SPSY	P52788		
SPTB2	Q01082	SPTCS	Q96JI7		
STAG1	Q8WVM7	SRPK1	Q965B4		
STAM1	Q92783	TAF2	Q6P1X5		
STAU1	Q95793	TBA4A	P68366		
TACC2	Q95359	THOC4	Q86V81		
TCPZ	P40227	TM110	Q86TL2		
TITIN	Q8WZ42	TNAP2	Q03169		
TLN2	Q9Y4G6	WDR7	Q9Y4E6		
TMEM5	Q9Y2B1	ZN148	Q9UQR1		
TOPRS	Q9NS56	ZN276	Q8N554		
TTK	P33981				
USP9X	Q93008				
WWOX	Q9NZC7				
ZFHX4	Q86UP3				

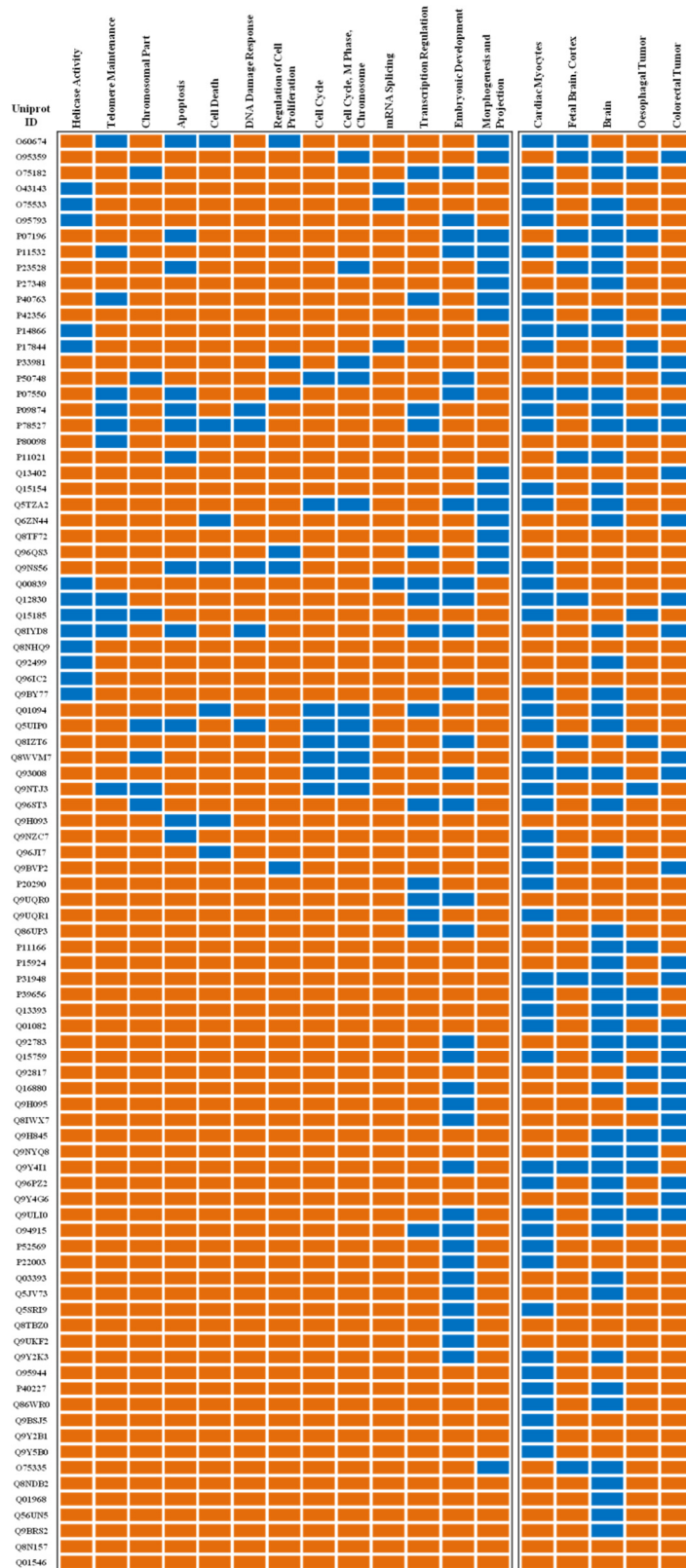


Figure 5.11. SMYD2 interactors and selected gene ontology terms. Correspondence of SMYD2 interactors with role and molecular function (left), as well as localization and

disease (right). Gene ontology terms cross referenced with SMYD2 interactors are indicated in blue, while those absent are indicated in orange.

The dataset was filtered using the DAVID Bioinformatics Resources Functional Annotation Tool [Huang *et al.*, 2009] to assess the enrichment of gene ontology terms. Putative targets sharing known roles with SMYD2 such as the regulation of cell cycle, as well as cell proliferation [Brown *et al.*, 2006; Cho *et al.*, 2012; Saddic *et al.*, 2010], gene transcription [Abu-Farha *et al.*, 2008; Brown *et al.*, 2006; Cho *et al.*, 2012; Saddic *et al.*, 2010], and apoptosis and cell death [Huang *et al.*, 2006]. Additional putative targets with biological activities related to mRNA splicing, helicase activity, as well as chromosome segregation, chromosomal part, telomere maintenance and DNA damage response were also included [Abu-Farha *et al.*, 2011] as well as those displaying higher levels of expression in fetal brain and cortex, brain and cardiac myocytes [Brown *et al.*, 2006; Diehl *et al.*, 2010], and during embryonic development, morphogenesis and projection [Diehl *et al.*, 2010; Kawamura *et al.*, 2008; Sese *et al.*, 2013]. To further refine and reduce the list of putative SMYD2 targets, protein sequences were searched for conservation of the target recognition motif lysine in humans (*Homo sapeins*), mice (*Mus musculus*), frogs (*Xenopus laevis*), chicken (*Gallus gallus*), and fish (*Danio rerio*) [Figure 5.12].

Uniprot ID	Helicase Activity	Telomere Maintenance	Chromosomal Part	Apoptosis	Cell Death	DNA Damage Response	Regulation of Cell Proliferation	Cell Cycle	Cell Cycle, M Phase, Chromosome	mRNA Splicing	Transcription Regulation	Embryonic Development	Morphogenesis and Projection	Cardiac Myocytes	Fetal Brain, Cortex	Brain	Oesophageal Tumor	Colorectal Tumor	K* Species Conservation
O60674																			
O75182																			
O43143																			
O95793																			
P11532																			
P17844																			
P07550																			
P78527																			
Q6ZM44																			
Q9N856																			
Q15185																			
Q8IYD8																			
Q92499																			
Q9BY77																			
Q01094																			
Q5UIP0																			
Q8WVM7																			
Q96ST3																			
Q15759																			
Q8IWX7																			
Q9ULI0																			
Q9Y5B0																			

Figure 5.12. Refined list of SMYD2 interactors. Of the list of potential SMYD2 substrates, 11 share (blue) conservation of the recognition motif target lysine (K* Species Conservation, right column) in humans (*Homo sapeins*), mice (*Mus musculus*), frogs (*Xenopus laevis*), chicken (*Gallus gallus*), and fish (*Danio rerio*) while 11 candidates do not (orange).

Of the selected putative targets, 10 proteins were expressed as FLAG-tagged constructs and incubated with recombinant SMYD2 and ³H-labeled SAM. Autoradiography confirmed that SMYD2 methylates four of the ten tested transcription factors SIX1 (L₋₁-K*-A₊₁-K₊₂) [Heanue *et al.*, 1999] and SIX2 (L₋₁-K*-A₊₁-K₊₂) [Boucher *et al.*, 2000], the transcriptional co-repressor SIN3B (L₋₁-K*-A₊₁-K₊₂) [Ayer *et al.*, 1995] and the RNA helicase DHX15 (L₋₁-K*-K₊₁-L₊₂) [Wen *et al.*, 2008] [Figure 5.13].

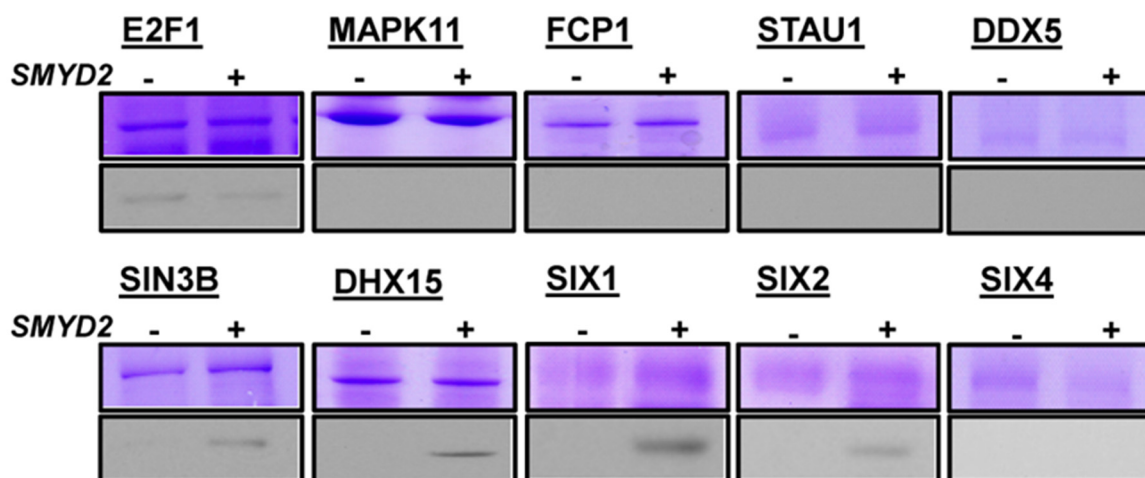


Figure 5.13. Discovery of four novel SMYD2 substrates. Autoradiography of FLAG-tagged proteins incubated with (+) and without (-) recombinant SMYD2 and ^3H -labelled SAM. Candidate substrates include: transcription factor E2F1, mitogen-activated protein kinase 11 (MAPK11), RNA polymerase II subunit A C-terminal domain phosphatase (FCP1), staufen 1 (STAU1), DEAD box RNA helicase 5 (DDX5), SIN3B, DEAD box RNA helicase 15 (DHX15), and sine oculis homeobox homolog 1, 2, and 4 (SIX1, SIX2, and SIX4). Coomassie stained SDS-PAGE indicates the presence of each substrate (top gel, blue) while autoradiography (bottom gel, grey) demonstrates that SIX1, SIX2, SIN3B, and DHX15 are substrates of SMYD2.

Furthermore, SMYD2 cannot methylate single-point mutants of SIN3B and SIX1 in which the predicted target lysine is substituted for arginine, confirming that our motif correctly predicts novel protein methylation sites [Figure 5.14]. Notably, SMYD2 did not methylate six of the proteins identified by gene ontology analysis which contain the recognition motif. It is likely that the putative target lysine of these proteins is either not solvent exposed or engaged in intramolecular interactions with other residues of the motif, preventing methylation by SMYD2. All novel SMYD2 substrates identified contain a recognition sequence that is included in both the SPOT array and MSD motifs, demonstrating the utility of MSD for the discovery of novel lysine methylation targets for a PKMT.

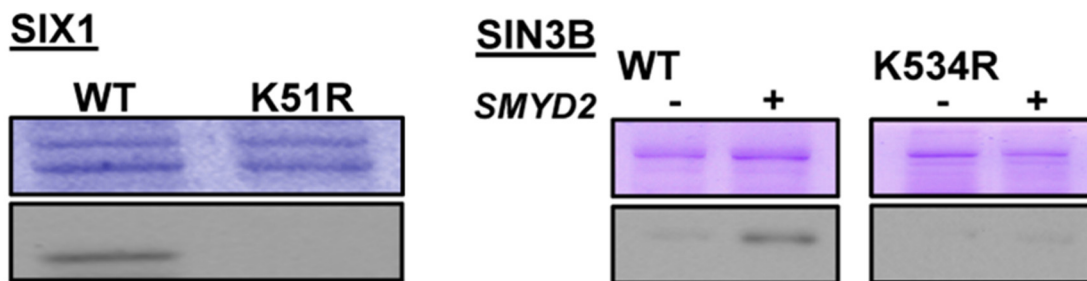


Figure 5.14. SMYD2 methylates the recognition motif of novel substrates. Mutation of the recognition motif target lysine of SIX1 and SIN3B to arginine results in a loss of methylation activity demonstrating that the recognition motif is the target site of lysine methylation. Coomassie stained SDS-PAGE indicates the presence of each substrate (top gel, blue) while autoradiography (bottom gel, grey) displays a loss of methylation activity with mutant SIX1 (K51R) and SIN3B (K534R).

The sequence surrounding the methylation site ($L_{-1}-K^*_{266}-H_{+1}-K_{+2}$) on the Estrogen Receptor α is included in our motif [Zhang *et al.*, 2013]. Accordingly, corresponding residues on this substrate peptide bind to SMYD2 in a conformation nearly identical to that seen in the SMYD2-p53 complex [Jiang *et al.*, 2014]. In contrast, our motif would not have predicted the methylation of HSP90 ($L_{-1}-K^*-A_{+1}-Q_{+2}$) and pRb ($L_{-1}-K^*-S_{+1}-P_{+2}$ and $L_{-1}-K^*-R_{+1}-S_{+2}$) by SMYD2. Most likely, the recognition sites of these protein substrates adopt alternate peptide conformations not probed by our PertMin ensemble or consist of divergent sequences not considered in the single-mutation space evaluated by MSD and not covered in our SPOT arrays. Alternatively, it is also possible that the recognition motif may extend beyond the peptide positions defined here. These observations suggest that substrates included in our motif bind in a conformation resembling the structure of p53 when bound to SMYD2 and that SMYD2 also recognizes substrates (HSP90 and Rb) adopting an alternative binding mode. These findings are consistent

with previous findings showing that SMYD2 is controlled by a specificity-switch enabling the methylation of distinct pools of substrates [Abu-Farha *et al.*, 2011].

5.6. Conclusions on the Application of Positive Multistate Design

We showed that SMYD2 methylates the sequence motif [LFM]₋₁-K*-[AFYM_{SHRK}]₊₁-[LYK]₊₂ with position -1 and +2 exhibiting the highest degree of selectivity. Combining different bioinformatics and biochemical approaches with the motifs defined by MSD and SPOT arrays, a novel set of substrates for SMYD2 was discovered. Interestingly, the methylation sites in SIX1, SIX2, and SIN3B are located in the protein-protein interaction domain. In SIX1 and SIX2, the methylation site K51 is located at the surface of the SIX1 domain of both transcription factors in a region known to interact with the EYA transcriptional activator [Patrick *et al.*, 2013]. In SIN3B, the methylation site is located in the interaction domain that contacts the adaptor protein SUDS3 [Alland *et al.*, 2002]. Methylation of p53, HSP90, pRb and ER α by SMYD2 was already shown to control the association of their binding partners [Cho *et al.*, 2012; Huang *et al.*, 2006; Jiang *et al.*, 2014; Voelkel *et al.*, 2013]; our findings suggest that this mechanism may be shared by numerous substrates of SMYD2. These four substrates also provide promising avenues to explore the critical role of SMYD2 in organ development [Diehl *et al.*, 2010; Donlin *et al.*, 2012] and genetic regulation [Abu-Farha *et al.*, 2008; Brown *et al.*, 2006].

High throughput methyltransferase assays showed that MSD correctly characterized the substrate sequence space recognized by SMYD2. In addition to biochemical characterization, CPD techniques can thus be used to study the spectrum of

specificity of other post-translational modifying enzymes. In contrast with current methods, MSD provides a rapid and inexpensive mean to probe the specificity of a post-translational modifying enzyme on the basis of its enzyme-substrate interface. This novel method combining multistate CPD and k-means cluster analysis is the first demonstration that CPD can be applied to the discovery of previously unknown substrates of PTM enzymes.

CHAPTER 6

Applied Negative Multistate Design: Predicting the Stability of Proteins and their Mutants

Accurate predictions of protein stability have great potential to accelerate progress in CPD, yet the correlation of predicted and experimentally determined stabilities remains a significant challenge. To address this problem, we have developed a computational framework based on negative MSD in which sequence energy is evaluated in the context of both native and non-native backbone ensembles. This framework was validated experimentally with the design of 10 variants of *Streptococcal* protein G domain $\beta 1$ that retained the WT fold, and showed a very strong correlation between predicted and experimental stabilities ($R^2 = 0.86$). When applied to four different proteins spanning a range of fold types, similarly strong correlations were also obtained. Overall, the enhanced prediction accuracies afforded by this method pave the way for new strategies to facilitate the generation of proteins with novel functions by CPD.

Some of the language and figures in this chapter were adapted from a research article coauthored with Dr. Roberto A. Chica.

Davey JA, Damry AM, Euler CK, Goto NK, and Chica RA (2015) Prediction of Stable Globular Proteins Using Negative Design with Non-native Backbone Ensembles. *Structure*, 23(11):2011-2021 [Davey *et al.*, 2015].

I would like to formally acknowledge the contributions of these authors to the completion of this chapter. Specifically:

James A Davey: Calculation design, predictions, and comparisons with experiment

Adam M Damry: Experimental results including NMR data

Christian K Euler: Experimental results excluding NMR data

Dr. Natalie K Goto: Experimental design

Dr. Roberto A Chica: Calculation and experimental design

6.1. Accurate Prediction of Protein Stability

In spite of the impressive advances made in CPD, it is still necessary for libraries of CPD-predicted sequences to be generated and subjected to resource-intensive experimental screening in order to identify functional variants. This is because a significant proportion of predicted sequences fail to produce a protein with the targeted properties [Chica *et al.*, 2010; Jiang *et al.*, 2008; Röthlisberger *et al.*, 2008; Stranges and Kuhlman, 2013], highlighting limitations in current CPD approaches that prevent the routine design of functional proteins.

One of the principal causes of prediction inaccuracies in CPD arises from difficulties in discriminating sequences with compromised stability from those that will fold stably into the intended structure, as seen by the numerous predictions that produce unfolded species [Allen *et al.*, 2010; Siegel *et al.*, 2010]. The occurrence of a significant number of unfolded variants in the top-ranked sequences predicted by CPD reflects a longstanding challenge in the accurate prediction of free energy differences between WT and mutant sequences [Allen *et al.*, 2010; Kellogg *et al.*, 2011; Yin *et al.*, 2007a]. This problem is exacerbated by the fact that it is not yet possible to efficiently compute the absolute free energy of a protein [Grigoryan, 2013], and that accurate atomistic models of unfolded conformational states are not routinely accessible. To work around these difficulties, approximations have been made in CPD simulations, such as the use of potential energy as a surrogate of free energy [Boas and Harbury, 2007; Gordon *et al.*, 1999] and amino acid reference state energies as implicit models of the unfolded state [Ali *et al.*, 2005; Pokala and Handel, 2005; Renfrew *et al.*, 2012].

Although these approaches have given rise to improvements in prediction accuracies, the correlation of protein stabilities predicted by CPD with experimentally determined values remains a significant challenge. Directed efforts to improve the correlation between predicted and experimental stabilities have shown some success, in particular through optimization of potential energy functions [Kellogg *et al.*, 2011; Pokala and Handel, 2005], systematic testing of conformational sampling methodologies [Kellogg *et al.*, 2011; Yin *et al.*, 2007a], and use of alternate input templates [Howell *et al.*, 2014]. However, none of these methods explicitly considers the unfolded state, resulting in the omission of critical information that comes from the energy contribution of the unfolded state ensemble in calculations of protein stability. In addition, none of these methods have yet been validated for stability predictions in newly-designed proteins. These limitations highlight the need for the development of more general CPD methodologies that can accurately predict protein stability for any type of protein fold and target function.

Herein, we develop and validate a generally applicable computational framework for the prediction of protein stability based on MSD, an emerging methodology in CPD that enables the explicit consideration of multiple conformational states during sequence optimization [Davey and Chica, 2012]. Previously, we showed that MSD with backbone ensembles approximating the native folded state (i.e. positive design) decreased the number of false negative predictions by helping to address the fixed-backbone approximation, but failed to provide strong correlation between predicted and experimental stabilities [Davey and Chica, 2014]. To improve the accuracy of

stability predictions by CPD, we developed a general strategy based on negative design, a procedure in which sequence energy is evaluated in the context of both native and non-native backbone ensembles intended to approximate the folded (positive) and unfolded (negative) states. Because the unfolded state is structurally heterogeneous and dynamic, the structural features of the unfolded state that are important for calculating protein stabilities are unknown. Therefore, several atomistic models of non-native states were prepared *in silico* and a training set of sequences was used to identify ensembles that could be utilized as a reference state for stability predictions in CPD. Using our negative design approach, we identified an ensemble pair capable of correlating calculated and experimental stabilities for a training set of G β 1 mutant sequences with an R^2 of 0.82 when used in negative design, a substantial improvement over positive design ($R^2 = 0.42$). To validate our computational framework, 10 new mutant sequences were designed and all were found to adopt the G β 1 structure with stabilities that were comparable to the WT protein. Most impressively, correlation of predicted and experimental stabilities was maintained, with $R^2 = 0.83$. We then extended our framework to the recapitulation of sequence stability for four alternate protein folds, and showed that negative design provides improved accuracy over positive design in all cases. To our knowledge, this work represents the first successful application of negative design with native and non-native backbone ensembles to the design of protein stability, providing a general approach that could be incorporated into a range of CPD strategies.

6.2. A State Independent Framework

Our goal was to develop a computational framework to accurately predict the stability of mutant sequences designed with CPD that would not require fine-tuning of the potential energy function or comprehensive evaluation of a broad range of conformational sampling methodologies. In addition, we wanted a computational framework that would discard unfolded species from the top-ranked predictions such that only sequences that adopt the intended fold were included. To this end, we elected to use MSD to optimize the rotameric configuration of mutant sequences on ensembles of fixed backbone templates using a physics-based potential energy function [Mayo *et al.*, 1990].

Previously, we showed that MSD employed in a positive design approach improves the quality of predicted sequences by using ensembles approximating conformational flexibility as input templates, but fails to correlate predicted with experimental stabilities of proteins [Figure 6.1.A] [Davey and Chica, 2014]. Here, we hypothesized that MSD using a negative design approach in which sequences are evaluated in the context of both an ensemble approximating the native folded state and an ensemble approximating a high-energy but accessible non-native state would result in improved correlation [Figure 6.1.B] and fewer unfolded species compared to MSD used in positive design.

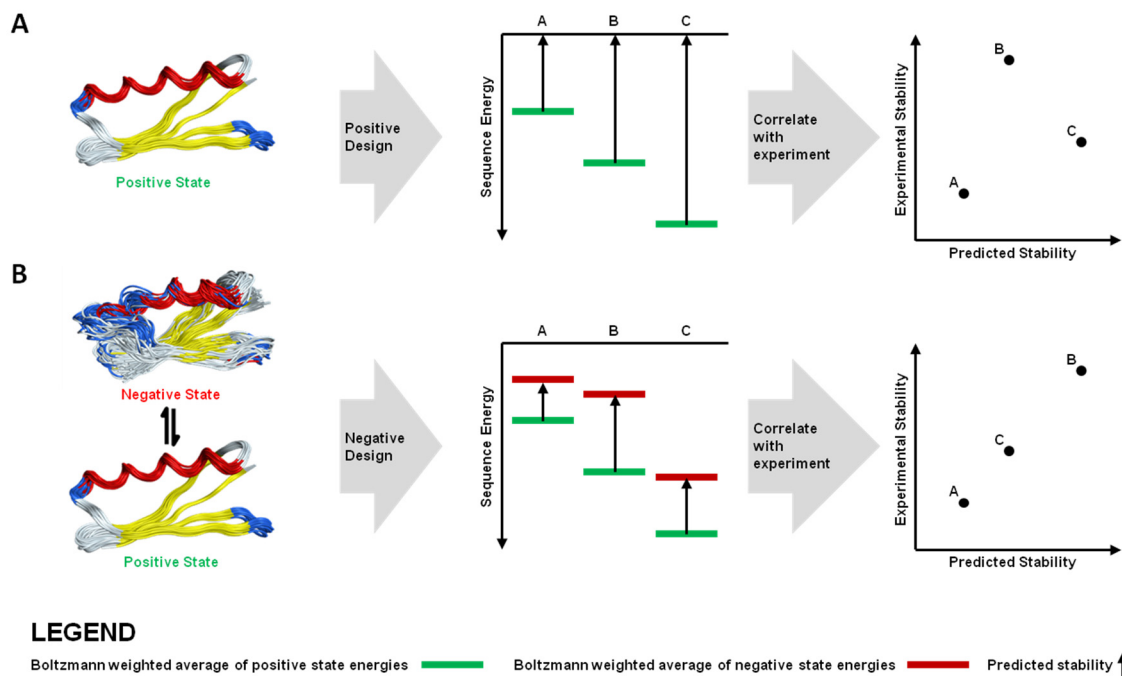


Figure 6.1. Multistate design methodologies. In MSD, the stability of sequences (1, 2, and 3) is evaluated in the context of a backbone ensemble approximating conformational flexibility. In positive design (A), a single backbone ensemble approximating the native folded state (positive state) is used to compute sequence fitness values, in this case the Boltzmann-weighted average of energies on individual backbones included in the ensemble. To compute sequence stability (arrow), fitness values are compared with respect to a reference value, typically represented as either an arbitrarily fixed energy value (as shown) or a composition-dependent value derived from implicit amino acid reference energies. In contrast, negative design (B) incorporates an additional non-native backbone ensemble intended to approximate the unfolded state (negative state). In this case, fitness values are computed for each sequence on both the positive and negative states, and their difference corresponds to predicted stability. We hypothesize that incorporation of a non-native ensemble in negative design will lead to correlation of predicted and experimental stabilities, which is not the case in positive design.

Our hypothesis is based on previous observations where MSD with backbone ensembles closely resembling the crystal structure and having low potential energy (i.e. on-target ensembles) gave rise to top-ranked predictions enriched in stably folded proteins whereas MSD with backbone ensembles having low structural similarity to the

crystal structure and high potential energy (i.e. off-target ensembles) produced top-ranked predictions containing protein sequences that were largely unfolded [Allen *et al.*, 2010; Davey and Chica, 2014]. Based on the scoring behaviour of off-target ensembles, we postulated that they could serve as useful reference states in negative design, helping to discard unfolded sequences as these would be scored more favorably on off-target than on on-target ensembles.

In this computational framework, MSD is used to compute the fitness of sequences in the context of native (positive state) and non-native (negative state) ensembles. Fitness values are computed as the Boltzmann-weighted average energy ($T = 300\text{ K}$) for each sequence across individual members of an ensemble, reflecting how well each sequence stabilizes each ensemble. These fitness values are then converted to stability units, reported as C_m values using multiple linear regression against a training set of known values. In positive design, a single energy value for each sequence computed from the positive state is used in regression whereas in negative design, two energy values for each sequence derived from positive and negative states are used. In this way, regression provides coefficients (α and β for sequence fitness calculated from the first and second ensemble used in regression, respectively) and intercepts that are then used to convert energy values into C_m values, enabling the comparison of predicted and experimental stabilities.

6.3. Training Set Stability Correlation

As a starting point in the development of our computational framework, we generated a series of backbone ensembles [Figure 6.2] using computational methods such as the PertMin protocol [Davey and Chica, 2014], Backrub motions [Davis *et al.*, 2006; Lauck *et al.*, 2010], and MD simulations run at 100, 500, and 1000 K. Backbone ensembles were generated using input coordinates for GB1 retrieved from the PDB (PDB ID: 1PGA) [Gallagher *et al.*, 1994]. Following the removal of crystallographic water molecules, the 1PGA structure was prepared by the addition of hydrogens, solvent (H₂O cube of 50 Å in length), and counter-ions (Na⁺ and Cl⁻) using the Protonate3D utility [Labute, 2009] available in the MOE software package (2012) [MOE, 2012]. 64-member PertMin ensemble were created by truncated Newton energy minimization [Nash, 2000] of the prepared 1PGA structure subjected to random perturbation of heavy atom coordinates (± 0.001 Å along each Cartesian cardinal axis). Either 50 or 150 minimization iterations were applied to give two ensembles referred to as PertMin₅₀ and PertMin₁₅₀, respectively.

For the generation of the MD ensembles, the prepared 1PGA structure was first energy minimized by conjugate gradient minimization to a root-mean-square gradient below $0.001 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-1}$. NVT MD simulations [Bond *et al.*, 1999; Sturgeon and Laird, 2000] of 100, 500, or 1000 K were initiated from the energy minimized 1PGA structure. MD trajectories were heated over 10 ps and equilibrated for another 10 ps. This was followed by a 6.4, 64, or 640 ps production run samples at 0.1, 1, or 10 ps increments, respectively, to produce sets of three 64-member MD ensembles for each temperature

(100, 500, and 1000 K). All energy minimization and MD simulations were performed using the AMBER99 force field [Wang *et al.*, 2000], with a combined explicit and implicit reaction field solvent model set up using the MOE software package. In addition to PertMin and MD ensembles, a 50-member Backrub ensemble was created from the 1PGA crystal structure using the default RosettaBackrub protocol [Lauck *et al.*, 2010] made available on a server from the Kortemme lab (<http://kortemmelab.ucsf/backrub>). These methods were selected because they enable the generation of on- or off-target ensembles encompassing backbones possessing different structures and energies [Davey and Chica, 2014].

These ensembles served as input templates in MSD to compute fitness values for a training set of 18 G β 1 mutants [Table 6.1]. Training set sequences were selected because (i) they can be favorably scored on on-target ensembles [Davey and Chica, 2014], (ii) they are all folded, (iii) their stability is comparable to that of the WT, and (iv) their stability was measured by performing chemical denaturation experiments under identical conditions, providing a consistent set of C_m values [Allen *et al.*, 2010]. The training set of sequences consist of GB1 variant sequences containing between 1 and 4 mutations from the WT, with stabilities spanning a range of C_m values from 1.84 M to 2.98 M GuHCl.

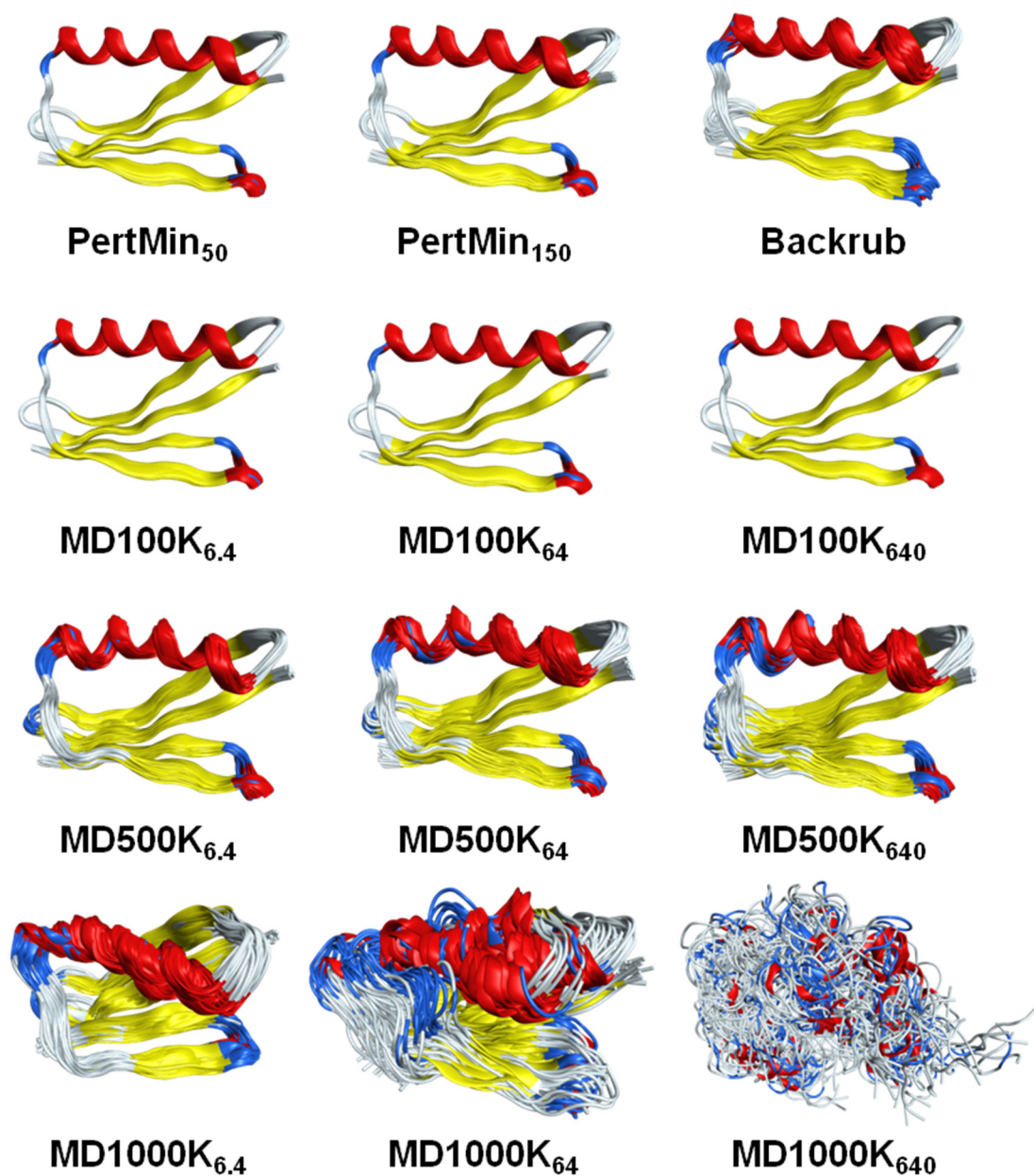


Figure 6.2. Gβ1 backbone ensembles. Superimposed members of the PertMin, Backrub, and MD (MD100K, MD500K, and MD1000K) ensembles are shown in cartoon representations. The number of energy minimization steps (PertMin ensembles) or duration of MD simulation in picoseconds (MD ensembles) are indicated as subscripts. Backbones are coloured according to their secondary structural elements: α-helices, β-strands, turns, and undefined secondary structures are coloured red, yellow, blue, and white, respectively. The Backrub ensemble comprises 50 individual backbones whereas all other ensembles comprise 64.

Table 6.1. Experimental stabilities for Gβ1 training set sequences

Sequence ID	Mutations from WT	Cm ^a (M)
1	Y3F/L7I/V39I	2.98
2	Y3F/L7I/V39L/V54I	2.66
3	Y3F/L7V/V39I	2.63
4	Y3F/L7I	2.67
5	Y3F/L7I/V39I/V54I	2.85
6	Y3F/L7I/V39L	2.43
7	Y3F/L7V/V39I/V54I	2.47
8	Y3F/L7V/V39L/V54I	2.44
9	Y3F/V39I	2.40
10	Y3F/L7I/V54I	2.47
11	Y3F	2.35
12	Y3F/L7V	2.23
13	Y3F/L7V/V39L	2.13
14	Y3F/V39L/V54I	2.04
15	Y3F/L7V/V54I	2.06
16	Y3F/V39I/V54I	1.98
17	Y3F/V39L	2.01
18	Y3F/V54I	1.84

^a Experimental stability retrieved from [Allen *et al.*, 2010] and reported as the concentration of guanidium chloride at the midpoint of denaturation (Cm)

All CPD calculations were performed with PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012] using the FASTER algorithm [Allen and Mayo, 2006; Allen and Mayo, 2010] for sequence optimization. The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 [Dunbrack and Cohen,

1997] was used to provide side chain conformations to be threaded onto each fixed backbone template. Side chain rotamers of core residues (positions 3, 5, 7, 30, 34, 52, and 54 according to 1PGA residue numbering) were optimized on each fixed backbone template using hydrophobic amino acids (Ala, Val, Leu, Ile, and Phe) found at these positions in the G β 1 sequences published by Mayo and coworkers [Allen *et al.*, 2010] as well as the WT Tyr amino acid at position 3: 3 (Tyr and Phe), 5 (Ala, Ile, and Leu), 7 (Val, Ile, Leu, and Phe), 30 (Ile, Leu, and Phe), 39 (Val, Ile, and Leu), 52 (Val, Ile, Leu, and Phe), and 54 (Ala, Val, and Ile). Side chain rotamers of neighboring core residues (positions 20, 26, 34, and 43) were also optimized but their identity was not allowed to vary.

The design sequence space thus consisted of 2592 G β 1 mutant sequences. Sequences were scored using a potential energy function consisting of a van der Waals term from the Dreiding II force field with atomic radii scale by 0.9 [Mayo *et al.*, 1990], a direction specific hydrogen-bond term having a well depth of 8.0 kcal·mol⁻¹ [Dahiyat and Mayo, 1997], and electrostatic energy term modelled using Coulomb's law with a distance dependent dielectric of 40, and a surface area-based solvation penalty term [Lazaridis and Karplus, 1999; Street and Mayo, 1998]. This scoring function does not include the backbone template energy nor does it account for the difference in entropy of folding for each sequence. Multistate sequence fitness was reported as a Boltzmann weighted average at 300 K of the individual energies for each backbone from the ensemble. Fitness values of sequences belonging to the G β 1 training set were extracted from the ranked list of scored sequences obtain as described above.

Computed fitness values for each training set sequence on each ensemble [Table 6.2] were negative, indicating that there is at least one backbone in each ensemble where these sequences can be scored favorably. Following MSD, multiple linear regression was applied to model the relationship between fitness and Cm values for the 18 training set sequences using either a single ensemble or pairs of ensembles in positive or negative design, respectively [Figure 6.3].

Table 6.2. MSD of Gβ1 training set sequences using various ensembles

Sequence ID	Ensemble Boltzmann Weighted Average Energy (kcal·mol ⁻¹)											
	PertMin ₅₀	PertMin ₁₅₀	Backrub	MD100K _{6.4}	MD100K ₆₄	MD100K ₆₄₀	MD500K _{6.4}	MD500K ₆₄	MD500K ₆₄₀	MD1000K _{6.4}	MD1000K ₆₄	MD1000K ₆₄₀
1	-86.1	-88.0	-78.7	-78.8	-78.4	-77.0	-80.5	-77.9	-78.0	-78.8	-75.8	-73.6
2	-81.4	-80.7	-72.9	-74.6	-78.7	-78.6	-75.8	-76.0	-75.8	-72.6	-66.6	-75.5
3	-82.6	-84.4	-76.0	-75.5	-74.9	-73.9	-76.9	-74.2	-74.3	-76.2	-81.1	-71.8
4	-83.4	-85.4	-79.0	-76.7	-76.7	-75.6	-76.8	-75.1	-75.3	-75.8	-72.3	-72.6
5	-81.0	-81.0	-73.9	-73.3	-76.7	-76.6	-74.1	-72.6	-73.1	-75.6	-66.6	-67.4
6	-84.3	-85.5	-74.3	-77.2	-75.8	-75.9	-79.9	-76.5	-76.6	-74.6	-72.3	-75.3
7	-77.5	-77.6	-71.2	-69.8	-73.0	-73.3	-70.8	-68.7	-70.2	-72.3	-63.6	-66.1
8	-78.0	-78.2	-70.6	-70.7	-74.9	-74.8	-73.4	-72.3	-72.1	-70.6	-66.4	-72.8
9	-83.3	-85.3	-74.5	-77.8	-77.0	-75.3	-76.8	-77.0	-77.0	-78.3	-83.5	-77.4
10	-80.1	-79.3	-74.4	-73.0	-77.8	-77.6	-71.5	-73.2	-74.4	-73.1	-65.8	-69.8
11	-81.4	-83.9	-75.5	-76.5	-76.7	-75.8	-74.1	-74.0	-74.0	-75.7	-79.8	-74.1
12	-80.1	-81.7	-76.7	-73.7	-73.8	-72.7	-73.2	-71.4	-71.5	-72.8	-77.7	-69.8
13	-81.0	-82.5	-73.6	-73.8	-72.1	-71.9	-76.4	-72.8	-72.9	-72.4	-72.3	-72.6
14	-75.7	-80.3	-69.0	-72.8	-76.8	-76.7	-74.8	-75.0	-76.0	-74.6	-70.6	-72.9
15	-76.5	-75.6	-71.7	-69.6	-74.0	-73.7	-69.5	-69.4	-70.9	-69.8	-68.5	-68.4
16	-77.8	-75.8	-70.7	-73.0	-75.8	-75.6	-71.6	-71.3	-74.8	-74.1	-67.7	-70.0
17	-80.2	-83.1	-75.0	-76.2	-74.5	-74.1	-75.9	-75.4	-75.8	-72.8	-77.4	-72.5
18	-79.1	-75.3	-71.3	-72.8	-77.0	-76.8	-71.4	-72.1	-72.0	-73.3	-73.4	-73.3

Linear regression of predicted and experimental stabilities was performed using the least-squares algorithm. Regression returns a linear equation allowing for the conversion of calculated sequence fitness into predicted stability values. In positive design, sequence stability is calculated using a single ensemble (E_α) giving a single coefficient (α) and an intercept. The resulting linear equation has the form:

$$C_m = \alpha \cdot E_\alpha + \text{intercept} \quad (\text{EQ 6.1})$$

In negative design, sequence stability is evaluated on an ensemble pair (E_α and E_β), with regression giving two coefficients (α and β , for the first and second ensembles of the ensemble pair, respectively) and an intercept. The resulting linear equation has the form:

$$C_m = \alpha \cdot E_\alpha + \beta \cdot E_\beta + \text{intercept} \quad (\text{EQ 6.2})$$

The square of the Pearson correlation coefficient (R^2) is reported indicating the quality of fit between predicted and experimental stabilities, with values $0 \leq R^2 \leq 1$, where a value of unity indicates perfect agreement. Probability values (p-values) are also reported for each regression parameter as well as a regression F-test. Statistical analysis of variance provides p-value values that report on the probability that a regression coefficient has no effect in the correlation, with values $0 \leq \text{p-value} \leq 1$. Coefficients are considered to make a statistically significant contribution to regression if their corresponding p-values are lower than 2.4×10^{-4} (an alpha value of $0.05 \div 210$ hypothetical ensemble pairs). It is important to note that our computational framework does not enforce the identity of an ensemble as a positive or negative state: it is instead determined by the regression.

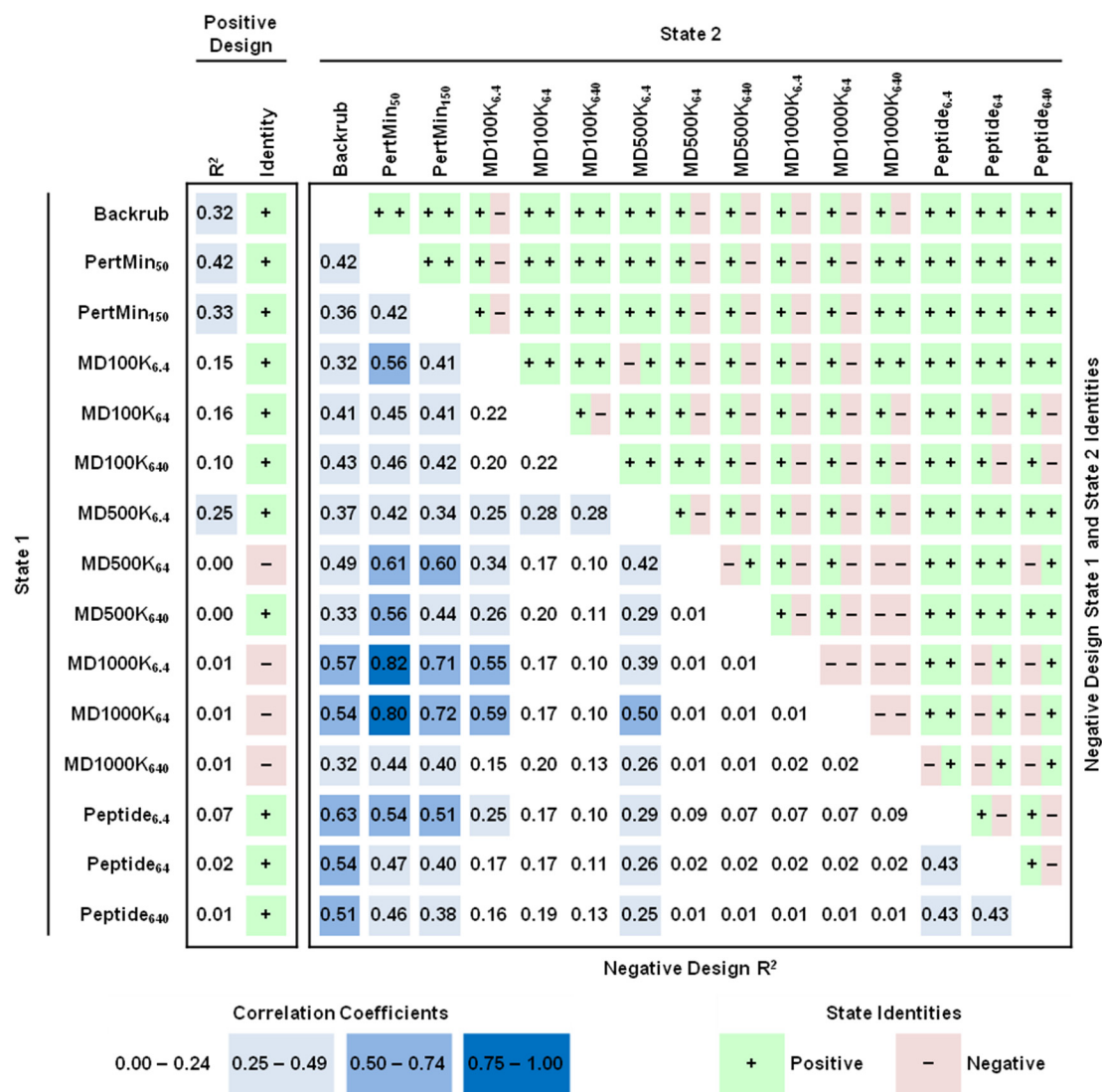


Figure 6.3. Correlation of MSD and experimental Gβ1 training set stabilities. Correlation coefficients (R²) resulting from regression of fitness values obtained for a training set of 18 Gβ1 mutants using a variety of backbones ensembles are shown. In positive and negative design, a single backbone ensemble (state 1) or a pair of ensembles (states 1 and 2) are used, respectively. Correlation coefficients are colored according to their values. Backbone ensembles are identified as positive (green) or negative (red) states according to the sign on their coefficient from multiple linear regression. Ensembles that behave as positive or negative states are indicated by + and – symbols, respectively. Only cases where one ensemble is a positive state and the other is a negative state are considered to be negative design (+/– or –/+).

Thus, analysis of α and β regression coefficients [Table 6.3] is necessary to determine whether ensembles are positive or negative states. For example, ensembles are considered positive states [Figure 6.3, green/+] if sequence fitness values obtained on these ensembles [Table 6.2] are multiplied by a negative coefficient, which would result in an increase in predicted C_m values. Conversely, ensembles are considered negative states [Figure 6.3, red/-] if fitness values obtained on these ensembles are multiplied by a positive coefficient, resulting in a decrease in predicted C_m values. Thus, only cases where one ensemble acted as a positive state and the other as a negative state are considered negative design and were analyzed further.

Table 6.3. Multiple linear regression for G β 1 training set sequences

State 1	State 2	R ²	Regression Parameters			Parameter P-Values			F-Test P-Value
			Intercept	α	β	Intercept	α	β	
Positive Design									
Backrub		0.3193	-2.4134	-0.0648		0.1861	0.0145		0.0145
PertMin ₅₀		0.4202	-3.5505	-0.0735		0.0580	0.0036		0.0036
PertMin ₁₅₀		0.3300	-1.5553	-0.0483		0.2828	0.0127		0.0127
MD100K _{6,4}		0.1492	-1.0421	-0.0460		0.6160	0.1133		0.1133
MD100K ₆₄		0.1645	-2.9712	-0.0704		0.3382	0.0949		0.0949
MD100K ₆₄₀		0.0982	-1.8097	-0.0555		0.5757	0.2055		0.2055
MD500K _{6,4}		0.2521	-2.6070	-0.0672		0.2416	0.0337		0.0337
MD500K ₆₄		0.0020	2.5429	0.0024		0.0198	0.8613		0.8613
MD500K ₆₄₀		0.0006	2.1776	-0.0027		0.2815	0.9232		0.9232
MD1000K _{6,4}		0.0053	2.6540	0.0048		0.0155	0.7741		0.7741
MD1000K ₆₄		0.0053	2.5472	0.0035		0.0008	0.7745		0.7745
MD1000K ₆₄₀		0.0086	2.6454	0.0208		0.0027	0.7148		0.7148
Negative Design									
Backrub	PertMin ₅₀	0.4233	-3.6417	-0.0111	-0.0645	0.0640	0.7833	0.1209	0.0161
Backrub	PertMin ₁₅₀	0.3636	-2.4720	-0.0340	-0.0286	0.1771	0.3878	0.3233	0.0337
Backrub	MD100K _{6,4}	0.3208	-2.2649	-0.0695	0.0067	0.2711	0.0706	0.8597	0.0550

Backrub	MD100K ₆₄	0.4130	-6.0260	-0.0582	-0.0541	0.0533	0.0236	0.1426	0.0184
Backrub	MD100K ₆₄₀	0.4317	-6.9979	-0.0663	-0.0594	0.0410	0.0096	0.1055	0.0144
Backrub	MD500K _{6.4}	0.3698	-3.8176	-0.0474	-0.0363	0.0971	0.1148	0.2902	0.0313
Backrub	MD500K ₆₄	0.4889	-2.7148	-0.0948	0.0265	0.1041	0.0018	0.0414	0.0065
Backrub	MD500K ₆₄₀	0.3275	-1.8493	-0.0669	0.0100	0.4196	0.0165	0.6758	0.0510
Backrub	MD1000K _{6.4}	0.5652	-3.0384	-0.1056	0.0404	0.0546	0.0005	0.0107	0.0019
Backrub	MD1000K ₆₄	0.5374	-3.6874	-0.1007	0.0267	0.0322	0.0008	0.0179	0.0030
Backrub	MD1000K ₆₄₀	0.3197	-2.3286	-0.0645	0.0047	0.2633	0.0193	0.9239	0.0556
PertMin ₅₀	PertMin ₁₅₀	0.4211	-3.4883	-0.0677	-0.0050	0.0775	0.1452	0.8803	0.0166
PertMin ₅₀	MD100K _{6.4}	0.5644	-2.9020	-0.1521	0.0941	0.0869	0.0018	0.0416	0.0020
PertMin ₅₀	MD100K ₆₄	0.4498	-5.3509	-0.0655	-0.0323	0.0628	0.0138	0.3839	0.0113
PertMin ₅₀	MD100K ₆₄₀	0.4566	-5.7984	-0.0692	-0.0344	0.0590	0.0067	0.3321	0.0103
PertMin ₅₀	MD500K _{6.4}	0.4231	-3.8093	-0.0673	-0.0103	0.0797	0.0522	0.7882	0.0161
PertMin ₅₀	MD500K ₆₄	0.6090	-3.9143	-0.1026	0.0274	0.0185	0.0002	0.0168	0.0009
PertMin ₅₀	MD500K ₆₄₀	0.5551	-2.2370	-0.0989	0.0467	0.2054	0.0006	0.0499	0.0023
PertMin ₅₀	MD1000K _{6.4}	0.8205	-5.1250	-0.1330	0.0543	0.0002	< 0.0001	< 0.0001	< 0.0001
PertMin ₅₀	MD1000K ₆₄	0.8041	-6.1831	-0.1305	0.0379	0.0001	< 0.0001	< 0.0001	< 0.0001
PertMin ₅₀	MD1000K ₆₄₀	0.4389	-4.4363	-0.0791	-0.0326	0.0586	0.0040	0.4910	0.0131
PertMin ₁₅₀	MD100K _{6.4}	0.4073	0.1264	-0.0943	0.0731	0.9454	0.0220	0.1823	0.0199
PertMin ₁₅₀	MD100K ₆₄	0.4076	-4.8686	-0.0426	-0.0498	0.0944	0.0254	0.1814	0.0197
PertMin ₁₅₀	MD100K ₆₄₀	0.4183	-5.4627	-0.0476	-0.0526	0.0809	0.0116	0.1521	0.0172
PertMin ₁₅₀	MD500K _{6.4}	0.3442	-2.4130	-0.0375	-0.0234	0.2636	0.1673	0.5778	0.0423
PertMin ₁₅₀	MD500K ₆₄	0.6002	-1.7974	-0.0835	0.0362	0.1291	0.0003	0.0062	0.0010
PertMin ₁₅₀	MD500K ₆₄₀	0.4393	0.0779	-0.0657	0.0423	0.9625	0.0038	0.1079	0.0131
PertMin ₁₅₀	MD1000K _{6.4}	0.7119	-2.1107	-0.0953	0.0551	0.0432	< 0.0001	0.0005	< 0.0001
PertMin ₁₅₀	MD1000K ₆₄	0.7222	-3.3985	-0.0966	0.0404	0.0043	< 0.0001	0.0003	< 0.0001
PertMin ₁₅₀	MD1000K ₆₄₀	0.3996	-3.6698	-0.0627	-0.0707	0.1022	0.0069	0.2071	0.0218
MD100K _{6.4}	MD100K ₆₄	0.2150	-3.6708	-0.0301	-0.0502	0.2542	0.3416	0.2799	0.1627
MD100K _{6.4}	MD100K ₆₄₀	0.1984	-3.5816	-0.0390	-0.0406	0.3011	0.1910	0.3528	0.1904
MD100K _{6.4}	MD500K _{6.4}	0.2530	-2.5632	0.0062	-0.0728	0.2695	0.8925	0.1694	0.1122
MD100K _{6.4}	MD500K ₆₄	0.3409	-2.4069	-0.0966	0.0331	0.2393	0.0141	0.0542	0.0439
MD100K _{6.4}	MD500K ₆₄₀	0.2642	-0.2511	-0.0849	0.0512	0.9029	0.0345	0.1466	0.1002
MD100K _{6.4}	MD1000K _{6.4}	0.5479	-4.3133	-0.1453	0.0692	0.0284	0.0007	0.0024	0.0026
MD100K _{6.4}	MD1000K ₆₄	0.5946	-6.4900	-0.1574	0.0546	0.0050	0.0003	0.0010	0.0011
MD100K _{6.4}	MD1000K ₆₄₀	0.1537	-1.5184	-0.0494	-0.0164	0.5811	0.1295	0.7814	0.2859
MD100K ₆₄	MD100K ₆₄₀	0.2216	-2.3460	-0.1998	0.1385	0.4550	0.1439	0.3110	0.1528
MD100K ₆₄	MD500K _{6.4}	0.2791	-4.1411	-0.0337	-0.0534	0.1856	0.4649	0.1434	0.0859
MD100K ₆₄	MD500K ₆₄	0.1678	-2.7706	-0.0708	0.0031	0.4017	0.1043	0.8116	0.2522

MD100K ₆₄	MD500K ₆₄₀	0.1963	-2.6002	-0.0862	0.0217	0.4127	0.0753	0.4531	0.1942
MD100K ₆₄	MD1000K _{6.4}	0.1749	-2.6668	-0.0717	0.0068	0.4130	0.0995	0.6698	0.2364
MD100K ₆₄	MD1000K ₆₄	0.1716	-2.7958	-0.0708	0.0040	0.3864	0.1032	0.7259	0.2438
MD100K ₆₄	MD1000K ₆₄₀	0.1994	-2.9566	-0.0778	0.0430	0.3466	0.0781	0.4312	0.1886
MD100K ₆₄₀	MD500K _{6.4}	0.2773	-4.3066	-0.0297	-0.0599	0.1986	0.4803	0.0730	0.0875
MD100K ₆₄₀	MD500K ₆₄	0.0989	-1.9971	-0.0565	-0.0015	0.5975	0.2234	0.9155	0.4580
MD100K ₆₄₀	MD500K ₆₄₀	0.1052	-1.5076	-0.0606	0.0096	0.6615	0.2054	0.7369	0.4346
MD100K ₆₄₀	MD1000K _{6.4}	0.1002	-1.5789	-0.0548	0.0030	0.6584	0.2277	0.8567	0.4530
MD100K ₆₄₀	MD1000K ₆₄	0.0988	-1.6942	-0.0548	0.0012	0.6308	0.2312	0.9184	0.4582
MD100K ₆₄₀	MD1000K ₆₄₀	0.1265	-1.8176	-0.0624	0.0388	0.5807	0.1752	0.4963	0.3627
MD500K _{6.4}	MD500K ₆₄	0.4190	-3.5023	-0.1057	0.0271	0.1001	0.0051	0.0555	0.0170
MD500K _{6.4}	MD500K ₆₄₀	0.2948	-1.7796	-0.0800	0.0247	0.4547	0.0244	0.3558	0.0729
MD500K _{6.4}	MD1000K _{6.4}	0.3895	-3.0342	-0.0955	0.0282	0.1527	0.0077	0.0860	0.0247
MD500K _{6.4}	MD1000K ₆₄	0.4955	-4.9590	-0.1198	0.0300	0.0266	0.0017	0.0168	0.0059
MD500K _{6.4}	MD1000K ₆₄₀	0.2579	-2.3540	-0.0668	0.0172	0.3273	0.0403	0.7357	0.1068
MD500K ₆₄	MD500K ₆₄₀	0.0054	2.1438	0.0045	-0.0077	0.3047	0.7907	0.8221	0.9599
MD500K ₆₄	MD1000K _{6.4}	0.0057	2.6262	-0.0018	0.0065	0.0271	0.9387	0.8156	0.9580
MD500K ₆₄	MD1000K ₆₄	0.0057	2.4845	-0.0017	0.0046	0.0309	0.9405	0.8166	0.9583
MD500K ₆₄	MD1000K ₆₄₀	0.0130	2.9564	0.0037	0.0241	0.0562	0.7996	0.6884	0.9068
MD500K ₆₄₀	MD1000K _{6.4}	0.0118	2.1054	-0.0107	0.0085	0.3125	0.7574	0.6857	0.9147
MD500K ₆₄₀	MD1000K ₆₄	0.0110	1.9664	-0.0097	0.0057	0.3588	0.7730	0.6973	0.9206
MD500K ₆₄₀	MD1000K ₆₄₀	0.0092	2.9105	0.0031	0.0240	0.3247	0.9245	0.7235	0.9331
MD1000K _{6.4}	MD1000K ₆₄	0.0055	2.6124	0.0026	0.0018	0.0498	0.9511	0.9527	0.9592
MD1000K _{6.4}	MD1000K ₆₄₀	0.0228	3.3074	0.0086	0.0323	0.0580	0.6472	0.6120	0.8413
MD1000K ₆₄	MD1000K ₆₄₀	0.0225	3.1084	0.0061	0.0320	0.0263	0.6506	0.6148	0.8431

As shown on Figure 6.3, ensembles previously identified as on-target (PertMin₅₀, PertMin₁₅₀, Backrub, MD100K_{6.4}, MD100K₆₄, and MD100K₆₄₀) [Davey and Chica, 2014] result in moderate to strong correlations ($R^2 = 0.10 - 0.42$ or $R = 0.31 - 0.65$) when used in positive design, and all behave as positive states when used in negative design with the exception of the MD100K_{6.4} ensemble, which is a negative state when used with the MD500K_{6.4} ensemble. On the other hand, ensembles previously identified as off-target

(MD500K and MD1000K ensembles) result in negligible correlations (R^2 of 0.00 – 0.01 or R of 0.00 – 0.10) when used in positive design, with the exception of the MD500K_{6,4} ensemble which results in a strong correlation ($R^2 = 0.25$ or $R = 0.50$). As expected, MD500K and MD1000K ensembles act as negative states when paired with on-target ensembles in negative design, the only exception being the MD500K_{6,4} ensemble that acts as positive state when paired with the MD100K_{6,4} ensemble.

Analysis of regression parameters obtained by negative design using all ensemble pairs [Table 6.3] identifies 11 on-target/on-target ensemble pairs as positive/positive states, 9 on-target/off-target pairs as positive/positive states, 4 on-target/on-target pairs as positive/negative states, 26 on-target/off-target pairs as positive/negative states, 1 on-target/off-target pair as a negative/positive state, 10 off-target/off-target pairs as positive/negative states, and 5 off-target/off-target pairs as negative/negative states. These results suggest that while not all on-target/off-target ensemble pairs behave as positive/negative states in negative design, the majority (26/41) of ensemble pairs where one ensemble acted as a positive state and the other as a negative state belonged to an on-target and off-target ensemble positive/negative combination. The best correlations ($R^2 \geq 0.70$) were obtained from four ensemble pairs, with the positive state being either the PertMin₅₀ or PertMin₁₅₀ on-target ensembles and the negative state being either the MD1000K_{6,4} or MD1000K₆₄ off-target ensembles.

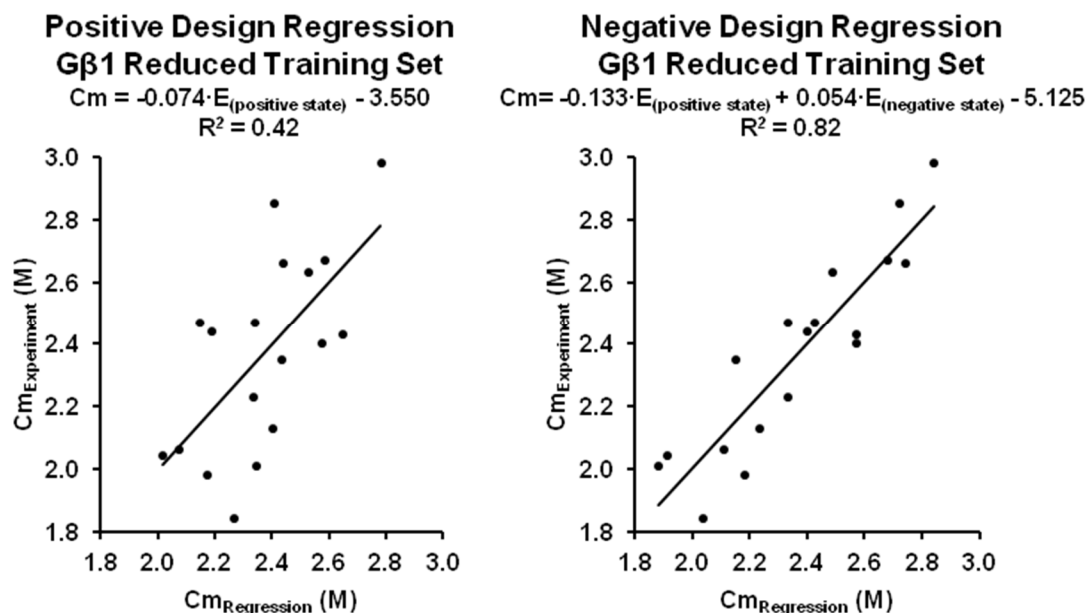


Figure 6.4. Training set sequence stabilities by positive and negative MSD

In all cases where an on-target (Backrub, PertMin or MD100K) and an off-target (MD500K or MD 1000K) ensemble are used in negative design as positive and negative states, respectively, the correlation coefficient is equal to or greater than that obtained by positive design. The best correlation ($R^2 = 0.82$) between predicted and experimental C_m was obtained when the PertMin₅₀ ensemble was used as a positive state in conjunction with the MD1000K_{6.4} negative state ensemble [Figure 6.4], a significant improvement in correlation compared to positive design ($R^2 = 0.42$). Of note, the regressed C_m values for a majority of training set sequences are closer to the experimental values when using negative design with the PertMin₅₀ and MD1000K_{6.4} ensembles than when using positive design with the PertMin₅₀ ensemble [Table 6.4], with an average error of approximately 5%, a value roughly two-fold lower than that obtained with positive design. The smaller error on predicted C_m and the significantly

improved R^2 value obtained by negative design demonstrates that this approach can improve calculation accuracy, leading to improved correlation of predicted with experimental stabilities.

Table 6.4. Calculated stabilities for G β 1 training set sequences

Sequence ID	Experimental Cm ^a (M)	Regression Cm Value ^b (M)		Regression Cm Error ^c (%)	
		Positive Design	Negative Design	Positive Design	Negative Design
1	2.98	2.78	2.84	7	5
2	2.66	2.43	2.74	9	3
3	2.63	2.52	2.49	4	5
4	2.67	2.58	2.68	3	0
5	2.85	2.41	2.72	15	5
6	2.43	2.65	2.57	9	6
7	2.47	2.14	2.33	13	6
8	2.44	2.18	2.40	11	2
9	2.40	2.57	2.57	7	7
10	2.47	2.34	2.43	5	2
11	2.35	2.43	2.15	3	9
12	2.23	2.33	2.34	4	5
13	2.13	2.40	2.24	13	5
14	2.04	2.02	1.91	1	6
15	2.06	2.07	2.11	0	2
16	1.98	2.17	2.19	10	11
17	2.01	2.34	1.88	16	6
18	1.84	2.27	2.04	23	11

^a Experimental stability retrieved from [Allen *et al.*, 2010] and reported as the concentration of guanidium chloride at the midpoint of denaturation (Cm)

^b Regression stability reported for linear regression using positive design with the PertMin₅₀ ensemble and multiple linear regression using negative design with the PertMin₅₀ and MD1000K_{6.4} ensembles

^c Numbers in parentheses indicate the percent difference between regressed and experimental Cm values

Because it is possible that these improvements could be due solely to the inclusion of an additional parameter in the regression model, p-values for regression coefficients and F-tests were examined [Table 6.3]. Ensemble pairs consisting of a PertMin₅₀, PertMin₁₅₀, Backrub, or MD100K_{6.4} ensemble (positive state) and a MD1000K_{6.4} or MD1000K₆₄ ensemble (negative state) have p-values for their regression coefficients and F-tests less than or equal to 2.4×10^{-4} , suggesting that each coefficient provides a statistically significant contribution to the regression. Additionally, not all ensemble pairs lead to enhanced correlation [Figure 6.3], demonstrating that these improvements are not due solely to the inclusion of an additional parameter in the regression model.

6.4. Test Set Design and Correlation

To validate our computational framework, we next sought to design a test set of 10 new Gβ1 mutant sequences using negative design. First, positive MSD was performed as described before during stability correlation with the training set sequences using the PertMin, Backrub, MD100K, MD500K, and MD1000K ensembles serving as input templates. Following MSD, the ranked list of scored sequences obtained from each ensemble were compared, and all sequences common to all ensembles with the exception of the MD1000K ensembles were subjected to backbone drift modeling [Howell *et al.*, 2014] in order to find the best scores (i.e. minimum fitness value obtained for that sequence on any of the ensembles except the MD1000K ensembles) for each unique sequence contained within this collection of common sequences. All sequences were then re-ranked using the backbone drift modeling score, resulting in a new ranked

list of sequences. The top 10 sequences that had not been previously characterized and whose energy on each ensemble with the exception of the MD1000K ensembles was both lower than 0 kcal·mol⁻¹ and lower than the energy computed on the MD1000K ensembles were included in the test set. Specifically, all ten test set sequences scored (-81.6 – -77.2 kcal·mol⁻¹) better than or similar to the WT sequence (-77.5 kcal·mol⁻¹) as evaluated on the potential positive state ensembles and were found within the top 30 sequences. Selection of test set sequences in this manner enabled unbiased evaluation of all potential positive states.

Table 6.5. Experimental and calculated Gβ1 test set sequence stabilities

ID	Sequence		Predicted Cm ^b (M)		Regression Cm ^c (M)	
	Mutations from WT	Cm ^a (M)	Positive Design	Negative Design	Positive Design	Negative Design
A	L7I/V39I	2.88 ± 0.03	2.31 (20%) ^d	2.31 (20%)	2.50 (13%)	2.62 (9%)
B	L7I/V39L	2.30 ± 0.04	2.18 (5%)	2.06 (10%)	2.29 (1%)	2.31 (1%)
C	L7I	2.61 ± 0.02	2.11 (19%)	2.14 (18%)	2.18 (17%)	2.53 (3%)
D	Y3F/L7I/V39I/V54A	1.48 ± 0.02	2.10 (42%)	1.85 (25%)	2.17 (46%)	2.01 (36%)
E	V39I	2.14 ± 0.06	2.09 (2%)	2.02 (6%)	2.15 (0%)	2.33 (9%)
F	Y3F/F30L/V39I	1.64 ± 0.02	1.93 (18%)	1.50 (8%)	1.89 (15%)	1.58 (4%)
G	L7V/V39I	2.31 ± 0.07	2.05 (11%)	1.95 (16%)	2.08 (10%)	2.26 (2%)
H	Y3F/L7I/F30L/V39I	2.03 ± 0.02	2.15 (6%)	1.84 (9%)	2.24 (10%)	1.95 (4%)
I	Y3F/L5I/V39L	0.85 ± 0.01	1.19 (40%)	0.66 (23%)	0.72 (16%)	0.88 (4%)
J	Y3F/F30L	1.60 ± 0.03	1.77 (11%)	1.28 (20%)	1.64 (3%)	1.36 (15%)

^a Concentration of guanidium chloride at the midpoint of denaturation

^b Predicted stability calculated by applying regression parameters from the Gβ1 training set using positive design with the PertMin₅₀ ensemble and multiple linear regression using negative design with the PertMin₅₀ and MD1000K_{6.4} ensembles

^c Regression stability reported for linear regression using positive design with the PertMin₅₀ ensemble and multiple linear regression using negative design with the PertMin₅₀ and MD1000K_{6.4} ensembles

^d Values in parentheses indicate the percent difference of predicted or regressed Cm value from experimental Cm

The test set sequences contain between one to four mutations [Table 6.5], including several belonging to the training set as well as two (F30L and V54A) that were observed exclusively in non-native folding G β 1 variants [Allen *et al.*, 2010]. All test set sequences were expressed and purified, and their CD spectra were found to be similar to that of the WT sequence, suggesting that they all adopt the G β 1 fold [Figure 6.5].

The experimental work, conducted by *Adam M Damry* and *Christian K Euler*, required that DNA sequences for the WT and mutant G β 1 cloned in pJ441-01 plasmids be purchased from DNA 2.0. Plasmids were transformed into chemically competent *E. coli* BL21(DE3)-Gold cells (Stratagene), and the complete coding regions were verified by DNA sequencing. Proteins studied by chemical denaturation and CD were expressed using LB broth containing 100 μ g/mL ampicillin. 1 L cultures of *E. coli* BL21(DE3)-Gold cells (Stratagene) transformed with a pJ441-01 plasmid (DNA 2.0) containing one of the G β 1 variants were grown at 37 °C with shaking at 220 rpm until an optical density of 0.6 at 600 nm was reached. At this point, protein expression was induced with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG), and LB or M9 cultures were incubated overnight at 16 °C or 37 °C, respectively, with shaking at 220 rpm. Following incubation, cells were harvested by centrifugation and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography according to the manufacturer's protocol.

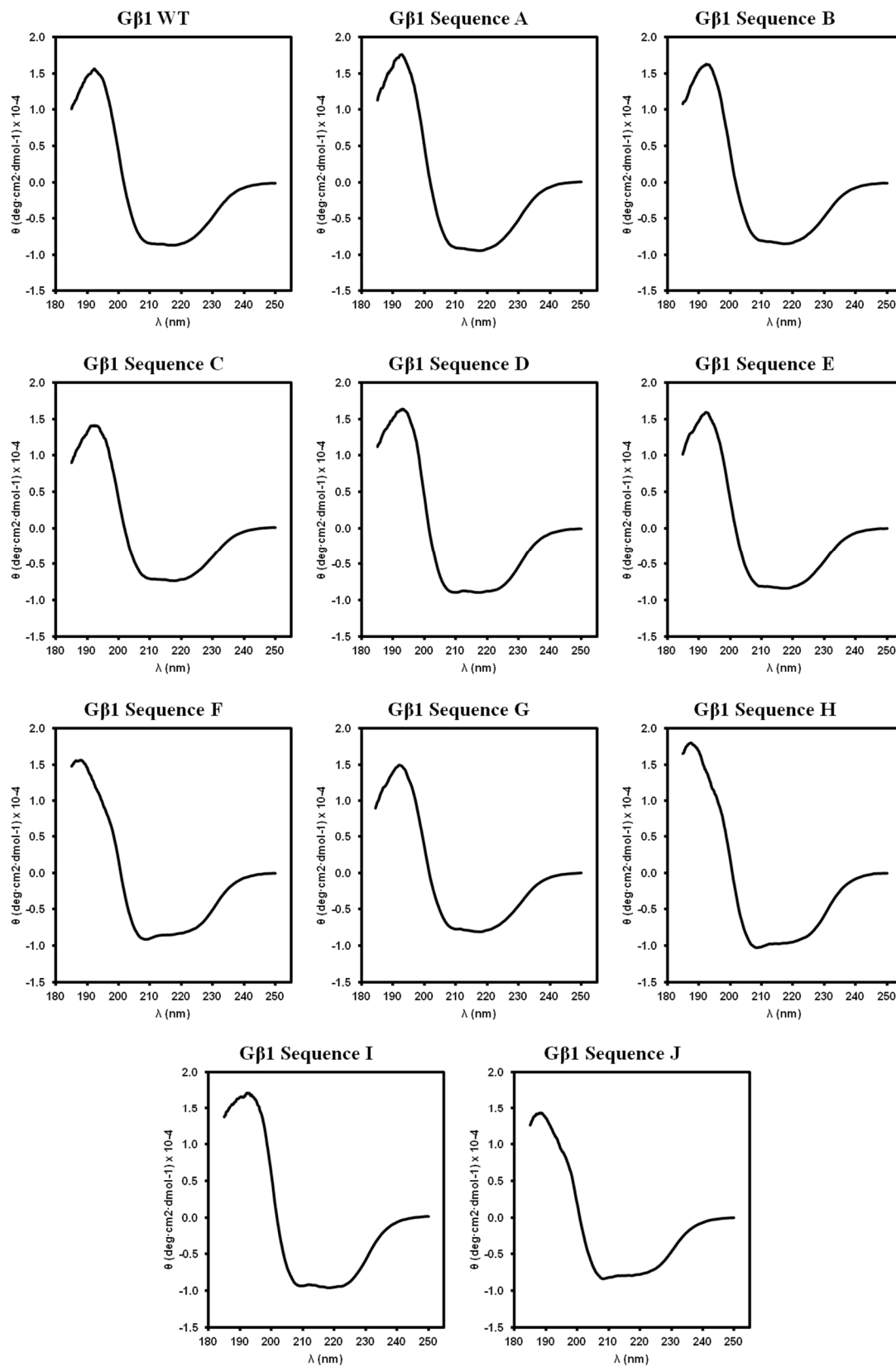


Figure 6.5. Far-UV CD spectroscopy of Gβ1 test set sequences

The CD measurements were performed using 40 μ M protein samples. A 650 μ L aliquot of each protein sample was placed in a 1-mm path length quartz cuvette (Jasco) and CD spectra from 185 nm to 250 nm (10 nm/min scan speed) were acquired with a Jasco J-815 spectrometer. All measurements were performed at 25 °C in 10 mM sodium phosphate buffer (pH 7.4). It should be noted that samples for CD or NMR spectroscopy were further purified and desalted by size exclusion chromatography on a Biologic DuoFlow system (Bio-Rad) using an ENrich SEC 650 column (Bio-Rad) with 10 mM sodium phosphate buffer (pH 7.4) as eluent. Protein concentrations were determined by their absorption at 280 nm following denaturation in 5 M guanidium chloride.

Because sequence H shows the largest discrepancy between its CD spectrum and the WT, the secondary structure of sequence H was confirmed by NMR [Figure 6.6.A] with chemical shift index measurements for the mutant being virtually identical to that obtained for the WT sequence [Figure 6.6.B]. To collect the NMR spectra of sequence H the preparation protocol required substitution of LB broth for M9 minimal medium containing 100 μ g/mL ampicillin, 1 g/L $^{15}\text{NH}_4\text{Cl}$, and 3 g/L of either unlabeled or ^{13}C -labeled glucose for preparation of ^{15}N - or $^{13}\text{C}/^{15}\text{N}$ -labeled samples, respectively. ^{15}N - and ^{13}C -labeled G β 1 samples were diluted to concentrations of 0.8 – 3 mM in 10 mM sodium phosphate buffer (pH 7.4) prepared in 10% D_2O . NMR spectra were collected using a Varian INOVA 500 MHz spectrometer fitted with a 5 mm triple resonance inverse probe. All spectra were acquired at 25 °C. Processing was performed with NMRPipe [Delaglio *et al.*, 1995] and spectra were analyzed with NMRViewJ (One Moon Scientific) [Johnson and Blevins, 1994]. Backbone chemical shift assignments were obtained from a standard

suite of 3D triple resonance experiments, including HSQC, HNCO, HNCACB and CBCA(CO)NH spectra. Secondary structure predictions from backbone chemical shifts [Wishart *et al.*, 1992] were performed in NMRViewJ.

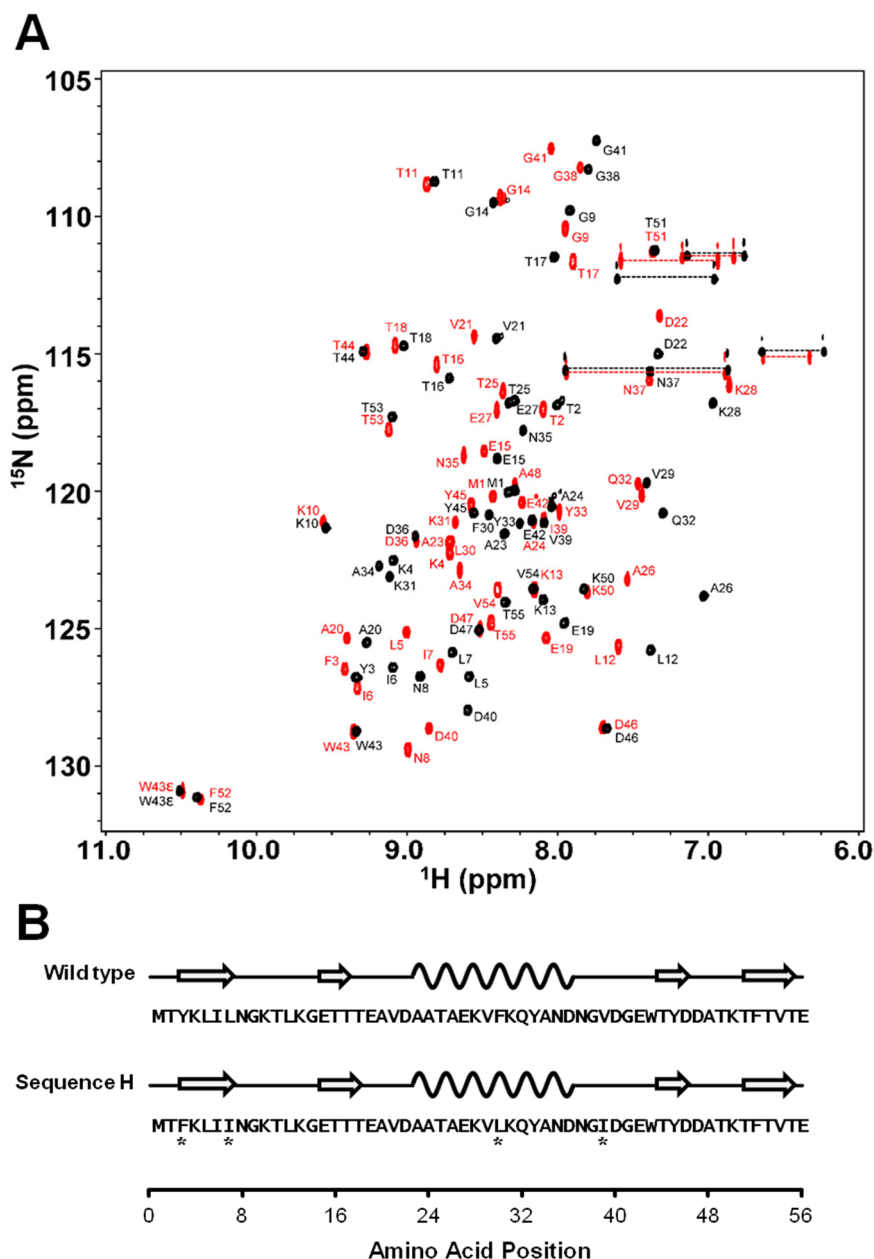


Figure 6.6. NMR study of G β 1 wild type and mutant sequence H sequences. ^1H - ^{15}N -HSQC spectra (A) for WT (black) and sequence H (red) sequences. Side chain amide resonances from asparagine and glutamine residues are connected by horizontal lines. Chemical shift indices (B) demonstrated conservation of secondary structural elements

between the WT and mutant sequences confirming they adopt the same folded structure.

Chemical denaturation assays were performed using protein samples normalized to 1 mg/mL. Protein aliquots of 25 μ L in individual wells of UV-transparent 96-well plates (Greiner Bio-One) were mixed with 175 μ L of 0 – 5 M guanidium chloride solutions (12 points, evenly spaced) and incubated at room temperature for an hour. Following chemical denaturation, fluorescence emission spectra were measured from 300 nm to 450 nm (excitation at 295 nm and step size of 2 nm) using an Infinite M1000 plate reader (Tecan). Spectra were then converted from integrated fluorescence to fraction of unfolded protein by assuming that maximum fluorescence intensity corresponded to 100% folded protein and that minimum fluorescence intensity corresponded to 0% folded protein. C_m values were determined by fitting a 3-term sigmoid using nonlinear least-squares regression.

Specifically, samples for chemical denaturation assays were then exchanged into 10 mM sodium phosphate buffer (pH 7.4) using Amicon Ultra-15 Centrifugal Filter Units (EMD Millipore). Chemical denaturation assays [Figure 6.7] showed that all 10 test set sequences unfold according to a two-state model and that five of the ten sequences [Table 6.5, sequences A, B, C, E, and G] have C_m values approximately equal to or greater than the WT sequence ($C_m = 2.23 \pm 0.04$ M) while those containing the F30L and V54A mutations are destabilized.

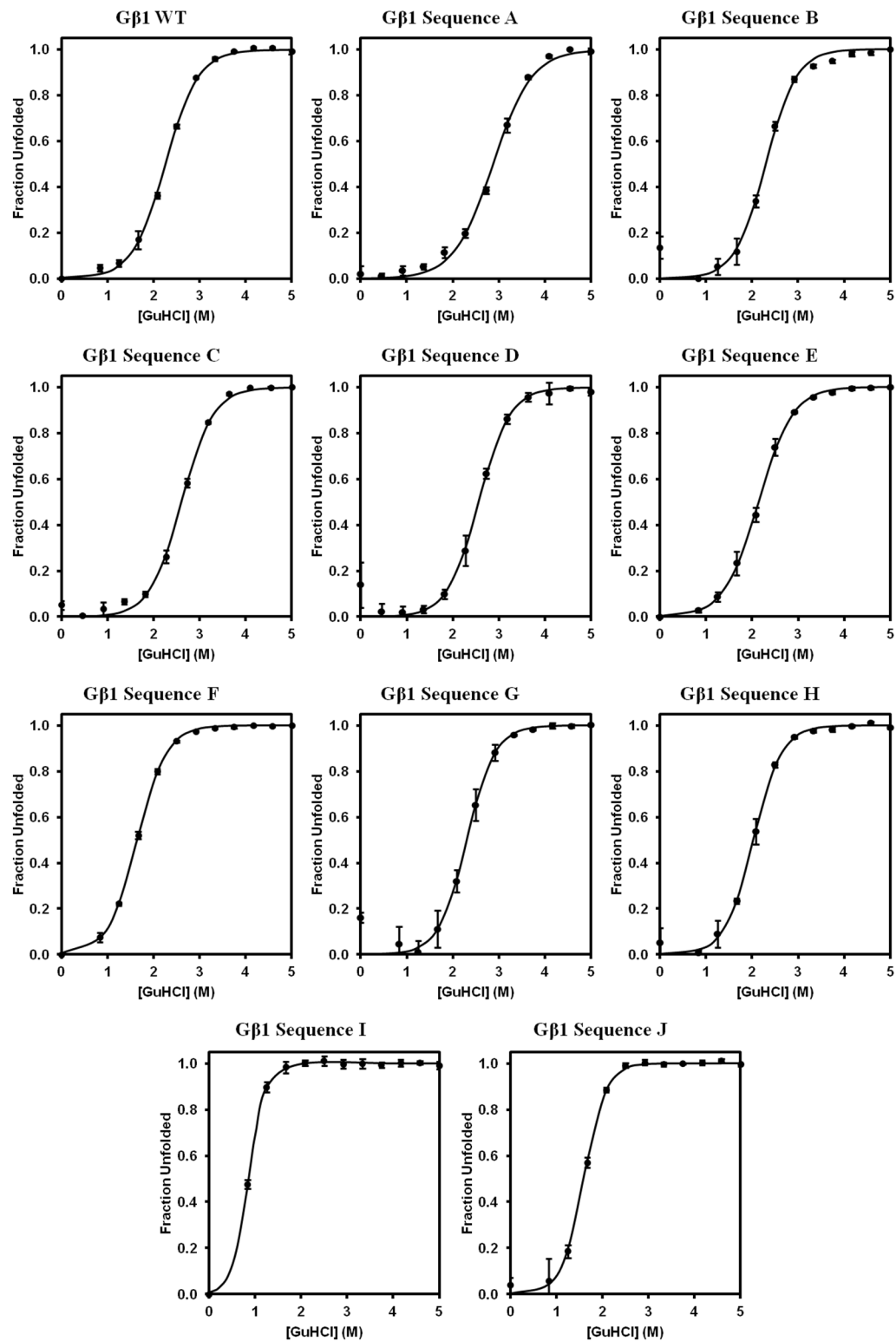


Figure 6.7. Chemical denaturation of G β 1 test set sequences. Chemical denaturation with guanidium chloride demonstrates that protein unfolding can be fit to a two state model for all sequences.

We next compared the experimental stability of the 10 test set sequences with predicted values using the regression parameters obtained for the training set using the PertMin₅₀ and MD1000K_{6.4} ensembles [Table 6.3]. As expected, negative design significantly improves correlation with experiment ($R^2 = 0.83$) compared to positive design ($R^2 = 0.70$) [Figure 6.8].

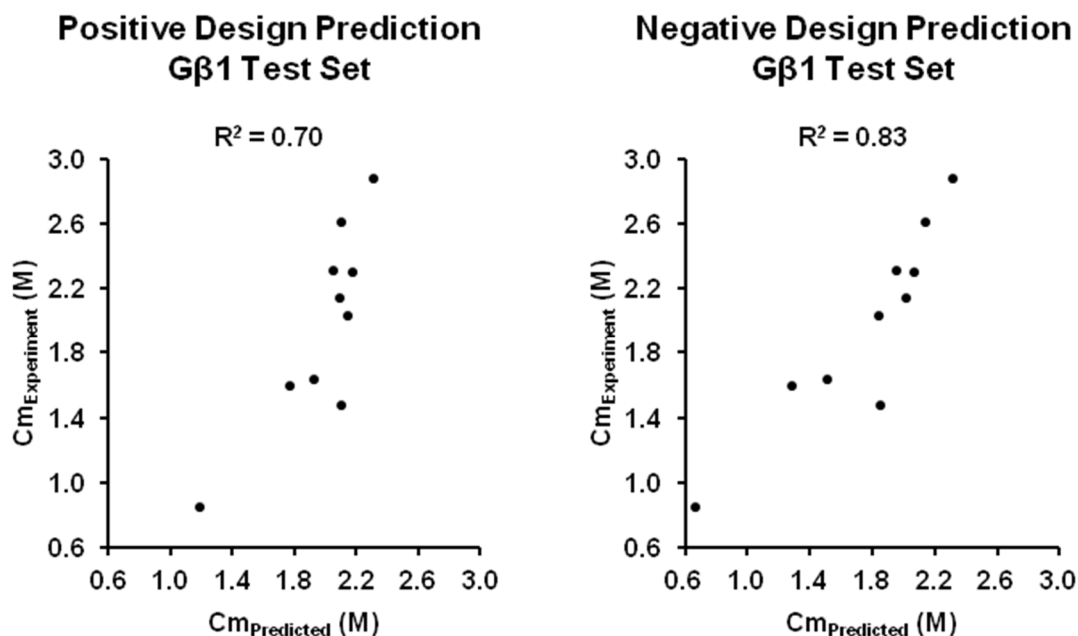


Figure 6.8. Prediction of Gβ1 test set sequence stabilities. Predictions of protein stability by positive and negative MSD using parameters obtained from regression of Gβ1 training set sequences with the PertMin₅₀ (positive state) and MD1000K_{6.4} (negative state) ensembles.

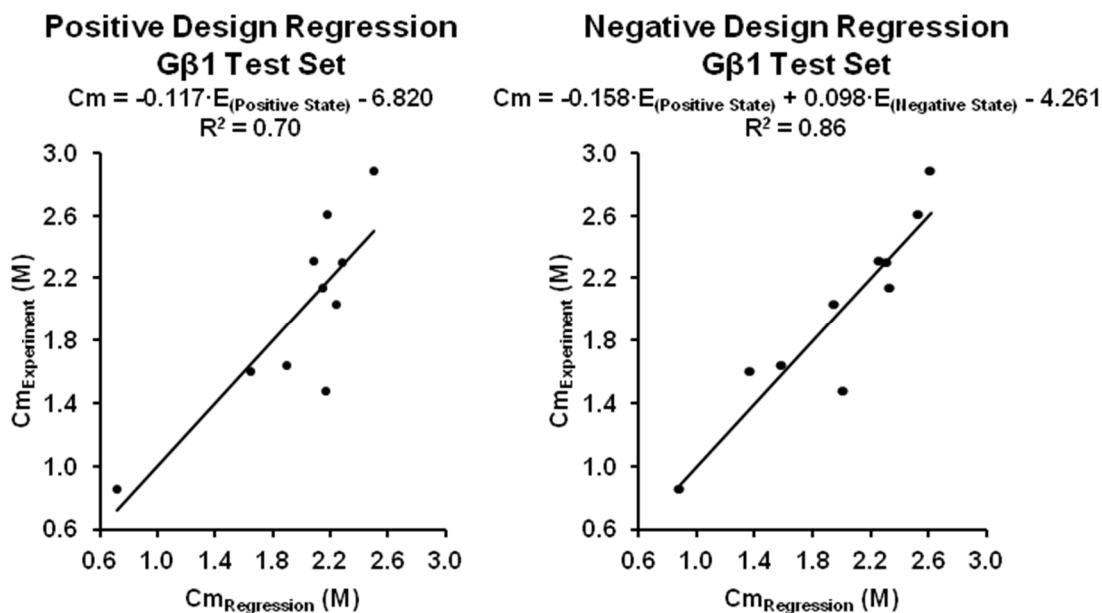


Figure 6.9. Regression of Gβ1 test set sequence stabilities. Correlation and regression of protein stability by positive and negative MSD giving new regression parameters for the PertMin₅₀ (positive state) and MD1000K_{6.4} (negative state) ensembles.

If regression analysis is performed with experimental data obtained from just the test set sequences [Figure 6.9], the negative design correlation is 0.86. The increased accuracy in predicted Cm values obtained by negative design is maintained, with approximately 80% of the 28 Gβ1 sequences (training and test sets) having their experimental Cm predicted within an error of 10%, a value significantly higher than that obtained using positive design (50%) [Figure 6.10]. These results further confirm that negative design provides more accurate predictions of protein stability than positive design.

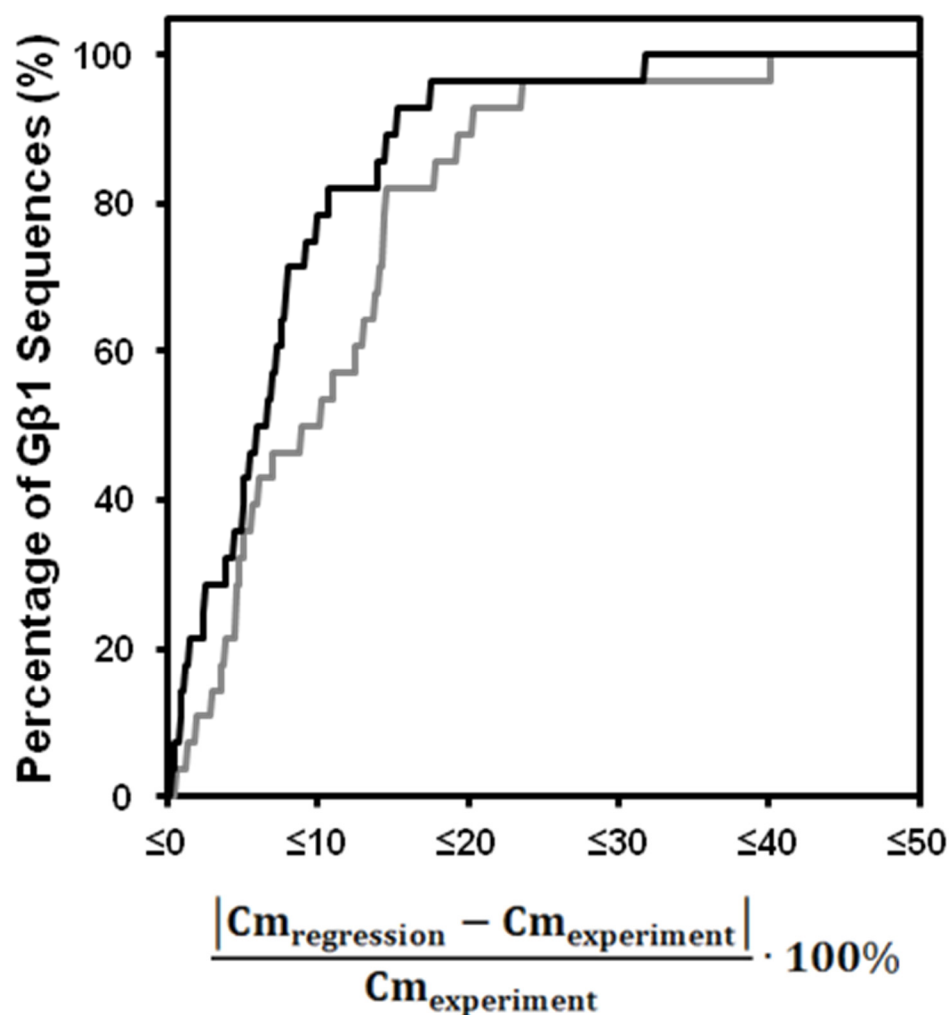


Figure 6.10. Accuracy of positive and negative MSD predictions. Correct prediction of C_m values for 28 mutant GB1 sequences (training and test sets) at a given error threshold is indicated for positive (grey) and negative (black) design. In negative design, the C_m values of fifty percent of GB1 sequences are predicted within an error of 6% while a value of roughly half is obtain with positive design (10%).

6.5. Stability Predictions Using Amino Acid Reference State Energies

The results described above suggest that off-target ensembles such as those prepared using high-temperature MD simulations can serve as useful negative states in MSD. Traditionally, peptide models have been used in CPD to derive amino acid

reference energies intended to approximate the stability of the unfolded state [Ali *et al.*, 2005; Busch *et al.*, 2008]. Because PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012], the CPD suite used here, does not employ reference energies and instead implicitly considers the energy of the unfolded state to be constant, we investigated whether use of amino acid reference energies derived from pentapeptide models of the unfolded state [Table 6.6] would yield improvements to correlation similar to those obtained by using fitness values computed on off-target ensembles.

Table 6.6. Amino acid reference energies obtained from pentapeptide models

Amino Acid	Energy ^a (kcal·mol ⁻¹)		
	Peptide _{6,4}	Peptide ₆₄	Peptide ₆₄₀
A	-0.6002	-0.6783	-0.5964
C	-1.1231	-1.4634	-1.2099
D	-1.4565	-5.3495	-5.7825
E	-3.6200	-4.3994	-11.6935
F	-2.3633	-3.2314	-2.7280
G	-0.1350	-0.1423	-0.1293
H	-4.3595	-7.2437	-5.4299
I	-2.0036	-2.1870	-1.7867
K	-5.4343	-7.5081	-5.5621
L	-1.7704	-2.1340	-1.7813
M	-2.5221	-2.7284	-2.2003
N	-2.7928	-5.5640	-4.5272
Q	-3.6573	-4.8969	-9.3046
R	-7.7290	-9.2388	-7.1308
S	-0.9084	-1.0455	-0.8698
T	-1.1927	-1.5989	-1.2578
V	-1.3931	-1.6685	-1.3646
W	-3.5646	-3.5589	-3.2540
Y	-2.4171	-3.4980	-2.8459

^a Reference energies are computed as the Boltzmann weighted average across 64-member pentapeptide ensembles

Table 6.7. Reference state energies for Gβ1 training set and test set sequences

Amino Acid	Energy ^a (kcal·mol ⁻¹)		
	Peptide _{6,4}	Peptide ₆₄	Peptide ₆₄₀
Training Set Sequences			
1	-19.6	-23.5	-20.0
2	-20.0	-23.9	-20.4
3	-19.0	-23.0	-19.5
4	-19.0	-23.0	-19.5
5	-20.2	-24.0	-20.4
6	-19.4	-23.4	-19.9
7	-19.6	-23.5	-20.0
8	-19.4	-23.4	-19.9
9	-19.4	-23.4	-19.9
10	-19.6	-23.5	-20.0
11	-18.8	-22.9	-19.5
12	-18.4	-22.4	-19.1
13	-18.8	-22.9	-19.5
14	-19.8	-23.9	-20.4
15	-19.0	-23.0	-19.5
16	-20.0	-23.9	-20.4
17	-19.2	-23.4	-19.9
18	-19.4	-23.4	-19.9
Test Set Sequences			
A	-19.7	-23.7	-20.1
B	-19.4	-23.7	-20.1
C	-19.1	-23.2	-19.6
D	-18.8	-22.5	-19.2
E	-19.4	-23.7	-20.1
F	-18.8	-22.3	-19.0
G	-19.1	-23.2	-19.6
H	-19.0	-22.4	-19.0
I	-19.4	-23.4	-19.9
J	-18.2	-21.8	-18.6

^a Reference energies are computed as the Boltzmann weighted average across 64-member pentapeptide ensembles

Amino acid reference energies intended to approximate the energetic contribution of the unfolded state to calculated stability were derived from ensembles created from a capped Gly-Gly-Ala-Gly-Gly pentapeptide model [Ding *et al.*, 2003]. A total of three 64-member pentapeptide ensembles were created by NVT MD simulations [Bond *et al.*, 1999; Sturgeon and Laird, 2000] at 300 K in explicit solvent (as described for G β 1 MD ensembles) sampled at 0.1, 1, and 10 ps increments, referred to as Peptide_{6,4}, Peptide₆₄, and Peptide₆₄₀, respectively. MSD on the three ensembles was performed by mutating *in silico* the pentapeptide Ala residue to all proteinogenic amino acids. This, the reference energy for an amino acid is the Boltzmann weighted average energy of that mutation in the context of one of the three pentapeptide ensembles. The application of these reference energies in negative design involves summing their values according to the sequence identity of G β 1 residues found at positions 3, 5, 7, 20, 26, 30, 34, 39, 43, 52, and 54 (numbering based on the 1PGA crystal structure [Gallagher *et al.*, 1994]). Reference state energies for training set and test set sequences are reported in Table 6.7.

Table 6.8. Regression of training set sequence with reference state energies

State 1	State 2	R ²	Regression Parameters			Parameter P-Values			F-Test P-Value
			Intercept	α	β	Intercept	α	β	
Positive Design									
	Peptide _{6,4}	0.0709	-1.0315	-0.1756		0.7419	0.2854		0.2854
	Peptide ₆₄	0.0150	0.2606	-0.0903		0.9521	0.6282		0.6282
	Peptide ₆₄₀	0.0062	0.9881	-0.0695		0.8239	0.7560		0.7560
Negative Design									
Backrub	Peptide _{6,4}	0.6295	-12.4515	-0.0945	-0.4050	0.0012	0.0003	0.0029	0.0006
Backrub	Peptide ₆₄	0.5432	-14.1998	-0.0967	-0.4044	0.0074	0.0008	0.0161	0.0028

Backrub	Peptide ₆₄₀	0.5141	-13.8079	-0.0960	-0.4574	0.0128	0.0013	0.0269	0.0045
PertMin ₅₀	Peptide _{6,4}	0.5371	-8.2990	-0.0779	-0.2268	0.0124	0.0015	0.0706	0.0031
PertMin ₅₀	Peptide ₆₄	0.4744	-7.9662	-0.0779	-0.1739	0.0616	0.0025	0.2329	0.0080
PertMin ₅₀	Peptide ₆₄₀	0.4594	-7.4022	-0.0776	-0.1773	0.0899	0.0029	0.3139	0.0099
PertMin ₁₅₀	Peptide _{6,4}	0.5087	-7.8704	-0.0574	-0.2877	0.0183	0.0023	0.0338	0.0048
PertMin ₁₅₀	Peptide ₆₄	0.4039	-6.8509	-0.0541	-0.2067	0.1168	0.0069	0.1928	0.0207
PertMin ₁₅₀	Peptide ₆₄₀	0.3804	-5.9909	-0.0530	-0.2040	0.1792	0.0088	0.2870	0.0276
MD100K _{6,4}	Peptide _{6,4}	0.2504	-5.4952	-0.0508	-0.2113	0.1584	0.0776	0.1754	0.1152
MD100K _{6,4}	Peptide ₆₄	0.1743	-3.9086	-0.0477	-0.1173	0.4209	0.1095	0.5097	0.2377
MD100K _{6,4}	Peptide ₆₄₀	0.1613	-3.0534	-0.0470	-0.0973	0.5345	0.1165	0.6486	0.2673
MD100K ₆₄	Peptide _{6,4}	0.1681	-3.3785	-0.0639	-0.0466	0.3486	0.2055	0.8030	0.2515
MD100K ₆₄	Peptide ₆₄	0.1746	-1.7489	-0.0809	0.0863	0.6835	0.1092	0.6746	0.2371
MD100K ₆₄	Peptide ₆₄₀	0.1857	-1.1367	-0.0852	0.1486	0.7927	0.0891	0.5418	0.2143
MD100K ₆₄₀	Peptide _{6,4}	0.1036	-2.1818	-0.0435	-0.0658	0.5411	0.4714	0.7682	0.4405
MD100K ₆₄₀	Peptide ₆₄	0.1120	-0.4978	-0.0743	0.1169	0.9079	0.2199	0.6355	0.4102
MD100K ₆₄₀	Peptide ₆₄₀	0.1274	0.1741	-0.0821	0.2008	0.9680	0.1695	0.4896	0.3600
MD500K _{6,4}	Peptide _{6,4}	0.2910	-4.8717	-0.0634	-0.1314	0.1603	0.0475	0.3786	0.0758
MD500K _{6,4}	Peptide ₆₄	0.2565	-3.6822	-0.0662	-0.0493	0.3975	0.0433	0.7696	0.1083
MD500K _{6,4}	Peptide ₆₄₀	0.2528	-3.0591	-0.0668	-0.0242	0.4891	0.0418	0.9045	0.1124
MD500K ₆₄	Peptide _{6,4}	0.0883	-2.8364	-0.0087	-0.2362	0.5482	0.6007	0.2519	0.4999
MD500K ₆₄	Peptide ₆₄	0.0154	-0.0449	-0.0013	-0.0994	0.9939	0.9364	0.6570	0.8899
MD500K ₆₄	Peptide ₆₄₀	0.0063	1.0890	0.0004	-0.0660	0.8536	0.9783	0.8026	0.9540
MD500K ₆₄₀	Peptide _{6,4}	0.0737	-1.5137	-0.0057	-0.1792	0.7041	0.8356	0.2938	0.5633
MD500K ₆₄₀	Peptide ₆₄	0.0155	0.0979	-0.0024	-0.0900	0.9840	0.9331	0.6408	0.8895
MD500K ₆₄₀	Peptide ₆₄₀	0.0066	0.8526	-0.0022	-0.0685	0.8625	0.9393	0.7675	0.9515
MD1000K _{6,4}	Peptide _{6,4}	0.0743	-1.6771	-0.0043	-0.1957	0.6961	0.8188	0.3071	0.5605
MD1000K _{6,4}	Peptide ₆₄	0.0158	0.5823	0.0020	-0.0817	0.9135	0.9127	0.6944	0.8873
MD1000K _{6,4}	Peptide ₆₄₀	0.0085	1.5054	0.0034	-0.0536	0.7794	0.8533	0.8276	0.9377
MD1000K ₆₄	Peptide _{6,4}	0.0719	-1.2759	-0.0016	-0.1839	0.7374	0.9027	0.3159	0.5715
MD1000K ₆₄	Peptide ₆₄	0.0164	0.5609	0.0019	-0.0816	0.9097	0.8858	0.6862	0.8833
MD1000K ₆₄	Peptide ₆₄₀	0.0090	1.3899	0.0026	-0.0561	0.7809	0.8394	0.8153	0.9343
MD1000K ₆₄₀	Peptide _{6,4}	0.0900	-0.9070	0.0315	-0.1908	0.7776	0.5829	0.2646	0.4928
MD1000K ₆₄₀	Peptide ₆₄	0.0246	0.4800	0.0220	-0.0935	0.9151	0.7061	0.6266	0.8294
MD1000K ₆₄₀	Peptide ₆₄₀	0.0149	1.2542	0.0209	-0.0701	0.7867	0.7212	0.7607	0.8936
Peptide _{6,4}	Peptide ₆₄	0.4260	11.3878	-1.8420	1.9147	0.0309	0.0051	0.0082	0.0156
Peptide _{6,4}	Peptide ₆₄₀	0.4260	9.7328	-1.4484	1.7823	0.0400	0.0047	0.0082	0.0156
Peptide ₆₄	Peptide ₆₄₀	0.4260	3.6407	-7.0483	8.3431	0.3176	0.0047	0.0051	0.0156

With the exception of negative design using the MD100K₆₄ and MD100K₆₄₀ ensembles as positive states, fitness values for training set sequences derived from pentapeptide reference energies [Table 6.6 and Table 6.7] add to the predicted stability of sequences [Table 6.8], indicating that they act as positive states when used with on- or off-target ensembles in negative design [Figure 6.3 and Table 6.8]. These results demonstrate that they cannot be used to approximate the unfolded state in our computational framework, suggesting that full-length models such as the non-native backbones included in the MD1000K_{6.4} ensemble represent a better computational proxy for this competing state [Figure 6.3 and Table 6.3]. This observation is in agreement with results by Pokala and Handel [Pokala and Handel, 2005], who showed that random protein backbone fragments of 10 amino acids or more extracted from hundreds of crystal structures are better reference states than shorter peptides, presumably because the unfolded state tends to be relatively compact, with transient sampling of longer-range interactions that may be better reflected by ensembles that retain some information about tertiary interactions in the global fold [Meng *et al.*, 2013; Mittag and Forman-Kay, 2007].

6.6. Stability Predictions for Alternate Protein Folds

Having demonstrated that we can design new G β 1 sequences that fold into the intended structure and improve the correlation between their predicted and experimental stabilities by using negative design, we next sought to investigate whether our computational framework could also be applied to recapitulate the stability of

proteins with alternate folds. Because the PertMin₅₀ and MD1000K_{6.4} ensembles were the positive and negative states that yielded the best correlation for G β 1 in negative design, we proceeded to generate 64-member PertMin₅₀ and MD1000K_{6.4} backbone ensembles using these methods for four additional proteins spanning a range of fold types, including: *Staphylococcus aureus* nuclease (PDB ID: 1STN) [Hynes and Fox, 1991], bacteriophage f1 gene V protein (PDB ID: 1VQB) [Skinner *et al.*, 1994], bacteriophage T4 lysozyme (PDB ID: 2LZM) [Weaver and Matthews, 1987], and barley seed chymotrypsin inhibitor 2 (PDB ID: 2CI2) [Mcphalen and James, 1987].

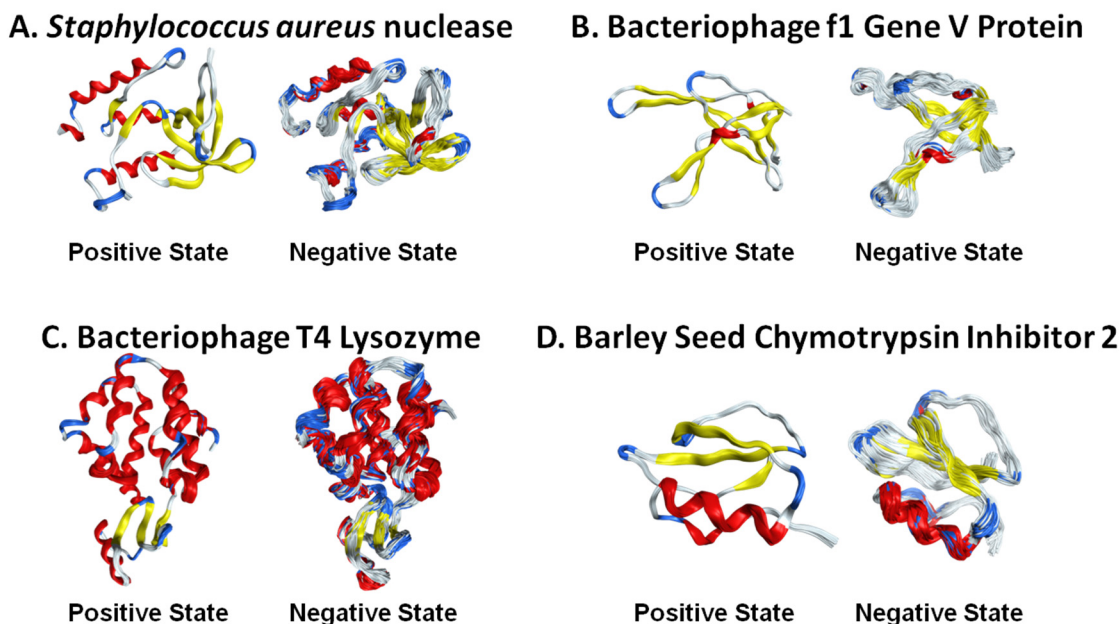


Figure 6.11. Positive and negative state ensembles for alternate protein folds. Positive (PertMin₅₀) and negative state (MD1000K_{6.4}) ensembles for: *Staphylococcus aureus* nuclease (A), bacteriophage f1 gene V protein (B), bacteriophage T4 lysozyme (C), and barley seed chymotrypsin inhibitor 2 (D). 64-member backbone ensembles used as positive and negative states for each protein are shown in cartoon representation and are colored according to their secondary structural elements: α -helices, β -strands, turns, and undefined secondary structures are colored in red, yellow, blue, and white, respectively.

These ensembles were generated using the procedures described previously with the exception that solvation of the crystallographic structures was completed by the addition of a water box having 6 Å depth from the protein surface, to reduce the size of the molecular systems improving the time to generate the set of PertMin and MD 1000 K ensembles. Using these ensembles [Figure 6.11], we computed fitness values for training sets of sequences of each protein [Table 6.9] and applied our computational framework to predict their stability.

Table 6.9. Calculated and experimental stabilities for alternate protein folds

Sequence	Cm (M)	M (kcal·mol ⁻¹ ·M ⁻¹)	Energy (kcal/mol) ^a	
			PertMin ₅₀	MD1000K _{6.4}
<i>Staphylococcus aureus</i> nuclease				
V23A ^b	0.33	1.18	-104.6	-60.5
L25A ^b	0.36	1.12	-102.2	-67.3
L25G ^b	0.12	1.18	-98.3	-66.7
Y27A ^b	0.18	1.10	-108.7	-61.2
F34A ^b	0.23	1.12	-99.7	-61.9
L36A ^b	0.26	1.13	-101.6	-57.1
L36G ^b	0.02	1.23	-113.3	-53.6
L37A ^b	0.63	0.89	-114.5	-72.7
L37G ^b	0.26	0.94	-108.5	-67.1
L38A ^b	0.60	0.94	-113.6	-63.6
L38G ^b	0.80	0.90	-113.2	-60.7
L39A ^b	0.49	0.99	-121.5	-91.8
L39G ^b	0.14	0.79	-116.8	-86.9
Y54A ^b	0.54	0.89	-105.3	-60.2
Y54G ^b	0.57	0.93	-101.7	-58.7
V66A ^b	0.44	1.10	-104.2	-58.9
V66G ^b	0.13	1.17	-98.8	-53.6
I72A ^b	0.05	1.29	-101.1	-62.5
I72V ^b	0.54	1.01	-108.8	-67.1
V74A ^b	0.31	1.13	-104.4	-69.2

L89A ^b	0.42	1.00	-106.1	-58.0
L89G ^b	0.34	0.96	-109.0	-54.5
Y91A ^b	0.02	1.01	-116.8	-65.8
I92V ^b	0.73	1.01	-108.9	-60.5
I92A ^b	0.18	1.18	-100.3	-61.4
V99A ^b	0.29	1.13	-103.2	-61.7
L103A ^b	0.13	0.96	-101.8	-61.3
V104A ^b	0.40	0.93	-106.2	-57.1
L125A ^b	0.12	0.66	-106.4	-76.0
Bacteriophage f1 gene V protein				
V35A/I47V ^c	1.50	-4.5	-6.7	-5.9
V35A/I47L ^c	1.85	-4.5	-6.4	-8.5
V35A ^c	2.01	-3.9	-9.5	-8.4
V35A/I47M ^c	1.47	-4.2	-10.0	-8.7
V35A/I47F ^c	1.69	-4.4	-1.9	-9.4
V35C/I47C ^c	0.99	-4.9	-8.3	-8.0
V35C ^c	2.21	-4.1	-6.0	-5.4
I47A ^c	0.87	-4.3	-9.8	-9.0
I47C ^c	1.33	-4.5	-12.3	-11.6
I47V ^c	1.91	-3.9	-12.3	-12.0
I47L ^c	2.43	-3.7	-15.1	-13.4
WT ^c	2.61	-3.6	-15.5	-11.2
I47M ^c	2.02	-3.9	-6.9	-14.3
I47F ^c	2.07	-4.0	-12.0	-12.1
V35L/I47V ^c	1.32	-4.2	-8.8	-11.6
V35L/I47L ^c	1.70	-4.3	-13.9	-14.6
V35L ^c	1.90	-4.5	-13.7	-11.9
V35L/I47M ^c	1.37	-5.0		-12.0
V35L/I47F ^c	1.60	-4.3	-12.7	-13.9
V35I/I47V ^c	1.80	-4.1	-8.6	-13.2
V35I/I47L ^c	2.27	-4.2	-15.3	-16.6
V35I ^d	2.44	-3.6	-15.9	-12.9
V35I/I47M ^c	1.87	-4.2	-7.4	-14.2
V35I/I47F ^c	2.05	-4.4	-13.6	-13.8
V35M/I47L ^c	2.14	-4.1	-17.1	-14.3
V35M ^c	2.33	-3.6	-17.7	-13.7
V35M/I47M ^c	1.65	-3.8	-8.3	-14.9

V35M/I47F ^c	1.99	-4.7	-14.6	
V35F/I47L ^c	1.51	-3.9	-15.4	
V35F ^c	1.76	-3.8	-15.4	
Bacteriophage T4 lysozyme				
WT ^d	6.3	-81.3	-30.4	
M6I ^d	3.9	-9.0	-32.6	
Y25G ^d	4.0	-68.9	-43.4	
C54Y ^d	3.8			
I58Y ^d	4.8			
L91P ^d	2.2	-15.7	-27.9	
A98T ^d	4.4	-64.6	-6.4	
A98V ^d	3.8	-54.7		
V103A ^d	5.3	-73.6	-28.9	
Q105G ^d	4.7	-81.8	-38.3	
W138Y ^d	4.9	-74.6	-34.2	
A146I ^d	4.2	-35.5		
A146T ^d	4.9	-72.7	-25.4	
A146V ^d	4.1	-67.4		
V149A ^d	4.9	-72.7	-38.8	
F153C ^d	4.7	-70.8	-29.6	
A98V/T152S ^d	3.6	-61.9		
Barley seed chymotrypsin inhibitor 2				
L8A ^e	2.62	2.15	-39.8	-29.7
G16A ^e	3.44	1.80	-53.6	-44.6
I20V ^e	3.33	1.99	-45.6	-36.0
K24A ^e	3.67	1.57	-38.7	-36.6
K24G ^e	2.36	1.75	-33.9	-33.9
V47A ^e	1.46	1.76	-41.4	-32.0
L49A ^e	2.01	1.98	-38.9	-30.3
I57V ^e	4.10	1.82	-44.3	-36.6
I57A ^e	1.79	1.93	-37.3	-29.4

^a Ensemble Boltzmann weighted average energy for each sequence is omitted if the sequence energy is greater than 0 kcal·mol⁻¹

^b [Shortle *et al.*, 1990]

^c [Sandberg and Terwilliger, 1991]

^d [Klemm *et al.*, 1991]

^e [Itzhaki *et al.*, 1995]

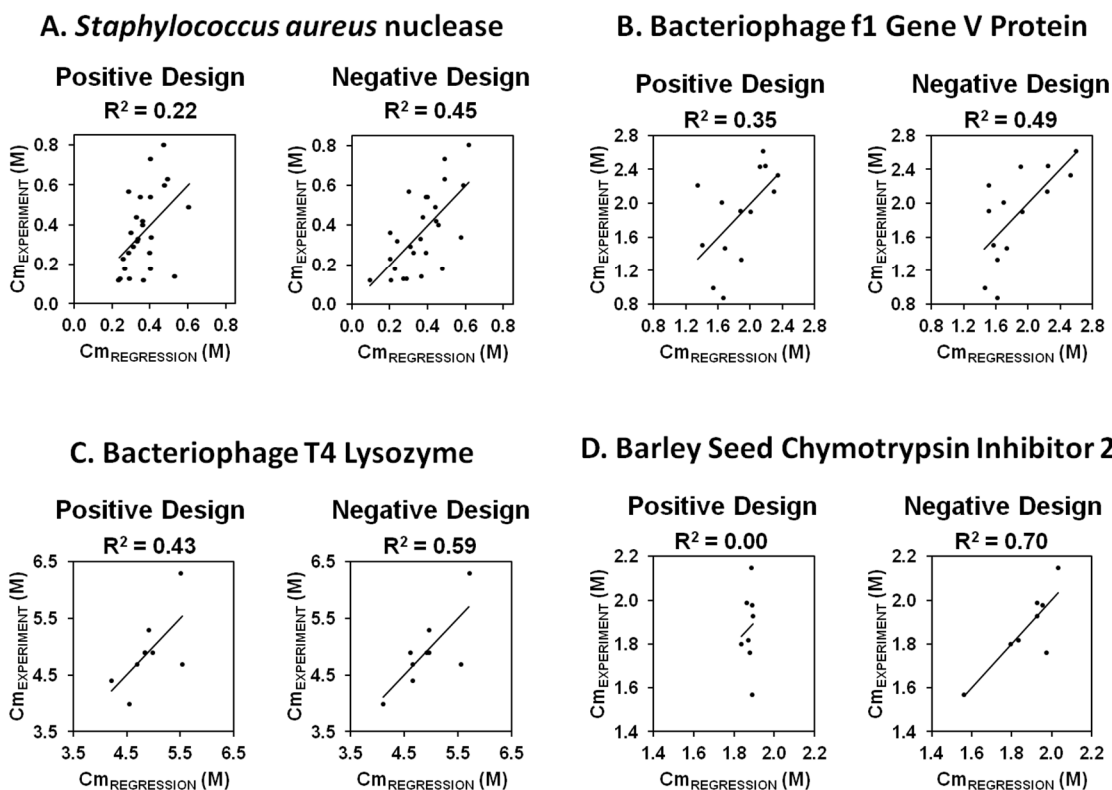


Figure 6.12. MSD correlations of stability for alternate protein folds. Stability correlations by positive and negative design using positive state (PertMin₅₀) and negative state (MD1000K_{6.4}) ensembles are reported for alternate proteins folds: *Staphylococcus aureus* nuclease (A), bacteriophage f1 gene V protein (B), bacteriophage T4 lysozyme (C), and barley seed chymotrypsin inhibitor 2 (D).

Regression parameters for each protein with their respective training set sequences are reported in Table 6.10. Correlation of experimental stabilities with those predicted using either positive or negative design [Figure 6.12 and Table 6.10] shows that in all cases, negative design results in strong to very strong correlations ($R^2 = 0.45 - 0.70$ or $R = 0.67 - 0.84$), an improvement over positive design. Additionally, in all cases, the PertMin₅₀ and MD1000K_{6.4} ensembles act as positive and negative states, respectively. Impressively, correlation coefficients obtained for two of the four proteins (bacteriophage T4 lysozyme and chymotrypsin inhibitor 2) are similar to those achieved

for G β 1 (0.59 and 0.70, respectively) even though we did not optimize the structure of the positive and negative state ensembles used as input templates for these proteins in negative design. These results demonstrate that our computational framework based on negative design is generally applicable to proteins of different structural characteristics.

Table 6.10. Stability regression parameters for alternate protein folds

Regression	R ²	intercept	α (PertMin ₅₀)	β (MD1000K _{6.4})
<i>Staphylococcus aureus</i> nuclease				
Positive Design	0.22	-1.328	-0.016	
Negative Design	0.45	-1.904	-0.030	0.014
Bacteriophage f1 gene V protein				
Positive Design	0.35	0.837	-0.085	
Negative Design	0.49	1.280	-0.267	0.254
Bacteriophage T4 lysozyme				
Positive Design	0.43	-0.733	-0.077	
Negative Design	0.59	-1.715	-1.715	0.026
Barley seed chymotrypsin inhibitor 2				
Positive Design	0.00	1.824	-0.001	
Negative Design	0.70	1.919	-0.048	0.060

6.7. Comparative Study of Alternate Stability Prediction Strategies

The computational framework described here involves the evaluation of sequence stability in the context of a variety of native and non-native ensembles intended to approximate the folded and unfolded states in a negative design approach. Typically, negative design aims to identify sequences that stabilize the positive state while simultaneously destabilizing the negative state. Because negative design optimizes the energy difference between the positive and negative states, sequences that maximize

unfavorable interactions such as steric clashes in the negative state tend to be preferred, leading to inflated energies that may not be physically meaningful. Although the application of energy thresholds applied to unfavorable interactions during optimization can partially alleviate this behaviour [Bolon *et al.*, 2005], the accurate selection of relevant sequences for the negative state can be difficult. To circumvent this issue, our computational framework utilizes MSD in a positive design approach to identify sequences that stabilize both the positive and the negative states. In this way, sequence stability is evaluated in our framework post MSD, resulting in meaningful values that can be used to identify sequences that are more stable in the positive than the negative state instead of focusing on identifying sequences that are highly unstable in the negative state. This approach, analogous to the “specificity sweep” procedure of the CLASSY framework [Grigoryan *et al.*, 2009], enables the identification and rejection of sequences that would be more stable on the negative state than on the positive state.

With our computational framework, we were able to achieve correlations between predicted and experimental stabilities with R^2 of up to 0.86 ($R = 0.93$). To our knowledge, this represents the first successful application of negative design with native and non-native backbone ensembles to the design of protein stability. Several approaches have been developed to improve the correlation of predicted and experimental stabilities [Kellogg *et al.*, 2011; Pokala and Handel, 2005; Yin *et al.*, 2007a], but these all involved modification of the potential energy function to recapitulate previously known experimental data. Unlike these methods, our computational

framework does not require force-field fine-tuning or systematic evaluation of a broad range of conformational sampling methodologies. However, as is the case with these other methods, a limitation of our framework is that solution of the regression parameters requires a training set of sequences with known stability. Perhaps the approach most similar to our work is the backbone drift modeling method developed by Wilson and coworkers [Howell *et al.*, 2014]. Backbone drift modeling is a multistate approach to CPD where sequence stability is evaluated on a set of available WT and mutant protein crystal structure templates. In this approach, the most favorable template to score each sequence is identified, allowing for a better correlation with its experimentally measured stability. Using backbone drift modeling in combination with an optimized scoring function, Wilson and coworkers were able to achieve a correlation of predicted and experimental stabilities for a training set of 100 mutant sequences of adenylate kinase with an R^2 of 0.89, a significant improvement compared to the correlation obtained by using a single backbone ($R^2 = 0.49$) [Howell *et al.*, 2014].

To evaluate whether backbone drift modeling could provide similar enhancements in correlations using our G β 1 data set, we created a positive state ensemble comprising 17 G β 1 backbone structures deposited in the PDB and used it to compute the energy of the 18 training set sequences. In backbone drift modeling, the energy of a sequence is the most favorable energy obtained from each of the backbone templates included in the ensemble. A total of 17 single-chain G β 1 protein structures for use in backbone drift modelling were retrieved from the PDB. Specifically, these structures were derived from a variety of x-ray, solution NMR, and solid state NMR PDB depositions: 1EM7 [Strop *et*

al., 2000], 1IGD [Derrick and Wigley, 1994], 1PGA [Gallagher *et al.*, 1994], 1PGB [Gallagher *et al.*, 1994], 2GB1 [Gronenborn *et al.*, 1991], 2IGD [Butterworth *et al.*, 1997], 2IGH [Lian *et al.*, 1992], 2J52 [Wilton *et al.*, 2008], 2J53 [Wilton *et al.*, 2008], 2KOP [Robustelli *et al.*, 2008], 2LGI [Wylie *et al.*, 2011], 2RMM monomers A and B [Jee *et al.*, 2007], 2RPV [Saio *et al.*, 2009], 3GB1 [Kuszewski *et al.*, 1999], and 3MP9 monomers A and B [Tomlinson *et al.*, 2010]. Following the removal of crystallographic water molecules, these 17 structures were prepared by the addition of hydrogen atoms, solvent (H₂O cube of 50 Å in length), and counter-ions (Na⁺ and Cl⁻) using the Protonate 3D utility [Labute, 2009] available in the MOE software package [MOE, 2012]. The prepared structures were then subjected to 0, 50, 100, or 150 iterations of conjugate gradient energy minimization using the AMBER99 force field [Wang *et al.*, 2000], resulting in four 17-member ensembles.

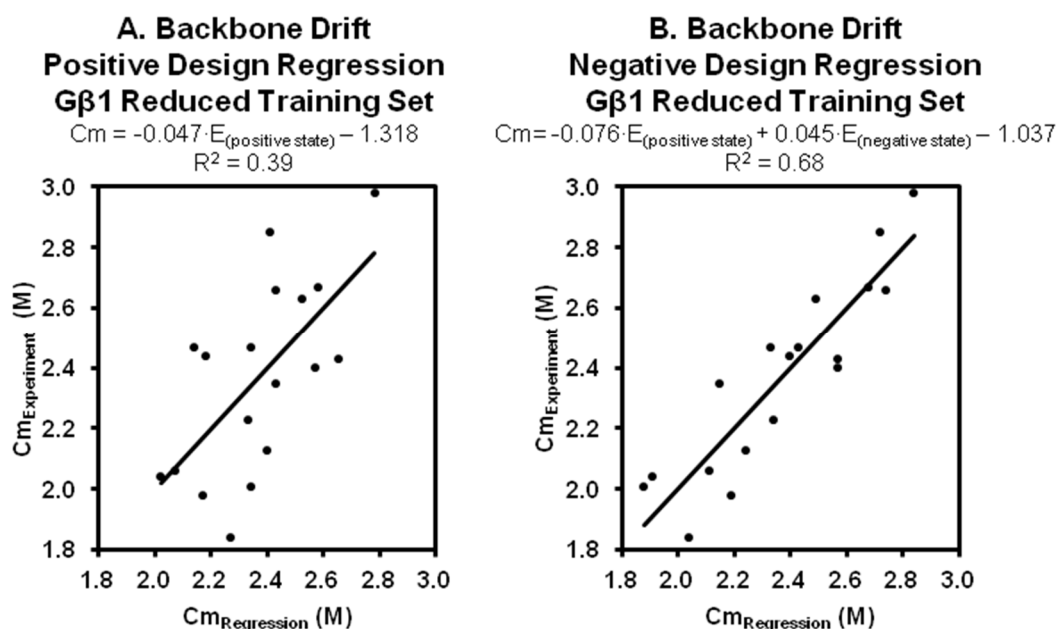


Figure 6.13. MSD of Gβ1 training set stability with backbone drift modeling

The resulting PDB ensembles were used as input templates to MSD with the exception that sequence fitness was the minimum energy obtained from each of the backbone templates included in the ensemble. Backbone drift modeling to recapitulate the stability of the 18 training set sequences yielded correlation coefficients (R^2) of 0.39, 0.03, 0.05, and 0.03 for PDB ensembles prepared with 0, 50, 100, and 150 iterations of energy minimization, respectively. For the purpose of this study, the best correlation obtained by backbone drift ($R^2 = 0.39$) is discussed [Figure 6.13]. By applying linear regression to the energy values obtained by backbone drift modeling and the experimental C_m values, we obtained a correlation of predicted and experimental stability with an R^2 of 0.39 [Figure 6.13.A], a value significantly lower than that obtained by Wilson and coworkers. However, this correlation coefficient is similar to the one that we obtained using positive MSD with the PertMin₅₀ ensemble ($R^2 = 0.42$) [Figure 6.4], demonstrating that an ensemble prepared *in silico* can perform as well in positive MSD as an ensemble derived from experimental data. When we apply our negative design approach by using the MD1000K_{6.4} ensemble in combination with the PDB ensemble, we significantly improve the correlation to $R^2 = 0.68$ [Figure 6.13.B], further demonstrating that negative design leads to improved prediction accuracy without the need for energy function or conformational sampling fine-tuning. It is possible that the lower correlation coefficient that we observe when we apply backbone drift modeling to G β 1 is caused by the use of a different protein and ensemble of structures. However, Wilson and coworkers achieved their stronger correlation by optimizing the potential energy function and remodeling the backbone geometry of ensemble members during

backbone drift modeling. Thus, additional improvements to correlation may be achieved by incorporation of these procedures into our negative design framework.

6.8. Features of Physical and Predictive Models

The results described above highlight how the successful correlation of predicted and experimental protein stabilities using our computational framework relies on the use of appropriate native and non-native ensembles to act as positive and negative states in negative design. Since not all off-target ensembles perform equally well as negative states [Figure 6.3], we were interested in identifying structural features that may be important for non-native ensembles to serve as useful reference states in negative design. Thus, we calculated the backbone RMSD to crystal structure, the solvent accessible surface area, and the secondary structure composition for each ensemble [Table 6.11]. As expected, on-target ensembles of G β 1 (PertMin, Backrub, and MD100K) have low backbone RMSD from crystal structure (0.46 – 0.73 Å) while off-target ensembles (MD500K and MD1000K) have significantly increased deviation (1.03 – 12.6 Å). Similarly, on-target ensembles retain a higher percentage of G β 1 secondary structure (94 – 99%) than off-target ensembles (21 – 88%). These trends are also observed with the four alternate protein folds [Table 6.11].

Interestingly, negative design with the PertMin₅₀ positive state ensemble and the MD1000K_{6.4} or MD1000K₆₄ negative state ensembles yielded the best correlations with $R^2 \geq 0.80$ [Figure 6.3]. However, the MD1000K₆₄₀ ensemble acted as a positive state when used in conjunction with the PertMin₅₀ ensemble, resulting in much weaker

correlation ($R^2 = 0.44$). Similar results were obtained by using the Backrub, PertMin₁₅₀, and MD100K_{6.4} ensembles with the three MD1000K ensembles [Figure 6.3]. Structurally, the MD1000K₆₄₀ ensemble differs significantly from the other MD1000K ensembles in that it is less compact and more conformationally diverse [Figure 6.2], as illustrated by drastically increased solvent accessible surface area ($3725 \text{ \AA}^2$ and $6226 \pm 350 \text{ \AA}^2$ for the crystal structure and MD1000K₆₄₀ ensemble, respectively) and backbone coordinate deviation ($12.6 \pm 1.6 \text{ \AA}$ RMSD from crystal structure), as well as significantly reduced retention of secondary structure ($21 \pm 32\%$).

Table 6.11. Structural characterization of backbone ensembles

Ensemble	RMSD _{XTAL} ^b (Å)	Solvent Accessible Surface Area (Å ²)	Retained Secondary Structure (%)	Average Secondary Structure (%)			
				α-helix	β-strand	Turn	Undefined
<i>Streptococcal</i> protein G Domain β1							
1PGA ^a		3725		25	43	11	21
PertMin ₅₀	0.46 ± 0.02	3703 ± 25	95 ± 19	30	43	6	21
PertMin ₁₅₀	0.53 ± 0.03	3673 ± 29	95 ± 20	30	43	5	21
Backrub	0.52 ± 0.10	3965 ± 71	99 ± 6	25	43	11	21
MD100K _{6.4}	0.73 ± 0.02	3759 ± 18	96 ± 17	29	43	6	21
MD100K ₆₄	0.72 ± 0.02	3759 ± 21	96 ± 17	29	43	6	21
MD100K ₆₄₀	0.71 ± 0.02	3764 ± 17	94 ± 19	29	41	6	23
MD500K _{6.4}	1.03 ± 0.08	3754 ± 34	88 ± 21	26	40	11	23
MD500K ₆₄	1.13 ± 0.12	3747 ± 43	88 ± 20	24	39	13	25
MD500K ₆₄₀	1.21 ± 0.13	3823 ± 92	87 ± 20	23	40	13	24
MD1000K _{6.4}	1.57 ± 0.25	3741 ± 54	80 ± 26	23	32	15	30
MD1000K ₆₄	2.34 ± 0.55	3880 ± 120	66 ± 28	14	27	18	41
MD1000K ₆₄₀	12.6 ± 1.6	6230 ± 350	21 ± 32	7	1	20	72
<i>Staphylococcus aureus</i> nuclease							
1STN ^a		7932		26	30	18	26
PertMin ₅₀	0.50 ± 0.02	7928 ± 36	95 ± 21	27	28	16	29
MD1000K _{6.4}	2.47 ± 0.09	8734 ± 104	59 ± 36	12	15	25	48
Bacteriophage f1 gene V protein							
1VQB ^a		6552		7	47	7	40
PertMin ₅₀	0.53 ± 0.02	6635 ± 32	99 ± 8	6	45	7	42
MD1000K _{6.4}	3.58 ± 0.08	6936 ± 137	66 ± 37	2	26	8	64
Bacteriophage T4 lysozyme							
2LZM ^a		8419		66	9	7	18
PertMin ₅₀	0.50 ± 0.02	8640 ± 41	93 ± 24	59	8	14	18
MD1000K _{6.4}	3.10 ± 0.12	9574 ± 135	59 ± 34	16	7	31	48
Barley seed chymotrypsin inhibitor 2							
2CI2 ^a		4591		22	22	15	42
PertMin ₅₀	0.50 ± 0.02	4588 ± 29	89 ± 31	21	21	15	42
MD1000K _{6.4}	1.86 ± 0.15	4736 ± 82	69 ± 33	13	13	18	56

^a PDB ID of crystal structures

^b Backbone RMSD from crystal structure

These results, coupled to what we observed with the amino acid reference energies, suggest that for a negative state to be useful in our negative design framework, it must retain a significant amount of folded structure. This is likely due to the potential energy function used here, which is parameterized to score folded protein structures but may be unable to accurately compute the energy of a fully unfolded protein.

Nevertheless, structural analysis of the negative state ensembles used here suggests that to be useful as reference states in our negative design framework, they should meet the following three criteria: (i) their backbone RMSD from crystal structure should be greater than 1 Å in order to be off-target [Davey and Chica, 2014], (ii) they should have reduced secondary structure compared to the crystal structure (50 – 90%), and (iii) they should have comparable solvent accessible surface area to the crystal structure in order to be properly scored by our potential energy function. Based on these observations, it is unlikely that our negative state ensembles are accurate atomistic models of real-world unfolded protein states. Instead, these ensembles may represent high-energy accessible states that serve as useful reference states in negative design calculations. Regardless, the computational framework presented here can be adapted to screen hypothesized models of the unfolded state, identifying those that both conform to the ideal structural characteristics and provide predicted stabilities in agreement with experiment.

6.9. Conclusions on the Application of Negative Multistate Design

We have developed a computational framework based on negative design with native and non-native backbone ensembles to accurately predict the stability of protein sequences. To our knowledge, this represents the first example of explicit negative design using backbone ensembles as negative states. Our computational framework enables very strong correlations between experimental and calculated stability for various fold types. Incorporation of this framework into future design strategies focussed on the development of novel protein functions holds great promise for the enhancement of prediction accuracies. Ultimately, the efficiencies gained by more reliable predictions of protein stability have the potential to reduce the scale of experimental screens that follow CPD and increase the range of functionalities that can be designed, taking us one step closer to an age of designer proteins ‘on-demand’.

CHAPTER 7

Development and Application of *meta*-Multistate Design: the Rational Design of Protein Dynamics

Prediction of protein conformational equilibrium and dynamics with atomic-level accuracy and quantitative energetic precision are important objectives for CPD. Traditional CPD methodologies evaluate sequences in the context of a single backbone template and do not consider these properties when engineering complex functions. To date, a general framework for the structure-based rational design of protein dynamics has yet to be developed. As a proof-of-concept, we predicted and experimentally validated sequences that stabilize and facilitate exchange between two non-native sidechain conformations of Trp43 in the G β 1 fold using an ensemble based MSD strategy. Four candidate G β 1 sequences were identified by MSD across a sequence space of 1,296 mutants, predicted to exchange between a two non-native conformations of Trp43 with various barrier heights (7-12 kcal·mol⁻¹). Study by ¹⁵N-HSQC and ZZ-exchange NMR spectroscopy demonstrates that three of the four candidates are dynamic, with two candidates exchanging between two distinct conformations on the 10 to 100 millisecond time scale. NMR solution models structures of a dynamic sequence and additional non-dynamic G β 1 variants suggest good agreement between calculation and experiment. The application of our *meta*-MSD strategy may facilitate the incorporation and consideration of protein dynamics in future design efforts toward increasingly complicated protein functions.

Two of the sections in this chapter include language and figures that were adapted from a manuscript in preparation coauthored with Adam M Damry, Dr. Natalie K Goto, and Dr. Roberto A. Chica. The following sections share co-authorship:

7.5. Experimental Characterization of DANCER Proteins

7.6. Validation of DANCER Conformational States

The manuscript in preparation is:

Davey JA[‡], Damry AM[‡], Goto NK^{*}, and Chica RA^{*} (2016) Computational design of dynamics into a stable globular protein. *Manuscript in Preparation*.

[‡]denotes co-first authorship of the manuscript

^{*}denotes co-corresponding authorship of the manuscript

I would like to formally acknowledge the contributions of these authors to the completion of this chapter. Specifically:

James A Davey: Calculation design, predictions, and comparisons with experiment, as well as experimental design excluding NMR

Adam M Damry: Experimental results

Dr. Natalie K Goto: Experimental design

Dr. Roberto A Chica: Experimental design

7.1. The Importance of Protein Dynamics

The dynamic behaviours of proteins pose a series of central and fundamental questions to physical biology. Current studies demonstrate that despite large differences in protein structure, the distinct functions of different protein folds can be influenced by thermal attenuation of common protein transitions [Daniel *et al.*, 1998; Frauenfelder *et al.*, 2009; Frauenfelder *et al.*, 1991; Rasmussen *et al.*, 1992]. And while it is understood that proteins exhibit an array of complex and coincidental hierarchical motions [Lewandowski *et al.*, 2015], their relationship to the function and evolution of proteins in general remains poorly understood. As a result, a coherent understanding consistent with results from a variety of protein dynamics studies using MD calculations [Tarek and Tobias, 2002; Vitkup *et al.*, 2000] as well as numerous different experimental methodologies, including: neutron scattering [Doster *et al.*, 1989], terahertz spectroscopy [Knab *et al.*, 2007], Mössbauer spectroscopy [Frauenfelder *et al.*, 2009; Parak and Formanek, 1971], dielectric spectroscopy [Frauenfelder *et al.*, 2009], x-ray crystallography [Rasmussen *et al.*, 1992; Weik and Colletier, 2010], and differential scanning calorimetry [Jansson *et al.*, 2011], has eluded the protein science community [Doster, 2008; Fenimore *et al.*, 2013].

Progress into the elucidation of the origin and functional implication of protein dynamics has been limited to the characterization of naturally occurring proteins and a small number of their mutant sequences. Ideally, sequence perturbations could be created at will having desired predictable effects on the dynamics of any given fold. Unfortunately, the complex nature of protein motions and the challenges associated

with their detailed modelling [Heath *et al.*, 2007] as well as the extensive experimentation required for their characterization limits the throughput and application of sequence mutagenesis methodologies. Because dynamics are implicated in a variety of protein functions involving conformational exchange, such as allostery [Guo and Zhou, 2016] and catalysis [Bhabha *et al.*, 2011], the ability to rationally design protein dynamics would serve as a valuable protein engineering tool.

With this objective in mind, we report in this chapter the first successful protein engineering effort directed toward the rational design of a protein that spontaneously interconvert between two designed states of a globular protein fold. The computational design of protein dynamics required the development of a new approach, referred to as *meta*-multistate design (*meta*-MSD), which builds upon current CPD methodologies [Davey and Chica, 2012] allowing for the physics based rational screening of sequences *in silico* on a scale that is experimentally impossible to achieve [Chica *et al.*, 2005]. Use of this *meta*-MSD approach, enabled the *in silico* search for dynamic sequences across a sequence space (1,296 total sequences) using the G β 1 scaffold [Gallagher *et al.*, 1994] for a previously unobserved exchange of the tryptophan 43 sidechain [Gronenborn *et al.*, 1991; Lewandowski *et al.*, 2015] between two non-native conformations. The result was the creation of four new proteins referred to as DANCER-0, DANCER-1, DANCER-2, and DANCER-3, so named for their predicted behaviour as Dynamic Native Conformation ExchangeR (DANCER) proteins. This series of DANCER proteins were subjected to experimental characterization to determine that they are stable at room temperature, adopt the G β 1 fold, and unfold both in the presence of chemical

denaturant (guanidium chloride) and by thermal denaturation, according to a two-state model. In cooperation with *Adam M Damry* and *Dr. Natalie K Goto*, solution NMR study of the DANCER proteins reveal that DANCER-1, DANCER-2, DANCER-3 exchange as predicted in agreement with the design models, on the millisecond time-scale. Specifically, while DANCER-0 does not adopt a single stable and observable state, as evidenced by significant peak broadening in its ^1H - ^{15}N -HSQC spectrum, DANCER-1, DANCER-2, and DANCER-3 proteins are observed in two distinct and exchanging conformational states, with ZZ-exchange NMR spectroscopy allowing for the determination of exchange kinetics of DANCER-1 and DANCER-3 proteins on the 10 and 100 millisecond time scales, respectively. The ability to rationally and reliably engineer dynamic motions at will into any protein scaffold promises to enable its consideration in the future study and design of complex protein functions, such as catalysis and association, with atomic level detail and energetic precision.

7.2. The Design Objective, Hypothesis, and Rationale

G β 1 was chosen as the donor protein scaffold [Gallagher *et al.*, 1994] for our rational design of dynamic protein sequences for two reasons. The first reason is that the scaffold is small, consisting of only 56 residues making it amenable to study with atomic resolution by NMR spectroscopy [Gronenborn *et al.*, 1991]. The second reason is that the WT protein contains a single tryptophan at position 43, useful for experimental characterization, that is observed to consistently occupy the same single conformation across a variety of structure characterization studies [Butterworth *et al.*, 1997; Derrick

and Wigley, 1994; Gallagher *et al.*, 1994; Gronenborn *et al.*, 1991; Jee *et al.*, 2007; Kuszewski *et al.*, 1999; Lian *et al.*, 1992; Robustelli *et al.*, 2008; Saio *et al.*, 2009; Strop *et al.*, 2000; Tomlinson *et al.*, 2010; Wilton *et al.*, 2008; Wylie *et al.*, 2011]. The objective of the work presented in this chapter involves the explicit design of the exchange of this tryptophan between two non-native conformations not yet observed experimentally.

The computational design of a dynamic protein that spontaneously interconverts between two distinct states requires the consideration of an ensemble of protein configurations along the trajectory of the reaction coordinate. Such a trajectory can be represented by a continuum of atomic configurations, referred to as sub-states that the dynamic protein adopts during the course of its exchange between stable protein conformations. To simplify this reaction coordinate, dynamic exchange can be conceptually restricted to three states: (1) a major state, which is the most populated end-state found in the dynamic trajectory, (2) a minor state, which is the alternate end-state conformation of reduced population occupancy and higher in energy than the major state, and (3) the transition state, which is the highest energy conformation that is transiently adopted as the polypeptide undergoes exchange between the major and minor states. By conceptualizing the exchange reaction with this formalism, the task of predicting dynamic protein sequences using a CPD framework can be simplified. Specifically, a framework that allows for the partitioning of sequence threaded ensemble member states (sub-states) into the defined set of the major, minor, and transition conformations (states) according to both their energy and geometry should

enable the design and evaluation of dynamic protein sequences [Dill, 1999; Keedy *et al.*, 2015].

The CPD of the dynamic exchange of tryptophan 43 between two non-native conformations into the rigid G β 1 scaffold requires that a set of design constraints be defined and expressed as a set of geometric configurations unique to each of the states. Because the relevant atomic structures have not been previously observed, suitable models of the desired states must be created. At the very least, geometric constraints for three states are required to explicitly design the dynamic exchange a two state transition. It is possible that many molecular systems, like that of the tryptophan sidechain, can have many *df* exceeding the desired set of three states required to define the exchange coordinate. To ensure that dynamic sequences are designed to have a specific desired exchange, it is important to consider all alternative undesired sub-states and states that may be accessible to each protein sequence variant.

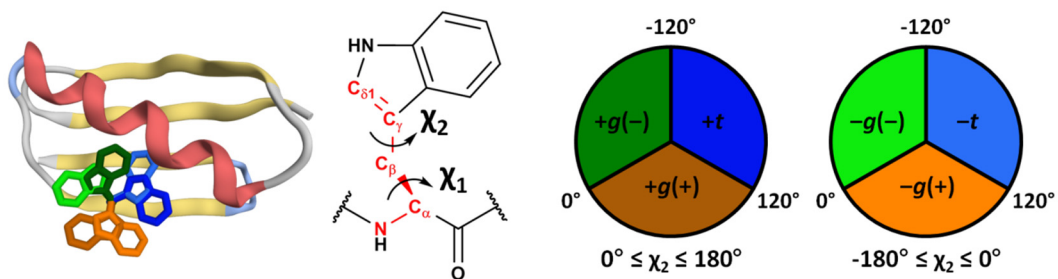


Figure 7.1. Tryptophan conformations for G β 1 design states. The geometric constraints for χ_1 (N-C α -C β -C γ) and χ_2 (C α -C β -C γ -C δ_1) dihedral angles of the tryptophan 43 residue used to partition design sub-states into their respective states. The six states are named: *plus gauche*⁻ [+g(-)] ($0^\circ \leq \chi_2 \leq 180^\circ$ and $-120^\circ \leq \chi_1 \leq 0^\circ$), *plus trans* [+t] ($0^\circ \leq \chi_2 \leq 180^\circ$ and either $-180^\circ \leq \chi_1 \leq -120^\circ$ or $-120^\circ \leq \chi_1 \leq 180^\circ$), *plus gauche*⁺ [+g(+)] ($0^\circ \leq \chi_2 \leq 180^\circ$ and $0^\circ \leq \chi_1 \leq 120^\circ$), *minus gauche*⁻ [-g(-)] ($-180^\circ \leq \chi_2 \leq 0^\circ$ and $-120^\circ \leq \chi_1 \leq 0^\circ$), *minus trans* [-t] ($-180^\circ \leq \chi_2 \leq 0^\circ$ and either $-180^\circ \leq \chi_1 \leq -120^\circ$ or $-120^\circ \leq \chi_1 \leq 180^\circ$), and *minus gauche*⁺ [-g(+)] ($-180^\circ \leq \chi_2 \leq 0^\circ$ and $0^\circ \leq \chi_1 \leq 120^\circ$).

Specifically, six states of tryptophan were considered during the computational evaluation of protein dynamics [Figure 7.1]. These six states were chosen because the sidechain of tryptophan contains two rotatable bonds, χ_1 and χ_2 , both single bonds defined by the sets of atoms N-C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$ and C $_{\alpha}$ -C $_{\beta}$ -C $_{\gamma}$ =C $_{\delta 1}$, respectively. The first bond, χ_1 , exists between the sp³ hybridized C $_{\alpha}$ and C $_{\beta}$ atoms having three minima, referred to as: *gauche*⁻ [*g*(-)], *trans* [*t*], and *gauche*⁺ [*g*(+)], at -60°, 180°, and 60° about its torsional coordinate ($\chi_1 = \theta_{\text{N-C}_{\alpha}\text{-C}_{\beta}\text{-C}_{\gamma}}$), respectively (with the abbreviated name listed in square parenthesis). The second bond, χ_2 , has a more complicated landscape as it exists between the sp³ and sp² hybridized C $_{\beta}$ and C $_{\gamma}$ atoms of the tryptophan sidechain with a molecular graph similar to that of a propene system where the three-fold sp³ center of the C $_{\beta}$ atom eclipses the π -bonded C $_{\gamma}$ -C $_{\delta 1}$ carbons of tryptophan's indole. The resulting torsional coordinate ($\theta_{\text{C}_{\alpha}\text{-C}_{\beta}\text{-C}_{\gamma}\text{-C}_{\delta 1}}$) can be simplified as having a collection of minima with the two predominant minima at approximately 60° and -60°. Thus, the set of all local minima about the χ_2 bond were named as *positive* [+] and *negative* [-] as they can be partitioned either $0^\circ \leq \textit{positive} \leq 180^\circ$ and $-180^\circ \leq \textit{negative} \leq 0^\circ$, respectively. Together, the nomenclature of these six tryptophan states follows a format describing their respective dihedral angles, with the χ_2 angle listed first and the χ_1 angle listed second. The six states are depicted in Figure 7.1 and named: *positive gauche*⁻ [+*g*(-)] ($0^\circ \leq \chi_2 \leq 180^\circ$ and $-120^\circ \leq \chi_1 \leq 0^\circ$), *positive trans* [+*t*] ($0^\circ \leq \chi_2 \leq 180^\circ$ and either $-180^\circ \leq \chi_1 \leq -120^\circ$ or $-120^\circ \leq \chi_1 \leq 180^\circ$), *positive gauche*⁺ [+*g*(+)] ($0^\circ \leq \chi_2 \leq 180^\circ$ and $0^\circ \leq \chi_1 \leq 120^\circ$), *negative gauche*⁻ [-*g*(-)] ($-180^\circ \leq \chi_2 \leq 0^\circ$ and $-120^\circ \leq \chi_1 \leq 0^\circ$), *negative trans* [-*t*]

($-180^\circ \leq \chi_2 \leq 0^\circ$ and either $-180^\circ \leq \chi_1 \leq -120^\circ$ or $-120^\circ \leq \chi_1 \leq 180^\circ$), and *negative gauche*⁺ [$-g(+)$] ($-180^\circ \leq \chi_2 \leq 0^\circ$ and $0^\circ \leq \chi_1 \leq 120^\circ$).

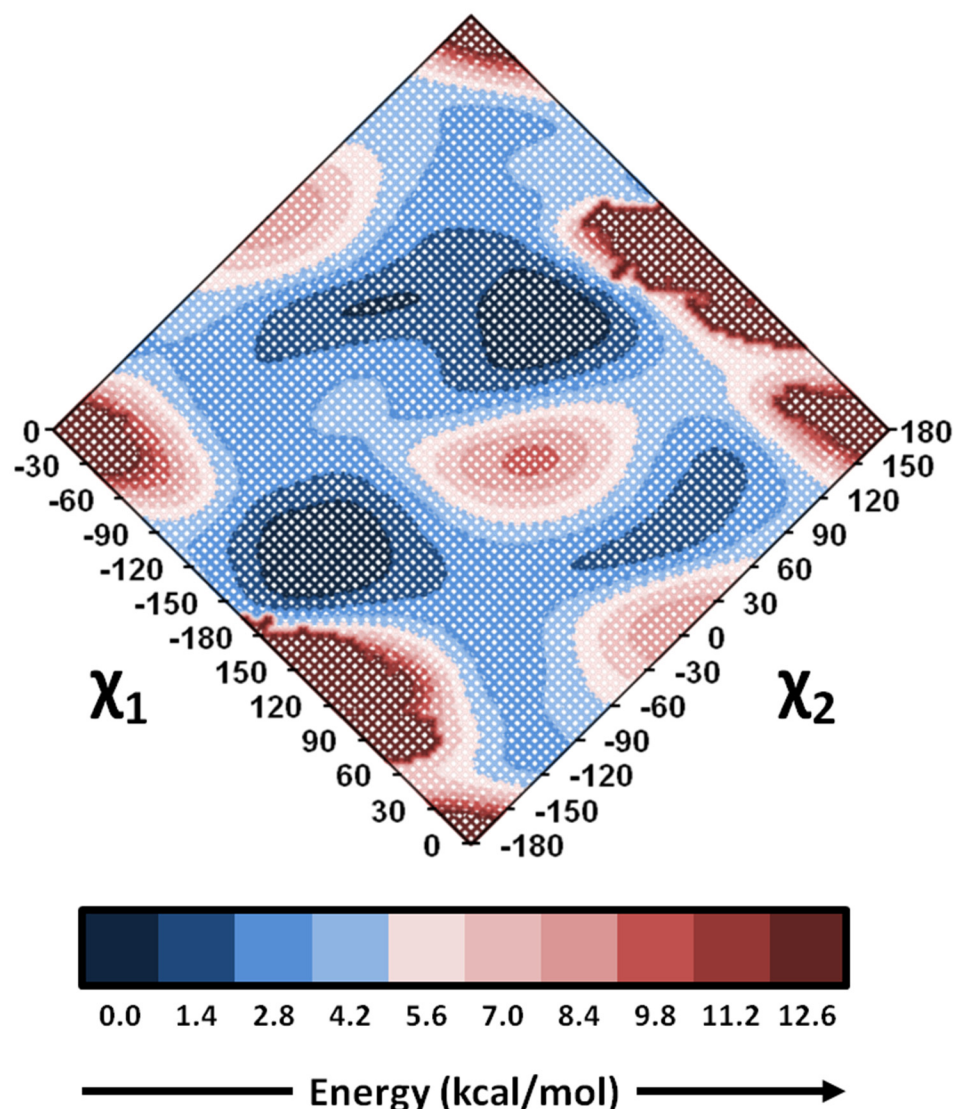


Figure 7.2. The tryptophan χ_1 and χ_2 potential energy landscape. PM6 calculation [Stewart, 2007] of gas phase potential energy ($\text{kcal}\cdot\text{mol}^{-1}$) at energy minimized geometries of tryptophan. Minimizations are initiated with fixed χ_1 and χ_2 dihedral angles searched in 5° increments. The backbone amine and carbonyl of tryptophan are capped and held fixed at dihedral angles corresponding to the crystallographic structure [Gallagher *et al.*, 1994] during calculation ($\phi = 155.335^\circ$ and $\psi = -115.236^\circ$). The energy of tryptophan conformations across the landscape are colored with respect to the global minimum energy conformation.

To facilitate the sequence dependent design of dynamics into a rigid protein scaffold, it is important that the exchange can occur between the defined states in the absence of the protein scaffold. For the purpose of the study presented here, the ideal tryptophan landscape would be degenerate; with the relative stability between states being solely dependent on the packing of protein sequences searched during calculation. Theoretical investigation of the tryptophan sidechain conformational energy landscape at the parameterized model 6 semi-empirical level [Stewart, 2007] of theory reveals a relatively flat potential energy surface permitting exchange between the defined states [Figure 7.2]. The calculations were conducted using the Gaussian software [G09, 2009] to minimize and evaluate the energy of tryptophan configurations at fixed χ_1 and χ_2 torsions in 5° increments in the gas phase in the absence of the G β 1 scaffold. Atoms belonging to the capped backbone of tryptophan were held fixed during calculation.

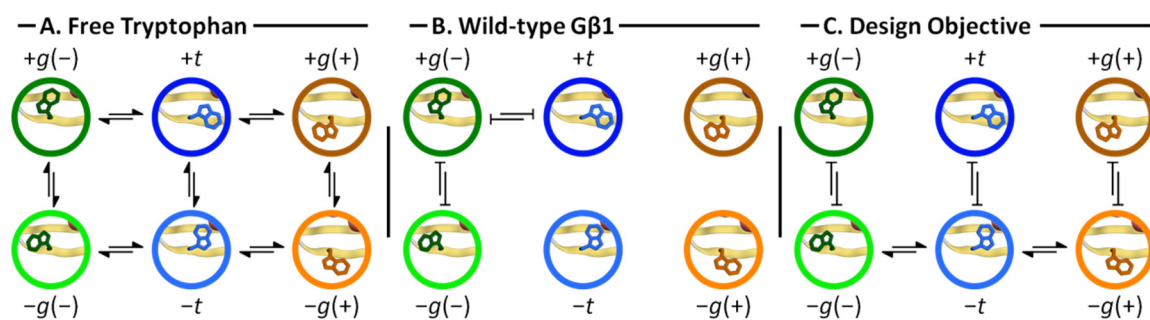


Figure 7.3. Permitted exchange of tryptophan between the defined states. Free of the G β 1 protein scaffold (A), a tryptophan residue is able to exchange between all six states. When packed within the core of the WT G β 1 protein scaffold (B) the tryptophan residue is kept fixed in a single state conformation $+g(-)$. The design objective (C) presented in this chapter involves the solution of G β 1 sequences that exhibit exchange between two tryptophan state conformations $-g(-) \leftrightarrow -g(+)$ through the $-t$ conformation.

Having defined the geometries of the states and demonstrated that the potential energy surface is permissive to exchange, the design objective can be defined [Figure 7.3]. Specifically, sequences were searched and evaluated to ensure the permissive exchange between two non-native tryptophan conformations, $-g(-)$ and $-g(+)$, through an exchange trajectory defined by the $-t$ state [Figure 7.3.C]. In this calculation framework, the $-t$ state intended to represent the collection of transitionally populated sub-states visited by the fold during interconversion of the major and minor states becomes a proxy for the transition state. The successful design of this specific reaction coordinate will produce G β 1 sequences with their tryptophan residue occupying conformations different from those previously observed in high resolution solution NMR and crystallographic structures of WT of mutant G β 1 proteins whose tryptophan solely occupies the $+g(-)$ state (average values of χ_1 and χ_2 across 14 structures measured at $-73.7 \pm 8.8^\circ$ and $75.4 \pm 10.9^\circ$, respectively) [Butterworth *et al.*, 1997; Gallagher *et al.*, 1994; Gronenborn *et al.*, 1991; Jee *et al.*, 2007; Kuszewski *et al.*, 1999; Saio *et al.*, 2009; Strop *et al.*, 2000; Tomlinson *et al.*, 2010; Wilton *et al.*, 2008; Wylie *et al.*, 2011]. The ability to design any conformational exchange at will is required to demonstrate that the method is generally applicable. By designing for tryptophan in conformations unobserved in natural G β 1 sequences we improve confidence in the *meta*-MSD methodology by attempting to eliminate any bias imparted by the WT sequence and tryptophan conformation in calculation. Including the three alternate states, $+g(+)$, $+t$, and the WT $+g(-)$ state [Figure 7.3.B], ensures that designed sequences do not adopt an unintended state or static/dynamic behaviour.

The design objective involves the creation of sequences exhibiting a dynamic exchange between conformations that have not yet been observed requiring the creation of suitable protein models *in silico*. To address the challenge of generating suitable models allowing for evaluation of the major, minor, and transition states, a step-wise procedure can be employed that includes template preparation and ensemble modelling strategies. Template preparation methods involve structural manipulation of experimentally derived atomic resolution structures, such as crystallographic structures, to produce models corresponding to the desired conformational states involved in exchange. Template generation procedures can be used to incorporate a variety of motions spanning small scale rearrangements of sidechain rotamers and backbone dihedrals to large scale changes in backbone topology. Suitable template generation methods include rotamer optimization followed by energy minimization (ROM) [Davey and Chica, 2015], homology modelling [Krieger *et al.*, 2003], and structure extension with native substructure graphs (SEWING) [Jacobs *et al.*, 2016]. To improve conformational modeling the local structure space about the set of templates can be further populated and refined by ensemble generation strategies such as the application of Backrub motions [Davis *et al.*, 2006; Smith and Kortemme, 2008] and PertMin [Davey and Chica, 2014]. Thus, the application of such a step-wise procedure should allow for the creation of relevant templates capable of reliably evaluating sequences in the context of all considered states.

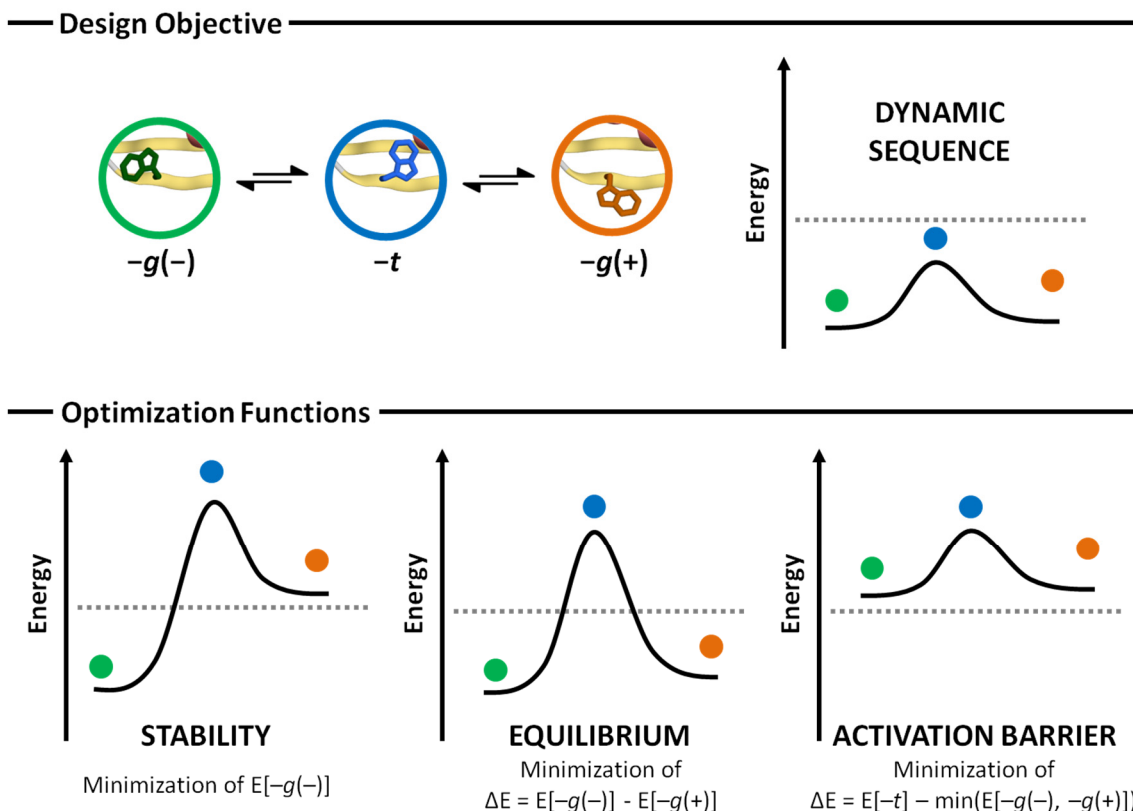


Figure 7.4. Optimization functions and the design of protein dynamics. The design of protein dynamics that exchange between three state conformations $[-g(-) \leftrightarrow -t \leftrightarrow -g(+)]$, requires that all states are stable at ambient temperature and pressure, that the major $[-g(-)]$ and minor $[-g(+)]$ state conformations are similar in energy, and that the activation barrier $[-t]$ is surmountable. Optimization of only one of these three parameters: stability, equilibrium, or the activation barrier, does not ensure that a dynamic sequence will be identified.

For a dynamic protein to be detectable by biophysical techniques, the energy difference (ΔE_{EQ}) between the major and minor states must be small in order for these end-states to be similarly populated, and the transition state barrier ($\Delta E^\ddagger$) must be sufficiently low to enable exchange on a measurable timescale, while sufficiently high to allow the protein to inhabit the two distinct major and minor states while only

transiently populating the structure space between the major and minor states. Thus, the design of proteins that can spontaneously exchange between two defined end-states requires not only the evaluation of sequences in the context of the defined states, but also the consideration of the relative stabilities of these states. This presents an additional challenge as the multivariable optimization of sequences across all defined states is complex. Not only must sequences be designed successfully to satisfy stability of the global protein fold, but their energy profiles must conform to that of a dynamic reaction coordinate. Specifically, all states must be thermally stable ($E_{\text{SEQ}} \leq E_{\text{WT}}$), the energy difference between the major and minor states must permit their coexistence as observable populations ($\Delta E_{\text{EQ}} \leq 4.2 \text{ kcal}\cdot\text{mol}^{-1}$), and the barrier to exchange must be high but surmountable to facilitate exchange on an experimentally observable timescale ($4.2 \text{ kcal}\cdot\text{mol}^{-1} \leq \Delta E^{\ddagger} \leq 16.8 \text{ kcal}\cdot\text{mol}^{-1}$).

Traditionally, CPD algorithms search sequences in the context of a single objective function – stability. However, the design of dynamics introduces an additional level of complexity as the geometry of each conformation, their absolute stability, as well as their relative stabilities must all be considered. The traditional approaches of exclusively evaluating sequence stability [Allen *et al.*, 2010], relative conformational stability [Ambroggio and Kuhlman, 2006], or transition state stability [Privett *et al.*, 2012], fail to describe the full set of optimization constraints for the solution of dynamic sequences. The exclusive optimization of either one of these three properties while ignoring the remaining two does not ensure the solution of a suitable dynamic sequence [Figure 7.4].

Multistate Design

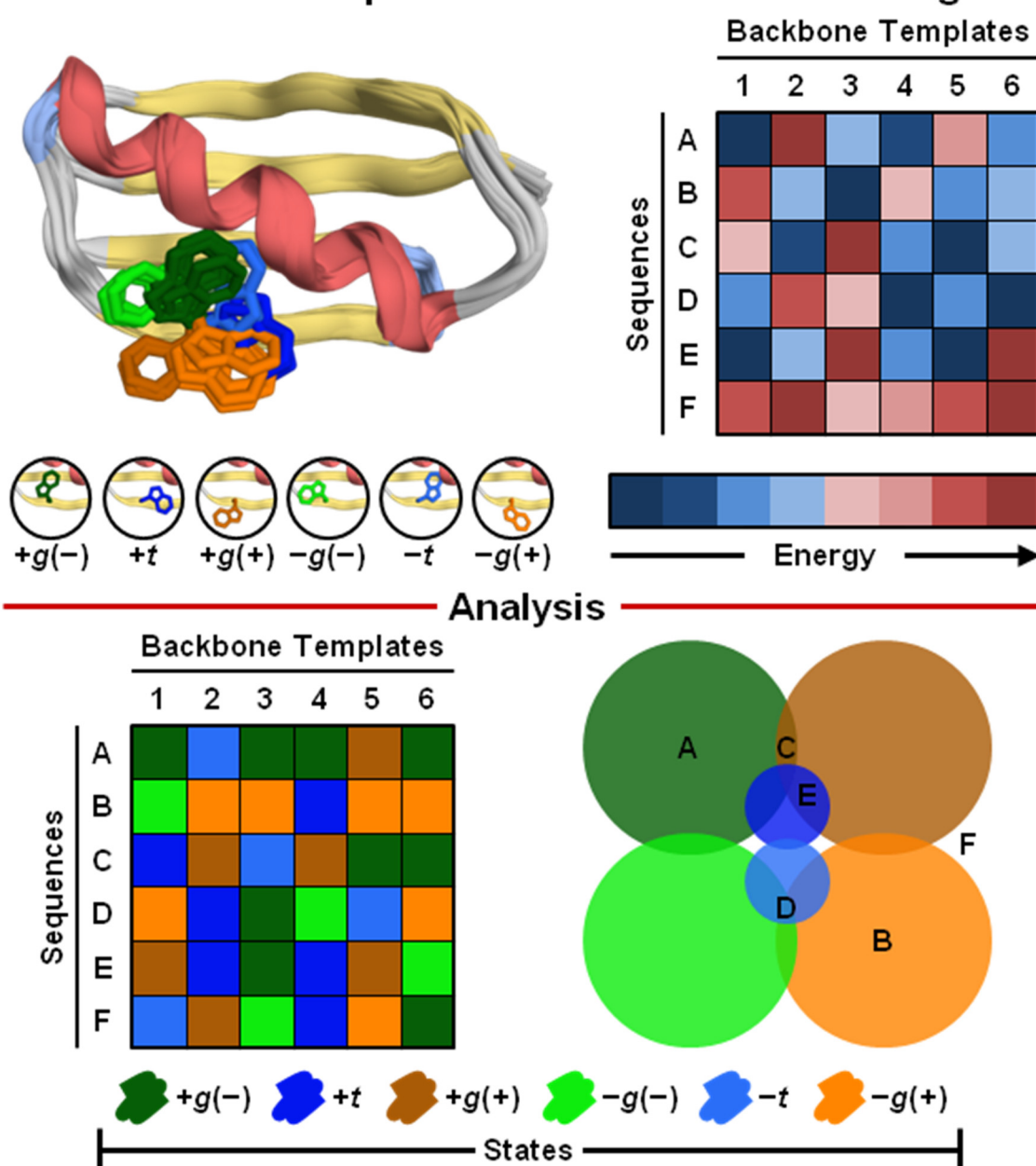


Figure 7.5. The design of protein dynamics by *meta*-multistate design. The computational framework for the design of protein dynamics is conducted in three steps: (1) backbone templates [1 – 6] are constructed to comprise a mega ensemble that allows for the evaluation of sequences in all state conformations, (2) MSD solves the set of sub-state conformations (energy and optimal rotamer configuration) across sequence space [A – F], and (3) MSA of the geometry of each sub-state is evaluated and partitioned into their respective states. The behaviour of each sequence is predicted as static or dynamic based on the energetic profile of its states. For this hypothetical

example, sequence F is predicted to be unstable in all states (unfolded), sequences A and B are predicted to adopt a single static state conformation [A: $+g(-)$ and B: $-g(+)$], sequences C and D are predicted to stabilize two state conformations but their exchange is not permitted [C: $+g(-)$ and $+g(+)$, and E: $+t$ and $+g(+)$]. Sequence D is the only sequence that can adopt three adjacent states with an energy profile corresponding to that of a dynamic sequence.

While it is possible to adapt these multistate strategies to n-dimensional MSD [Grigoryan *et al.*, 2009], a far simpler approach is to perform MSD with the goal of solving the lowest energy configuration (backbones and rotamers) of sub-states ignoring the design objective during sequence optimization. The resulting sets of sub-states are partitioned post calculation into their respective states allowing for the evaluation of the design objective to each sequence. We call this new CPD framework, involving the post-design assignment of state *meta-multistate design (meta-MSD)* [Figure 7.5].

A design procedure for creation of dynamic G β 1 sequences can be ordered into three principle steps. The first involves the preparation of a large collection of G β 1 backbone ensembles, using a step-wise template and ensemble generation strategy that will enable the quantitative evaluation of stable rotamer states for all states defined for all searched sequences. The second step, involves the solution of G β 1 sequences across the ensemble templates. This task will be executed by a series of MSD calculations across all backbone template ensembles to solve the lowest energy rotamer configuration of each sequence across each backbone template. The result yields the set of all sub-states considered in this procedure. The last step involves the prediction of protein dynamics. Specifically, the assignment of state is conducted post-MSD, involving the partitioning of all sub-states into their respective state. Sequence behaviour is then

assigned as either static or dynamic, based on the energetic profile for each sequence across the set of evaluated states.

To validate the *meta*-MSD framework, a set of candidate static and dynamic G β 1 proteins is selected for experimental characterization. Specifically, low resolution experimental characterization of the test set proteins by chemical denaturation and CD spectroscopy were carried out with the assistance of *Adam M Damry*. High resolution solution NMR experiments were designed by *Dr. Natalie K Goto* and carried out by *Adam M Damry* on the test set proteins.

7.3. Ensemble Preparation

To successfully engineer a set of dynamic G β 1 sequences that exhibit exchange of the tryptophan 43 residue via the set of $-g(-) \leftrightarrow -t \leftrightarrow -g(+)$ states, a set of backbone templates must be generated that facilitates the energetic evaluation of sequences across the set of all tryptophan states. While the evaluation of candidate dynamic sequences could hypothetically be conducted on a single template by a MSD framework where the tryptophan is oriented in a set of idealized coordinates approximating each state, it is unlikely that the local structure space would allow for the solution of a sequence that could be scored favorably in the context of all tryptophan state conformations [Davey and Chica, 2014]. Additionally, conventional ensemble generation strategies would likely prove inadequate at the same task as the structure space searched would be biased [Davey and Chica, 2015] by the identity of the WT sequence

present on the G β 1 scaffold. To circumvent these issues, a new step-wise ensemble generation strategy was developed [Figure 7.6].

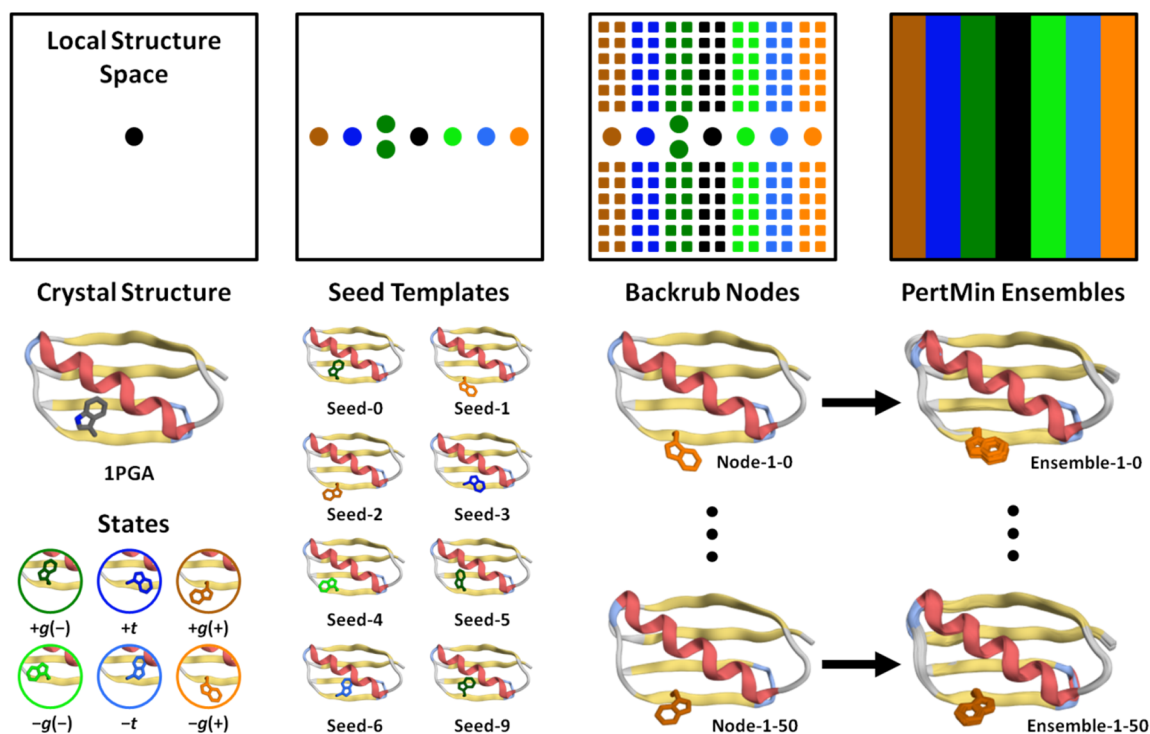


Figure 7.6. The step-wise mega ensemble generation strategy. Ensembles were constructed about the local structure space (large white square) of the crystallographic structure (1PGA) [Gallagher *et al.*, 1994] of the WT G β 1 fold (black circle). Rotamers of tryptophan corresponding to state conformations were threaded onto the fold at position 43 [Davey and Chica, 2015] to produce seed template structures (circles colored accordingly: *positive gauche*⁻ (dark green), *positive trans* (dark blue), *positive gauche*⁺ (dark orange), *negative gauche*⁻ (light green), *negative trans* (light blue), and *negative gauche*⁺ (light orange)). Backrub motions [Davis *et al.*, 2006] were applied [Lauck *et al.*, 2010] to each seed template to produce a set of node structure (squares colored according to their state conformation). PertMin ensembles [Davey and Chica, 2014] were constructed from each Backrub node. The collection of ensembles spans a total structure space of 12,648 templates.

The backbone template generation protocol illustrated in Figure 7.6 depicts the creation of a mega ensemble. The protocol is conducted across a series of three steps:

(1) seed structure generation, (2) node structure generation, and (3) ensemble generation. To ensure that the local structure space can be populated with templates that allow for the scoring of all tryptophan states, a set of seed structures were built from the 1PGA crystal structure [Gallagher *et al.*, 1994] whereby the geometry of the tryptophan 43 residue was altered. Specifically, the set of all tryptophan rotamers were threaded onto the G β 1 scaffold at position 43, ignoring all interactions with protein atoms, using the rotamer library available with the MOE program [MOE, 2012].

Table 7.1. Seed tryptophan rotamer state geometries

Seed	State Conformation	Tryptophan χ_1 Angle (°)	Tryptophan χ_2 Angle (°)
0 ^a	+g(-)	-68	74
1	-g(+)	63	-93
2	+g(+)	63	94
3	+t	-175	76
4	-g(-)	-61	-91
5	+g(-)	-61	111
6	-t	-173	-113
7 ^b	-g(-)	-67	-24
8 ^b	-t	-163	-28
9	+g(-)	-86	56

^aWT conformation found in crystallographic structure (1PGA) [Gallagher *et al.*, 1994]

^bSeed structures bearing tryptophan in conformations: $-130^\circ \leq \chi_2$ or $-50^\circ \leq \chi_2 \leq 50^\circ$ or $\chi_2 \geq 130^\circ$, are excluded from the structure set

To determine which rotamers would be used to create the seed structures, the set of all rotamers were clustered into groups less than 4 Å in heavy atom RMSD. For each

cluster, the rotamer having the lowest internal energy was kept while the remaining redundant rotamers pruned. Of the 1,723 total tryptophan rotamers available from the MOE rotamer library, 10 (including the original 1PGA crystallographic rotamer) were kept [Table 7.1]. A set of eight seeds structures encompassing the conformations of all six states were created by threading this set of seed rotamers [Table 7.1] excluding those whose χ_2 bond geometries forced the tryptophan 43 residue sidechain into the adjacent α -helix ($-130^\circ \leq \chi_2$ or $-50^\circ \leq \chi_2 \leq 50^\circ$ or $\chi_2 \geq 130^\circ$). Production of the seed structures was finalized with the addition of hydrogens using the Protonate 3D utility [Labute, 2009]. Specifically, hydrogens were added calculating solutions under a general Borne volume integral solvent model and a 12-6 Lennard-Jones potential

Having built a set of eight seed structures, an ensemble generation strategy sensitive to coordinate changes in the initial geometry of the seed structure was then applied to create a series of node structures [Howell *et al.*, 2014]. The Backrub [Davis *et al.*, 2006; Smith and Kortemme, 2011] ensemble generation protocol [Lauck *et al.*, 2010] was chosen because it can rapidly generate a population of backbone template members having high coordinate diversity while maintaining structural similarity to the original G β 1 fold [Davey and Chica, 2014]. Application of the Backrub ensemble generation procedure to create 50 backbone member templates from each of the 8 seed structures, and keeping the original set of 8 seed structures, yielded a total of 408 node structures. To finish the ensemble generation protocol, each of the 408 node structures were subjected to the perturbation followed by energy minimization (PertMin) algorithm [Davey and Chica, 2014]. The PertMin algorithm was chosen to

finish the ensemble generation procedure as it can readily produce low energy structures populating local minima about the structure space of the input structure. Specifically, random perturbations of ± 0.001 Å were applied to the cardinal Cartesian coordinates of each atom of each node structure, followed by energy minimization of each perturbed structure with the application of 50 iterations of the truncated Newton algorithm [Nash, 2000]. The algorithm was executed using the MOE program with the AMBER99 force field [Wang *et al.*, 2000] in a reaction field implicit solvent model. Prior to perturbation and energy minimization, each of the 408 node structures were explicitly solvated using the Protonate 3D utility [Labute, 2009]. The applied solvent was introduced in the form of a cube with a minimum of 6 Å of depth from the protein surface. Explicit solvent was included to force sufficient divergence of minimization trajectories [Figure 3.6]. The PertMin ensemble generation procedure was applied to generate 30 perturbed structures from each node. Minimization of the unperturbed node structures was also performed yielding ensembles of 31 backbone templates from each of the 408 nodes. A set of 12,648 total backbone templates were generated using this three step protocol with the tryptophan 43 residue occupying all six states [Table 7.2]. It is important to note that not all of the six states are populated equally throughout the total set of backbone template structures because application of the Backrub protocol involves the reconfiguration of sidechain coordinates. Together, the PertMin ensembles generated from the 408 node Backrub structures make up the mega ensemble used to search and identify potential dynamic Gβ1 proteins.

Table 7.2. Properties of the total model ensemble structure space

State State	Ensemble Members (n)	Backbone RMSD to 1PGA Crystal Structure (Å)	Tryptophan χ_1 Angle (°)	Tryptophan χ_2 Angle (°)
+g(-)	6073	0.72 ± 0.15 [0.42 : 1.55]	-73.6 ± 6.5 [-105.6 : -50.7]	83.0 ± 8.7 [16.9 : 164.9]
+t	32	0.48 ± 0.07 [0.42 : 0.86]	168.0 ± 13.5 [-122.5 : 160.0]	81.0 ± 15.1 [67.1 : 146.9]
+g(+)	1458	0.73 ± 0.13 [0.42 : 1.67]	66.6 ± 7.8 [4.6 : 117.7]	89.6 ± 11.1 [51.6 : 140.8]
-g(-)	341	0.66 ± 0.16 [0.46 : 1.28]	-86.2 ± 7.6 [-103.9 : -48.8]	-84.3 ± 11.9 [-107.7 : -53.2]
-t	94	0.68 ± 0.19 [0.40 : 0.99]	-179.8 ± 7.8 [-164.7 : 161.3]	-112.8 ± 14.2 [-139.7 : -86.1]
-g(+)	4650	0.72 ± 0.13 [0.46 : 1.77]	70.0 ± 6.8 [26.9 : 93.7]	-84.4 ± 7.0 [-123.1 : -13.3]

7.4. Sequence Space and Calculation Results

Having generated the mega ensemble of structures, the next step in the design procedure involved the evaluation of stability and the identification of the lowest energy rotamer configuration across a sequence space that could potentially satisfy the design objective. For the purposes of our design objective, a sequence space encompassing substitutions to seven core positions [Figure 7.7] known to belong to natively folded as well as stable but non-native sequences was searched [Allen *et al.*, 2010]. Specifically, the design calculations were conducted by introducing substitutions *in silico* (3: Phe and Tyr; 5: Ala, Leu, and Ile; 7: Val, Leu, Ile, and Phe; 30: Leu, Ile, and Phe; 34: Ala and Phe; 39: Val, Leu, and Ile; 54: Ala, Val, and Ile) while allowing rotamer conformational variability at positions 20, 26, 43, and 52 restricting their identity to that of the WT sequence.

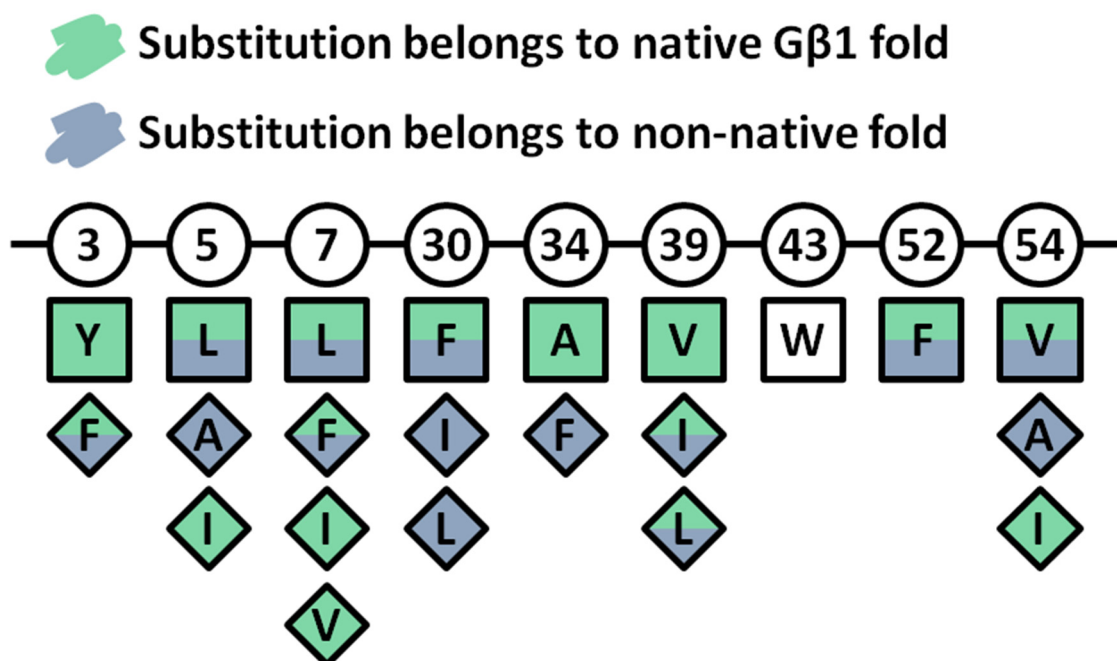


Figure 7.7. Gβ1 dynamics design sequence space. The rotamer space includes WT (square) and mutant (diagonal) residues searched at positions (circles) across the positions (circles) Gβ1 fold. Substitutions are colored according to their presence in sequences that are found to adopt the stable native Gβ1 fold (green), and those that are found in sequences that adopt non-native (blue) Gβ1 folds.

Key to this design procedure is that the conformational geometry of the tryptophan at position 43 is explicitly used to assign a state that is free to explore its rotamer conformational space to each design sub-state. In this manner, the design objective transition state of the exchange can never be sampled but only approximated by intermediate state $-t$. The sequence spaced searched totals 1,296 sequences.

All CPD calculations were performed with PHOENIX [Allen *et al.*, 2010; Chica *et al.*, 2010; Privett *et al.*, 2012] using the FASTER algorithm [Allen and Mayo, 2006; Allen and Mayo, 2010] for sequence optimization. The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2 [Dunbrack and Cohen,

1997] was used to provide side chain conformations to be threaded onto each fixed backbone template. Sequences were scored using a potential energy function consisting of a van der Waals term from the Dreiding II force field with atomic radii scale by 0.9 [Mayo *et al.*, 1990], a direction specific hydrogen-bond term having a well depth of 8.0 kcal·mol⁻¹ [Dahiyat and Mayo, 1997], and electrostatic energy term modelled using Coulomb's law with a distance dependent dielectric of 40, and a surface area-based solvation penalty term [Lazaridis and Karplus, 1999; Street and Mayo, 1998]. This scoring function does not include the backbone template energy or the internal energy of side chain conformations. Multistate sequence stabilities were calculated as a Boltzmann weighted average at 300 K of the individual energies for each backbone from the ensemble. Because the mega ensemble contains a large number of backbone templates (12,648), MSD across the mega ensemble as a whole is not computationally practical. Instead, the rotameric configuration of the 1,296 sequences was optimized by MSD-FASTER [Allen and Mayo, 2010] across smaller local structure ensembles. By restricting MSD to PertMin structures created from each of the 408 nodes generated by the Backrub protocol, the sequence space can be efficiently and tractably tackled. It is only after MSD that the set of all sub-states (16,391,808) can be partitioned into their respective states. In this way, the evaluation and design of dynamic protein sequences by *meta*-MSD is accomplished through the MSD of sub-states followed by MSA of states.

Of the 1,296 sequences only 195 were predicted to be stable having MSA Boltzmann weighted average energy less than that of the WT. Prediction of protein stability involved calculation of the Boltzmann weighted average energy across the set

of all designed sub-states for each sequence, regardless of their tryptophan conformation. Sequences having a lower energy than that of the WT were predicted to be stable [Davey and Chica, 2014]. For each of the 195 predicted stable sequences, the minimum energy sub-state for each state was retrieved [Table 7.3].

Table 7.3. State conformation energies for sequences predicted stable

Index No.	Gβ1 Sequence Mutations	State Conformation Minimum Energy (kcal·mol ⁻¹)					
		+g(-)	+t	+g(+)	-g(-)	-t	-g(+)
1	Y3F/L7I/V39I	-85.87	-67.11	-72.74	-77.73	-79.88	-75.57
2	Y3F/L5I/L7I/V39I	-85.22	-52.64	-66.02	-74.06	-76.17	-74.02
3	Y3F/L7I/V39L	-84.20	-67.07	-70.34	-76.51	-77.90	-75.07
4	Y3F/L5I/L7F/V39I	-83.44	-57.19	-68.96	-65.56	-76.25	-69.98
5	Y3F/L7I	-83.95	-64.42	-72.49	-75.65	-79.10	-74.39
6	Y3F/A34F/V39I/V54I	-69.21	-77.63	-83.47	-74.39	-67.17	-83.15
7	Y3F/V39L	-83.41	-66.99	-71.87	-75.72	-79.07	-73.39
8	Y3F/A34F/V39L	-77.77	-76.23	-75.52	-77.09	-73.00	-82.81
9	Y3F/A34F/V39I	-75.15	-75.21	-80.09	-75.73	-78.19	-82.76
10	Y3F/L7V/V39I	-83.39	-64.09	-69.57	-73.58	-76.63	-72.45
11	Y3F/V39I	-83.23	-67.89	-74.82	-75.36	-79.47	-74.69
12	Y3F/L5I/V39I	-82.61	-53.16	-66.37	-73.60	-75.34	-70.76
13	Y3F/A34F/V54I	-66.24	-75.36	-80.41	-72.65	-64.26	-82.21
14	Y3F/L7I/V39L/V54I	-82.57	-72.93	-73.79	-76.31	-81.30	-77.71
15	Y3F/L5I/L7I	-82.21	-52.66	-64.27	-71.48	-75.19	-71.68
16	Y3F/L7V/V39L	-82.25	-64.96	-66.84	-72.95	-76.24	-72.13
17	Y3F/A34F/V39L/V54I	-68.36	-77.81	-75.96	-75.15	-63.77	-81.67
18 (DANCER-1)	Y3F/L5A/L7I/A34F/V39L/V54I	-72.65	-81.21	-76.82	-78.25	-73.55	-81.92
19	Y3F/L5I/L7V/V39I	-81.65	-49.54	-62.64	-70.62	-72.04	-70.89
20	Y3F/L7V/A34F/V39I/V54I	-67.45	-77.98	-73.94	-73.61	-63.55	-81.49
21	Y3F/L5A/L7I/A34F/V39I/V54I	-73.51	-79.95	-79.95	-78.28	-72.94	-81.98
22	Y3F/L5I/L7F	-81.09	-57.36	-68.71	-62.46	-75.22	-67.64
23	Y3F/L5A/A34F/V39L/V54I	-72.46	-78.65	-77.95	-77.19	-70.16	-81.24
24	Y3F/A34F	-79.19	-74.17	-77.10	-74.20	-75.51	-81.12
25	Y3F/L5I/L7F/V39L	-80.93	-62.66	-67.87	-64.61	-69.49	-72.84
26	Y3F/L5I/V39L	-80.71	-66.72	-67.02	-74.28	-76.17	-71.29
27	Y3F	-81.04	-64.15	-71.96	-74.60	-77.84	-72.39
28 (DANCER-0)	Y3F/L5A/L7I/A34F/V39I	-74.96	-77.19	-78.97	-80.90	-73.12	-80.04
29	Y3F/L7F/V39I	-80.65	-68.94	-72.93	-72.02	-69.01	-73.73
30	Y3F/L7V	-80.77	-61.37	-69.12	-71.80	-75.86	-70.79
31 (K)	Y3F/L7I/A34F/V39I/V54I	-70.26	-78.96	-76.99	-76.18	-65.77	-80.45
32	Y3F/L5I	-80.32	-51.52	-64.70	-71.03	-74.39	-68.77
33	Y3F/L7I/A34F/V39I	-73.55	-77.50	-76.09	-77.86	-71.40	-80.31
34	Y3F/V39L/V54I	-80.50	-72.10	-73.21	-75.46	-80.76	-74.30
35	Y3F/L7V/A34F/V39I	-67.32	-75.05	-73.02	-79.80	-68.94	-79.63
36	Y3F/L5I/L7F/V39L/V54I	-79.70	-73.39	-70.29	-66.18	-77.41	-75.80
37	Y3F/L5A/A34F/V39L	-73.35	-75.63	-76.52	-79.83	-71.13	-76.83
38 (DANCER-2)	Y3F/L5A/L7I/A34F/V39L	-75.14	-75.59	-75.60	-80.20	-70.82	-78.92
39	Y3F/V39I/V54I	-80.31	-72.65	-77.21	-74.96	-80.16	-77.55
40	Y3F/L5I/L7I/V39L	-79.77	-53.50	-65.12	-74.86	-71.09	-72.19
41	Y3F/L7I/A34F/V39L	-73.08	-74.33	-74.12	-77.38	-67.79	-79.77
42	Y3F/L7I/V39I/V54I	-79.90	-74.20	-75.01	-76.08	-80.32	-77.76
43 (DANCER-3)	Y3F/L7I/A34F/V39L/V54I	-67.19	-77.67	-76.41	-76.83	-66.03	-79.71
44	Y3F/L5I/L7I/V39L/V54I	-78.73	-64.55	-72.79	-72.47	-79.34	-73.59
45	Y3F/L5A/L7V/A34F/V39L/V54I	-69.22	-77.64	-73.17	-74.67	-67.00	-79.31
46	Y3F/L5A/L7I/A34F/V54I	-70.76	-77.65	-79.54	-76.58	-72.62	-79.61
47	Y3F/L5A/A34F/V39I/V54I	-73.47	-78.83	-78.94	-77.14	-70.85	-79.65
48	Y3F/L5A/L7V/A34F/V39I/V54I	-71.96	-76.39	-76.43	-74.77	-66.54	-79.31

49	Y3F/L7V/V39L/V54I	-79.29	-69.99	-70.38	-72.68	-77.68	-74.74
50	L5I/L7I/V39I	-78.73	-46.91	-60.29	-67.13	-69.47	-67.37
51	Y3F/L7V/A34F/V54I	-65.60	-75.96	-72.37	-71.27	-61.23	-78.66
52	Y3F/L5A/L7F/A34F/V39I/V54I	-75.07	-73.66	-75.51	-72.47	-70.57	-78.87
53	Y3F/L5I/L7I/V39I/V54I	-75.90	-63.15	-72.42	-72.93	-78.75	-76.89
54	Y3F/L5I/L7V	-78.67	-49.05	-61.13	-68.05	-71.09	-68.12
55	Y3F/V54I	-77.62	-70.16	-74.33	-72.24	-79.23	-74.70
56	L7I/V39I	-79.07	-59.35	-66.84	-71.19	-73.02	-68.39
57	Y3F/L5A/A34F/V39I	-74.33	-76.26	-77.98	-79.06	-72.80	-78.64
58	Y3F/L7I/V54I	-77.66	-71.92	-74.64	-73.24	-78.99	-76.03
59	Y3F/L7V/F30I/A34F/V39I/V54I	-61.21	-75.13	-70.35	-70.53	-48.92	-78.16
60	Y3F/L7I/A34F/V54I	-68.10	-78.41	-75.03	-74.72	-63.44	-78.15
61	Y3F/L7F/A34F/V39I	-69.70	-70.03	-74.09	-74.49	-71.61	-78.00
62	Y3F/L5I/V39I/V54I	-74.94	-71.53	-72.17	-72.21	-78.22	-75.70
63	Y3F/L5A/L7I/A34F	-74.21	-74.11	-77.19	-78.40	-71.53	-78.52
64	Y3F/L5I/L7I/A34F/V39I/V54I	-69.18	-69.91	-77.90	-63.27	-60.64	-75.61
65	Y3F/L5I/L7F/A34F/V39I	-69.00	-65.11	-59.96	-70.28	-45.40	-77.80
66	Y3F/L5I/L7F/A34F/V39I/V54I	-61.48	-66.34	-60.22	-58.88	-24.99	-77.62
67	Y3F/L5I/V39L/V54I	-76.94	-65.00	-72.43	-71.66	-77.76	-72.32
68	Y3F/L7F/A34F/V39L/V54I	-66.01	-72.00	-74.85	-74.22	-54.55	-77.61
69	Y3F/L7F/A34F/V39L	-72.03	-69.75	-75.23	-73.10	-64.78	-77.43
70	Y3F/L5I/L7I/A34F/V39L/V54I	-61.37	-69.89	-77.57	-64.43	-67.11	-75.23
71	Y3F/L7F	-77.98	-61.15	-70.93	-71.11	-71.89	-70.71
72	Y3F/L7I/A34F	-73.66	-76.71	-74.30	-75.73	-69.77	-77.64
73	Y3F/L5I/L7I/F30L/V39I	-77.56	-57.47	-59.82	-70.73	-71.13	-64.87
74	Y3F/L7F/V39L	-78.03	-63.25	-72.96	-71.31	-68.70	-69.26
75	Y3F/L5I/L7F/A34F/V39L/V54I	-63.33	-64.87	-63.89	-57.74	-36.08	-77.11
76	Y3F/L7I/F30I/A34F/V39I	-64.25	-70.02	-71.81	-77.23	-59.56	-74.14
77	L7I/V39L	-77.61	-60.41	-63.55	-69.97	-71.00	-68.02
78	Y3F/L5A/A34F/V54I	-71.08	-76.66	-77.51	-75.48	-68.69	-78.07
79	Y3F/L5I/L7I/V54I	-75.66	-62.44	-69.85	-70.52	-77.36	-74.61
80	Y3F/L5I/L7I/A34F/V54I	-67.39	-67.67	-77.01	-61.85	-58.63	-72.81
81	Y3F/L7V/A34F/V39L/V54I	-66.94	-73.92	-74.06	-73.28	-63.76	-77.35
82	Y3F/L7V/A34F	-69.68	-74.25	-71.62	-77.27	-66.88	-76.65
83	Y3F/L7F/V39I/V54I	-77.41	-68.22	-75.34	-70.81	-73.15	-75.81
84	Y3F/L7V/A34F/V39L	-69.69	-70.58	-70.83	-74.78	-64.90	-77.28
85	Y3F/L5I/L7F/A34F/V39L	-68.45	-64.12	-66.15	-68.60	-67.14	-76.82
86	Y3F/A34F/V39I/V54A	-74.76	-67.48	-71.38	-68.01	-70.68	-76.89
87	Y3F/A34F/V39L/V54A	-76.55	-68.30	-72.12	-70.20	-71.35	-77.10
88	Y3F/L5I/L7F/V39I/V54I	-74.01	-72.95	-71.31	-67.64	-77.01	-74.89
89	Y3F/L5A/L7F/V39L	-77.10	-57.86	-62.75	-67.68	-69.59	-65.28
90	Y3F/L5A/L7V/A34F/V39I	-70.97	-73.53	-75.48	-77.45	-69.76	-77.07
91	Y3F/L5I/L7F/V39I/V54A	-76.97	-41.09	-58.50	-62.48	-66.77	-63.53
92	Y3F/L5I/L7F/V39L/V54A	-76.73	-41.88	-61.52	-62.02	-60.43	-63.46
93	Y3F/L5I/L7F/V54I	-73.45	-70.62	-71.38	-65.20	-76.74	-72.30
94	Y3F/L5A/L7F/A34F/V39L/V54I	-73.98	-74.44	-75.76	-74.72	-65.83	-77.05
95	Y3F/L5A/L7F/A34F/V54I	-72.00	-71.93	-73.76	-69.78	-65.10	-76.81
96	Y3F/L5I/L7F/F30L/V39I	-76.48	-54.91	-60.38	-61.76	-71.23	-63.13
97	Y3F/L5I/L7I/V39I/V54A	-76.65	-44.27	-54.67	-66.02	-66.84	-67.60
98	Y3F/L5A/A34F	-72.15	-73.14	-75.98	-77.12	-70.26	-77.18
99	Y3F/L7F/V54I	-76.44	-66.67	-73.40	-69.02	-72.95	-72.49
100	L7I	-76.98	-57.08	-65.66	-69.11	-72.20	-67.52
101	Y3F/L5A/L7V/A34F/V39L	-71.15	-71.19	-72.16	-75.92	-68.06	-76.93
102	Y3F/L7F/V39L/V54I	-76.77	-66.35	-74.85	-70.00	-75.36	-72.17
103	Y3F/L7V/F30I/A34F/V39I	-61.06	-67.52	-68.72	-74.21	-56.78	-76.28
104	A34F/V39I/V54I	-62.05	-71.16	-76.22	-68.40	-61.12	-76.45
105	V39L	-76.55	-61.47	-65.18	-68.67	-72.99	-66.83
106	Y3F/L7I/F30I/A34F/V39L/V54I	-59.76	-70.84	-71.97	-74.55	-53.77	-76.13
107	Y3F/L5I/A34F/V39I/V54I	-64.47	-68.88	-76.21	-58.77	-59.11	-75.67
108	Y3F/L7I/V39I/V54A	-76.74	-59.10	-64.69	-70.92	-71.26	-67.68
109	Y3F/L5I/L7I/A34F/V39I	-71.49	-65.99	-76.21	-71.69	-68.18	-75.81
110	Y3F/L5I/L7V/V39L	-76.42	-51.97	-61.75	-71.43	-67.56	-70.53
111	Y3F/F30I/A34F/V39L/V54I	-62.68	-71.45	-75.23	-76.10	-51.65	-72.91
112	Y3F/F30L/V39L	-75.98	-58.15	-59.82	-66.19	-69.17	-60.56
113	Y3F/L5I/L7V/V39L/V54I	-74.83	-61.95	-69.49	-69.54	-76.07	-73.19
114	Y3F/L7V/F30L/A34F/V39I	-65.22	-66.47	-64.99	-75.85	-62.11	-72.54

115	Y3F/L5I/V54I	-74.61	-63.71	-69.61	-69.82	-76.22	-73.72
116	L5I/L7F/V39I	-76.02	-50.51	-62.08	-58.96	-69.55	-62.90
117	Y3F/L5A/L7F/A34F/V39I	-76.14	-70.80	-74.12	-74.73	-72.60	-76.33
118	Y3F/L5A/L7V/A34F/V54I	-70.69	-74.10	-76.43	-73.50	-64.06	-76.55
119	Y3F/L5I/A34F/V39I	-71.35	-65.85	-74.38	-70.66	-66.72	-75.87
120	Y3F/F30I/V39L	-75.99	-62.70	-63.92	-65.19	-56.62	-64.02
121	Y3F/L7V/V39I/V54I	-76.28	-70.66	-71.93	-73.07	-76.82	-74.63
122	Y3F/L5I/F30L/V39I	-75.79	-55.77	-59.55	-64.14	-70.41	-61.55
123	Y3F/F30I/V39I	-76.00	-61.88	-63.55	-67.36	-57.74	-64.24
124	Y3F/L5I/A34F/V39L/V54I	-59.48	-69.83	-76.02	-62.56	-63.50	-75.15
125	Y3F/L7F/A34F/V39I/V54I	-66.94	-71.55	-74.31	-68.71	-56.66	-75.97
126	Y3F/L7I/F30I/A34F/V39L	-64.22	-69.41	-70.37	-75.65	-62.82	-73.07
127	Y3F/L5I/L7V/A34F/V39I/V54I	-55.73	-65.14	-75.58	-60.44	-64.25	-72.02
128	Y3F/L7F/F30I/V39I	-75.57	-53.62	-62.94	-63.79	-52.77	-63.20
129	L7V/V39I	-76.43	-56.32	-63.78	-67.04	-70.31	-65.07
130	L5I/L7I	-75.73	-45.99	-57.51	-64.55	-68.49	-65.03
131	Y3F/L5I/L7V/A34F/V39L/V54I	-56.41	-66.25	-75.61	-61.12	-64.67	-71.54
132	V39I	-76.24	-60.83	-68.23	-68.87	-73.39	-68.43
133	A34F/V54I	-59.46	-68.61	-73.16	-66.99	-58.21	-75.50
134	Y3F/L7V/F30I/A34F/V54I	-58.93	-73.14	-71.11	-68.19	-47.97	-75.33
135	Y3F/L7I/F30I/A34F/V39I/V54I	-62.05	-74.97	-73.11	-74.08	-51.99	-75.78
136	Y3F/L7V/V54I	-74.37	-68.39	-71.30	-70.29	-76.10	-72.37
137	Y3F/L5I/L7I/F30I/V39I	-75.47	-56.39	-59.68	-71.28	-43.86	-67.22
138	A34F/V39L	-71.87	-69.57	-69.78	-71.19	-67.37	-75.37
139	Y3F/L7F/A34F	-67.90	-68.97	-73.54	-72.39	-72.42	-75.40
140	Y3F/L7I/F30I/V39I	-75.42	-62.12	-66.06	-70.22	-56.82	-68.14
141	L5I/L7V/V39I	-75.17	-44.30	-56.91	-63.69	-65.34	-63.64
142	Y3F/L5A/L7F/A34F/V39L	-75.66	-72.24	-72.77	-74.17	-72.72	-74.65
143	L7I/A34F/V39I	-66.86	-70.83	-70.35	-70.68	-65.35	-75.09
144	Y3F/F30I/A34F/V39I	-67.13	-69.08	-72.04	-75.26	-58.00	-73.50
145	Y3F/L5I/L7V/V39I/V54I	-72.43	-63.16	-69.05	-69.81	-75.31	-74.27
146	L5I/V39I	-75.25	-47.20	-60.00	-66.67	-68.64	-64.11
147	Y3F/F30L/A34F/V39L/V54I	-63.59	-68.77	-71.06	-75.10	-65.00	-72.64
148	Y3F/V39I/V54A	-75.76	-57.24	-67.52	-69.11	-70.66	-68.37
149	Y3F/L5I/L7I/A34F/V39L	-70.73	-66.74	-75.00	-68.66	-69.19	-74.96
150	L7I/A34F/V39I/V54I	-64.22	-72.30	-71.26	-70.83	-59.05	-75.23
151	A34F/V39I	-69.04	-68.53	-72.84	-69.33	-72.13	-75.31
152	A34F/V39L/V54I	-61.89	-70.84	-69.14	-69.25	-58.13	-74.96
153	Y3F/L5I/A34F/V39L	-71.33	-66.57	-72.07	-69.81	-67.70	-74.88
154	Y3F/L5A/L7F/V39I	-75.48	-58.67	-65.36	-67.82	-71.44	-66.97
155	Y3F/L7I/V39L/V54A	-75.59	-61.73	-64.99	-68.50	-70.94	-67.30
156	Y3F/L5I/V39I/V54A	-74.98	-44.11	-55.53	-65.36	-66.11	-66.09
157	L7I/V39L/V54I	-75.43	-66.29	-66.80	-69.77	-74.51	-70.66
158	Y3F/L5I/L7I/F30L	-74.89	-54.87	-57.17	-68.29	-70.17	-62.53
159	L7V/V39L	-75.21	-56.92	-60.33	-66.21	-70.16	-64.75
160	Y3F/L7I/F30I/A34F	-61.84	-69.59	-73.17	-74.81	-59.45	-71.21
161	Y3F/L7V/F30L/A34F/V39I/V54I	-57.91	-74.66	-64.61	-69.81	-53.56	-74.42
162	Y3F/L5I/L7F/A34F	-66.13	-62.83	-66.57	-68.96	-54.17	-74.50
163	Y3F/L5I/L7F/A34F/V54I	-59.02	-64.04	-69.92	-58.04	-43.53	-74.50
164	L5A/L7I/A34F/V39L/V54I	-65.45	-73.72	-69.74	-71.70	-66.64	-74.87
165	Y3F/F30I/A34F/V39I/V54I	-62.81	-74.62	-72.98	-73.04	-56.99	-74.81
166	Y3F/L5I/A34F/V54I	-62.69	-66.86	-74.51	-60.82	-57.12	-72.57
167	Y3F/F30L/V39I	-74.98	-56.21	-61.04	-68.70	-68.80	-63.11
168	Y3F/L7V/F30I/A34F/V39L/V54I	-58.85	-70.71	-69.27	-70.55	-50.71	-74.64
169	Y3F/L5A/L7V/A34F	-69.57	-70.47	-73.70	-75.23	-67.98	-75.02
170	L5A/A34F/V39L/V54I	-66.33	-71.15	-71.37	-70.24	-63.57	-74.56
171	Y3F/F30I/A34F/V39L	-67.01	-68.06	-74.60	-73.67	-60.58	-72.11
172	Y3F/L5I/F30L/V39L	-74.41	-59.68	-60.36	-61.37	-71.16	-61.00
173	Y3F/L5I/L7F/F30I/V39I	-74.19	-55.05	-61.41	-62.51	-48.07	-63.44
174	Y3F/L5I/L7F/F30L	-74.13	-53.10	-59.89	-61.45	-70.20	-57.98
175	L5I/L7F/V39L	-74.41	-56.60	-62.00	-59.57	-62.80	-65.65
176	Y3F/L7I/V54A	-74.86	-62.34	-65.09	-67.94	-70.53	-66.57
177	Y3F/V39L/V54A	-75.24	-61.60	-65.07	-67.61	-70.92	-65.24
178	Y3F/L5I/L7V/A34F/V54I	-53.73	-63.24	-74.06	-58.25	-62.47	-69.45
179	Y3F/L7I/F30I/A34F/V54I	-59.79	-74.37	-74.02	-71.73	-50.90	-74.04
180	Y3F/L5A/L7I/A34F/V39I/V54A	-70.26	-64.78	-70.13	-74.42	-65.78	-71.44

181	Y3F/F30L/A34F/V39I/V54I	-64.45	-72.54	-67.22	-74.16	-65.64	-71.42
182	Y3F/L5I/L7F/F30L/V39L	-73.97	-55.79	-60.78	-61.91	-64.47	-66.34
183	L7F/V39I	-74.40	-56.64	-66.35	-65.70	-62.76	-67.47
184	Y3F/L5I/L7F/V54A	-74.18	-40.27	-59.94	-59.51	-65.82	-60.43
185	Y3F/L5A/L7F/F30L/A34F/V39I/V54I	-66.14	-64.64	-70.11	-65.64	-64.29	-74.05
186	Y3F/L7V/F30L/V39I	-74.07	-54.62	-58.52	-61.52	-65.56	-59.17
187	Y3F/L5I/V39L/V54A	-74.39	-45.53	-59.25	-65.64	-68.38	-64.52
188	Y3F/F30L/A34F/V39L	-64.46	-64.92	-70.31	-73.90	-62.87	-69.47
189	L5A/L7I/A34F/V39I/V54I	-66.72	-72.45	-73.37	-71.74	-65.64	-74.60
190	Y3F/L5I/L7I/F30L/A34F/V39L/V54I	-59.90	-66.64	-71.63	-66.03	-55.53	-74.04
191	A34F	-73.28	-67.50	-69.85	-68.30	-69.46	-74.06
192	Y3F/L7V/V39L/V54A	-74.39	-60.90	-61.89	-65.36	-68.58	-63.88
193	Y3F/L7I/F30L/A34F/V39L/V54I	-60.43	-67.74	-69.48	-72.29	-55.50	-73.95
194	Y3F/F30I/A34F/V54I	-60.38	-74.05	-73.77	-70.72	-55.83	-72.47
195	Y3F/L5I/L7V/F30L/V39I	-73.99	-53.74	-56.43	-62.01	-67.04	-61.15
WT		-74.75	-59.52	-65.36	-66.14	-71.15	-68.55
196 (H)	Y3F/L7I/F30L/V39I	-73.72	-58.42	-60.81	-61.51	-68.68	-65.64

The energetic profile of a sequence's set of states was used to predict whether the sequence was static or dynamic. The list of state stabilities tabulated in Table 7.3 was parsed to identify those that exhibit an energetic profile corresponding to the design objective. For a sequence to be predicted dynamic it must satisfy four design criteria. Consider the design of the objective dynamic exchange between $-g(+) \leftrightarrow -t \leftrightarrow -g(-)$, representing the major, transition, and minor states, respectively. First, the sequence must be stable ($E_{-g(+)} \leq E_{WT}$, kcal·mol⁻¹), second, the second lowest energy state must belong to the minor state ($E_{-g(-)} \leq E_{-g(+)} + 4.2$, kcal·mol⁻¹), third, the intermediate state serving as an energetic proxy of the transition state must inhabit an energetic window permitting the dynamic exchange to occur but preventing the state from being stably populated ($E_{-g(+)} + 4.2 \leq E_{-t} \leq E_{-g(+)} + 16.8$, kcal·mol⁻¹). Of the 195 sequences predicted to be stable (satisfying the first condition), only 36 sequences exhibit an energetic profile conforming to the design objective, where $-g(+)$ and $-g(-)$ serve as either of the major and minor states and $-t$ serves as the transition state proxy.

To investigate the predictive capabilities of the *meta*-MSD framework, a set of candidate G β 1 proteins was selected [Table 7.4]. This set of candidate mutant sequences, includes two sequences predicted to be static, H and K, with the tryptophan occupying the +*g*(−) and −*g*(+) conformations, respectively, and four randomly selected dynamic candidate sequences having an array of predicted energetic barriers (DANCER-0, DANCER-1, DANCER-2, and DANCER-3). The four dynamic candidates are named DANCER because they are Dynamic Native Conformation ExchangeR proteins. Of the DANCER proteins, the major species of DANCER-0 and DANCER-2 are predicted to occupy the −*g*(+) tryptophan conformation, while the major species of DANCER-1 and DANCER-3 are predicted to occupy the −*g*(−) tryptophan conformation. The test set sequences have between four to six mutations, approximately 7% to 11% of the WT sequence.

Table 7.4. G β 1 sequences predicted static and dynamic by *meta*-MSD

Sequence ID	Mutations From WT	<i>meta</i> -MSD Predictions		
		Behaviour	ΔE_{EQ} (kcal·mol ^{−1})	$\Delta E^\ddagger$ (kcal·mol ^{−1})
WT		+ <i>g</i> (−)	8.6	15.2
H	Y3F/L7I/F30L/V39I	+ <i>g</i> (−)	12.2	15.3
K	Y3F/L7I/A34F/V39I/V54I	− <i>g</i> (+)	4.3	14.7
DANCER-0	Y3F/L5A/L7I/A34F/V39I	− <i>g</i> (+) $\leftrightarrow$ − <i>g</i> (−)	0.9	7.8
DANCER-1	Y3F/L5A/L7I/A34F/V39L/V54I	− <i>g</i> (−) $\leftrightarrow$ − <i>g</i> (+)	3.7	8.4
DANCER-2	Y3F/L5A/L7I/A34F/V39L	− <i>g</i> (+) $\leftrightarrow$ − <i>g</i> (−)	1.3	9.4
DANCER-3	Y3F/L7I/A34F/V39L/V54I	− <i>g</i> (−) $\leftrightarrow$ − <i>g</i> (+)	2.9	13.7

The predicted dynamic exchange between state conformations of DANCER (Dynamic And Native Conformational ExchangeR) proteins are indicated: major species $\leftrightarrow$ minor species

It is also important to note that, although not experimentally characterized in this chapter, there are alternate sequence behaviours and conformations that may be of interest for future study. Specifically, four of the predicted stable sequences are predicted by *meta*-MSD to fold as static structures with the tryptophan 43 residue adopting the $-g(-)$ conformation. These four sequences, identified by their index numbers: 76, 111, 144, and 188 [Table 7.3], have not been previously characterized or reported in the literature. Sequences 76, 111, and 144, are all predicted to have large activation energies ($17.26 \text{ kcal}\cdot\text{mol}^{-1} \leq E^\ddagger \leq 24.45 \text{ kcal}\cdot\text{mol}^{-1}$) prohibitive to conformational exchange, but with the alternative $-g(+)$ tryptophan conformational state evaluated to be similar in energy to the major state ($1.76 \text{ kcal}\cdot\text{mol}^{-1} \leq E_{\text{EQ}} \leq 3.19 \text{ kcal}\cdot\text{mol}^{-1}$). The last sequence predicted static adopting the $-g(-)$ conformation, index number 188 [Table 7.3], has an energy barrier predicted permissive to exchange ($E^\ddagger = 11.03 \text{ kcal}\cdot\text{mol}^{-1}$) but a difference in energy between $-g(-)$ and $-g(+)$ conformations favoring the $-g(-)$ state ($E_{\text{EQ}} = 4.43 \text{ kcal}\cdot\text{mol}^{-1}$). While these predictions promise that there are potentially four static G β 1 variants adopting the $-g(-)$ state, difficulties in evaluating the relative stabilities of the tryptophan in the core $-g(-)$ and solvent partially exposed $-g(+)$ states may lead to inaccuracies resulting in sequences that actually adopt the $-g(+)$ state. A second set of dynamic sequences that exchange via the $+g(-) \leftrightarrow +t \leftrightarrow +g(+)$ trajectory can also be identified by *meta*-MSD, although the transition is satisfied by far fewer sequences (5 sequences total) than those that satisfy the design objective presented in this chapter ($-g(-) \leftrightarrow -t \leftrightarrow -g(+)$, 36 sequences total). Specifically, sequences identified by the index numbers: 39, 83, 99, 102, and 142

[Table 7.3], are all predicted to be dynamic. The stabilities of the first four of these sequences have been previously characterized by Allen and colleagues [Allen *et al.*, 2010] and are predicted by *meta*-MSD to have activation energies ($7.66 \text{ kcal}\cdot\text{mol}^{-1} \leq E^\ddagger \leq 10.42 \text{ kcal}\cdot\text{mol}^{-1}$) and energy differences between major and minor state conformations ($1.91 \text{ kcal}\cdot\text{mol}^{-1} \leq E_{\text{EQ}} \leq 3.10 \text{ kcal}\cdot\text{mol}^{-1}$) that should be permissive to exchange. The last sequence, identified as index number 148 [Table 7.3], is presumed to adopt a non-native G β 1 conformation or behaviour as it has also been characterized by Allen *et al.* and shown not to unfold in response to the addition of chemical denaturant. Interestingly, *meta*-MSD predicts very low values for both the activation barrier ($E^\ddagger = 3.42 \text{ kcal}\cdot\text{mol}^{-1}$) and difference in energy between the major and minor states ($E_{\text{EQ}} = 2.89 \text{ kcal}\cdot\text{mol}^{-1}$). This prediction, corroborated with experimental evidence, could be interpreted to suggest that the core of the protein is highly dynamic and does not adopt a single folded state. Instead, the sequence may adopt a molten globular G β 1 fold.

7.5. Experimental Characterization of DANCER Proteins

To validate predictions made using the *meta*-MSD framework, a series of experiments were performed on the test set of sequences [Table 7.4]. These experiments required the cloning, expression, and purification of test set sequences. Specifically, His-tagged G β 1 sequences were synthesized by DNA2.0 and cloned into the pJ414 plasmid (DNA2.0). These constructs were transformed into chemically competent *E. coli* BL21-Gold (DE3) cells (Agilent) and complete coding regions were verified by DNA sequencing.

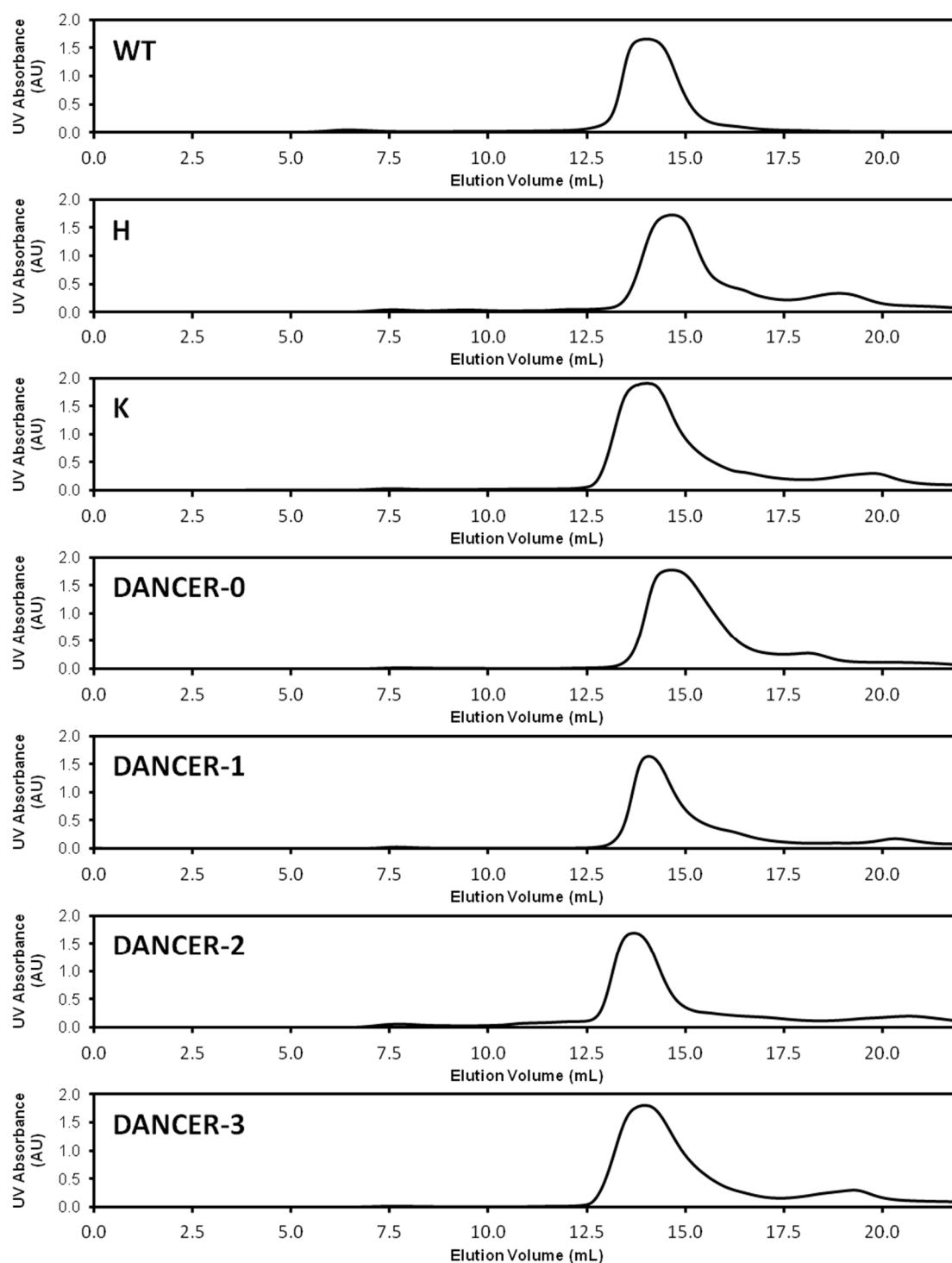


Figure 7.8. Size exclusion chromatograms of test set proteins. Study of chromatograms reveals that all test set G β 1 proteins elute with the same retention volume as the WT protein. A second species is eluted at a later fraction indicating potential degraded protein present. Protein absorbance is measured at 280 nm. Figure provided courtesy of Adam M Damry.

Proteins for size exclusion chromatography, chemical denaturation assays, and thermal denaturation monitored by CD measurements were expressed using LB broth containing 100 µg/mL ampicillin. Cultures were grown to an optical density of roughly 0.6 ($\lambda = 600$ nm) in an incubator at 37 °C with shaking, followed by the induction of expression with 1 mM IPTG. Induced cells were then incubated overnight at either 15 °C or 37 °C with shaking. Following incubation, cells were harvested by centrifugation and lysed using an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography according to the manufacturer's protocol, and then gel filtered using an ENrich SEC 650 size exclusion chromatography column (BioRad), using 10 mM sodium phosphate buffer at pH 7.4 as eluent. Purified samples were then concentrated using Amicon Ultracel-3K centrifugal filter units (EMD Millipore).

To determine whether the test set of sequences formed stable monomeric folds at room temperature, we designed and analyzed size exclusion chromatography experiments as well as chemical and thermal denaturation experiments which were conducted by *Adam M Damry*. Despite sizeable mutation to the sequence space of G β 1, the proteins express as soluble monomers as confirmed by size exclusion chromatography [Figure 7.8] and adopt a stable protein fold at room temperature as observed by thermal denaturation with CD spectroscopy [Figure 7.9]. It is important to note that data from size exclusion chromatography indicates that G β 1 sequence variants H and K, as well as the set of DANCER proteins, all exhibit chromatograms with

protein containing fractions eluted at retention volumes greater (> 15 mL) than those of the WT. Further characterization (not shown) by size exclusion chromatography of samples over a course of days demonstrated that these eluted fractions likely contained G β 1 fragments resulting from proteolysis.

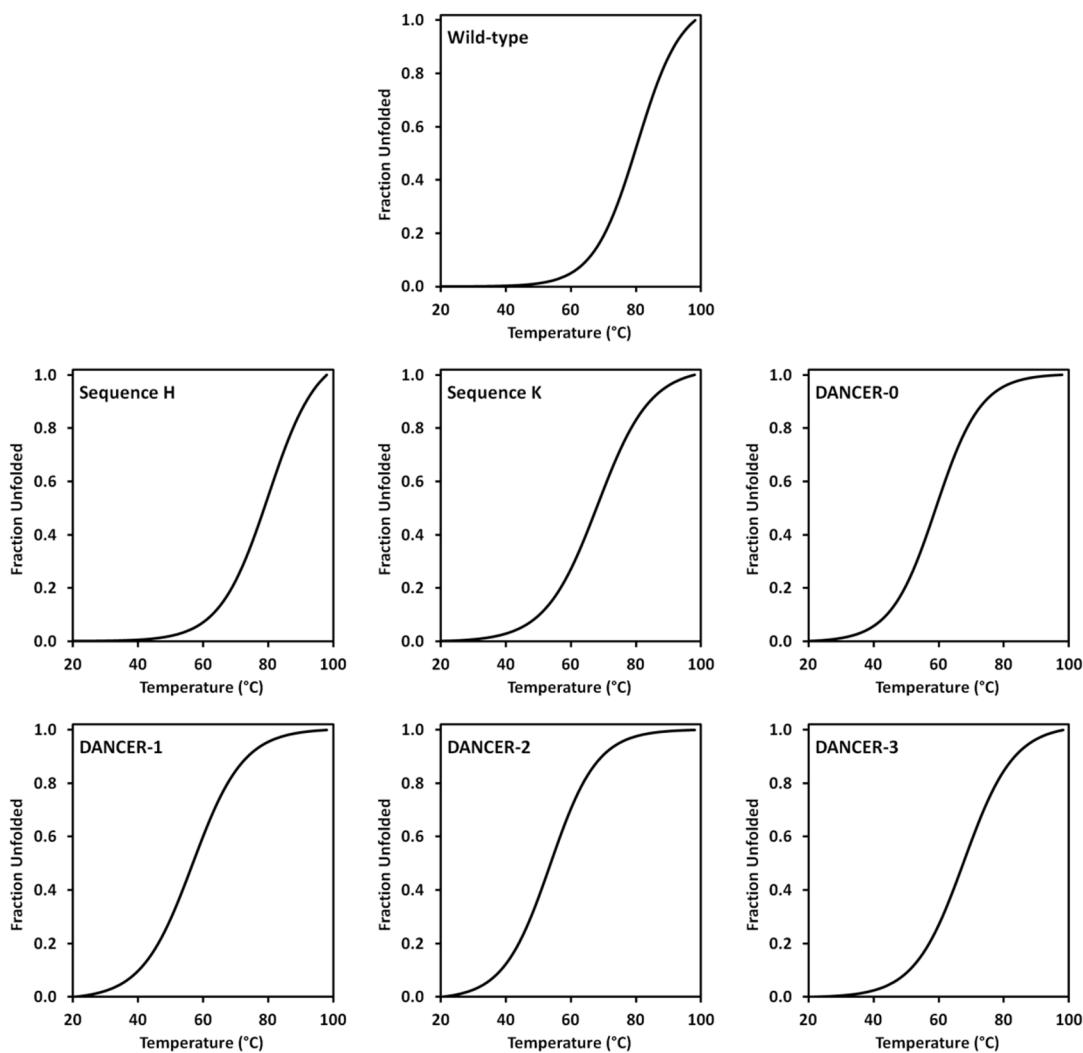


Figure 7.9. Thermal denaturation of test set proteins. Thermal denaturation was monitored by CD spectroscopy (208 nm) indicates that all test set proteins are stable at room temperature and pressure. Fraction unfolded values were calculated by computing loss of polarization with respect to the completely folded protein at atmospheric pressure and 20 °C. Thermal denaturation curves represent fits to experimental data (not shown).

Data detailing thermal denaturation of test set proteins was collected by CD spectroscopy. CD data was collected following purification and size exclusion chromatography using 40 μM samples. Aliquots of 650 μL for each sample were placed into 1 mm path length quartz cuvettes (Jasco) and melting curves were generated using a Jasco J-815 spectrometer. Samples were heated at a rate of $1\text{ }^{\circ}\text{C}\cdot\text{min}^{-1}$ and measuring CD at 208 nm following temperature equilibration at approximately $2\text{ }^{\circ}\text{C}$ intervals. All measurements were performed in 10 mM sodium phosphate buffer (pH 7.4). Spectra were converted to fraction of unfolded protein by assuming that the strongest dichroism signal corresponded to 100% folded protein and the weakest dichroism signal corresponded to 0% folded protein. The midpoint of thermal denaturation (T_m) was determined by fitting a 2-term (parameters T_m and a) sigmoid (EQ 7.1) using an in-house nonlinear least-squares regression program.

$$f(T) = \frac{1}{1 + e^{(T_m - T)/a}} \quad (\text{EQ 7.1})$$

The reported thermodynamic parameters were calculated by a Van't Hoff analysis of fitted curves using a plot of $\ln(K)$ and T^{-1} , y and x axes respectively (EQ 7.2). Equilibrium values (K) were calculated as the fraction of unfolded over folded signal at each data point.

$$\ln(K) = -\frac{\Delta H}{RT} + \frac{\Delta S}{T} \quad (\text{EQ 7.2})$$

Linear regression of each plot was performed solving slope (m) and intercept (b) parameters required to determine the enthalpy (ΔH_U , EQ 7.3) and the entropy (ΔS_U , EQ 7.4) of unfolding for each protein.

$$\Delta H_U = -R \times m \quad (\text{EQ 7.3})$$

$$\Delta S_U = R \times b \quad (\text{EQ 7.4})$$

Chemical denaturation assays were performed in triplicate using protein samples normalized to 1 mg/mL. Protein aliquots of 25 μL in individual wells of UV transparent 96-well plates (Greiner Bio-One) were mixed with 175 μL of 0 – 3 M or 0 – 5 M guanidium chloride solutions (12 points, evenly spaced), for proteins of decreased or similar stability with respect to the WT, respectively. These samples were incubated at room temperature for an hour. Following chemical denaturation, fluorescence emission spectra were measured from 300 nm to 450 nm (excitation at 295 nm and step size of 2 nm) using an Infinite M1000 plate reader (Tecan). Spectra were numerically integrated (EQ 7.5) and converted to fraction of unfolded protein by assuming that maximum fluorescence intensity corresponded to 100% folded protein and that minimum fluorescence intensity corresponded to 0% folded protein.

$$\int_{i=300\text{ nm}}^{j=450\text{ nm}} f(M) \cdot dx \cong \frac{1}{2} \sum_{k=1}^N (\lambda_{k+1} - \lambda_k) (f(k+1) + f(k)) \quad (\text{EQ 7.5})$$

Error on percent unfolded values is reported as standard deviation from three replicates at each denaturant concentration for each sequence. Values reporting on concentration of denaturant required to unfold half the population of protein (C_m) were determined by fitting a 2-term (parameters C_m and a) sigmoid (EQ 7.6) using the in-house nonlinear least-squares regression program.

$$f(M) = \frac{1}{1 + e^{(C_m - M)/a}} \quad (\text{EQ 7.1})$$

The free energy of unfolding at zero denaturant concentration (ΔG_U) and denaturant cooperativity (m-value) values were calculated by linear regression of ΔG (EQ 7.7) versus concentration of denaturant, where unfolded equilibrium values (K) are computed as a function of unfolded signal over folded signal.

$$G = -RT \cdot \ln(K) \quad (\text{EQ 7.1})$$

The ΔG_U and m-value parameters were determined from the intercept and slope of the linear regression.

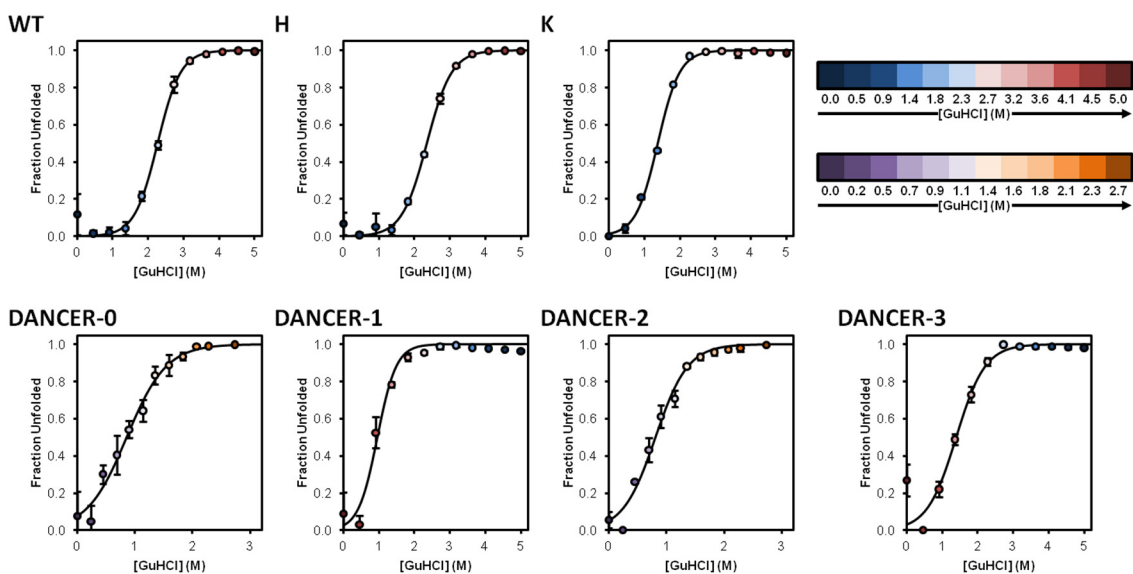


Figure 7.10. Chemical denaturation of test set proteins. Stability reported by chemical denaturation using guanidium chloride demonstrates that test set proteins unfold according to a two-state model. Fraction unfolded values calculated by monitoring integrated tryptophan fluorescence ($\lambda_{\text{excitation}} = 280 \text{ nm}$, $\lambda_{\text{excitation}} = 310 - 450 \text{ nm}$) relative to fluorescence at 0 M guanidium chloride.

Detailed study of the unfolding curves for the test set sequences produced by guanidium chloride chemical denaturation [Figure 7.10] reveals that each of the test set

proteins unfold with approximately the same cooperativity ($1.5 - 2.3 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{M}^{-1}$) and each according to a two state model [Table 7.5].

Table 7.5. Thermodynamic stability by thermal and chemical denaturation of test set proteins

Sequence	Chemical Denaturation			Thermal Denaturation (P $\approx$ 1 atm, T $\approx$ 298 K)			
	C _m (M)	ΔG_U (kcal $\cdot$ mol $^{-1}$)	m-value (kcal $\cdot$ mol $^{-1}\cdot$ M $^{-1}$)	T _m (°C)	ΔG_U (kcal $\cdot$ mol $^{-1}$)	ΔH_U (kcal $\cdot$ mol $^{-1}$)	ΔS_U (kcal $\cdot$ mol $^{-1}\cdot$ K $^{-1}$)
WT	2.26 $\pm$ 0.04	4.1 $\pm$ 0.2	1.8 $\pm$ 0.1	78.5	6.0	39.7	0.113
H	2.35 $\pm$ 0.02	4.0 $\pm$ 0.3	1.7 $\pm$ 0.1	77.8	5.4	36.3	0.103
K	1.38 $\pm$ 0.01	2.7 $\pm$ 0.1	1.9 $\pm$ 0.1	67.3	3.5	29.0	0.086
DANCER-0	0.85 $\pm$ 0.09	1.5 $\pm$ 0.2	1.8 $\pm$ 0.2	59.2	3.1	31.0	0.094
DANCER-1	0.95 $\pm$ 0.05	2.2 $\pm$ 0.1	2.3 $\pm$ 0.2	56.6	2.6	28.0	0.085
DANCER-2	0.81 $\pm$ 0.05	1.7 $\pm$ 0.1	2.1 $\pm$ 0.1	54.0	2.4	28.6	0.088
DANCER-3	1.39 $\pm$ 0.05	2.0 $\pm$ 0.3	1.5 $\pm$ 0.2	67.0	3.6	30.4	0.090

Having demonstrated that the test set proteins adopt stable, soluble, and monomeric protein folds, the next step in the characterization of the test set proteins involved solution NMR experiments designed by *Dr. Natalie K Goto* and conducted and analyzed by *Adam M Damry*. Proteins for NMR spectroscopy were expressed and purified as indicated above, with the exception that proteins for NMR spectroscopy were instead expressed using M9 minimal expression media supplemented with 1 g $\cdot$ L $^{-1}$ 15 N-ammonium chloride and/or 3 g $\cdot$ L $^{-1}$ 13 C-D-glucose for isotopic enrichment. 15 N- and 13 C-labelled GB1 samples comprised approximately 0.1 – 2.0 mM protein in a 10 mM sodium phosphate buffer (pH 7.4) with 10 μ M EDTA prepared in 10% D₂O for experiments requiring HN detection and 99% D₂O for the remainder. To address the proteolysis of protein samples [Figure 7.8] that complicates NMR data acquisition,

EDTA-free Protease Inhibitor Cocktail (Roche) was included when appropriate, specifically for long term study of DANCER-1 and DANCER-2 during NMR experiments. All NMR experiments were performed on either a Varian INOVA 500 MHz spectrometer equipped with a triple resonance inverse probe, or a Bruker AVANCEIII HD 600 MHz spectrometer equipped with a triple resonance cryoprobe.

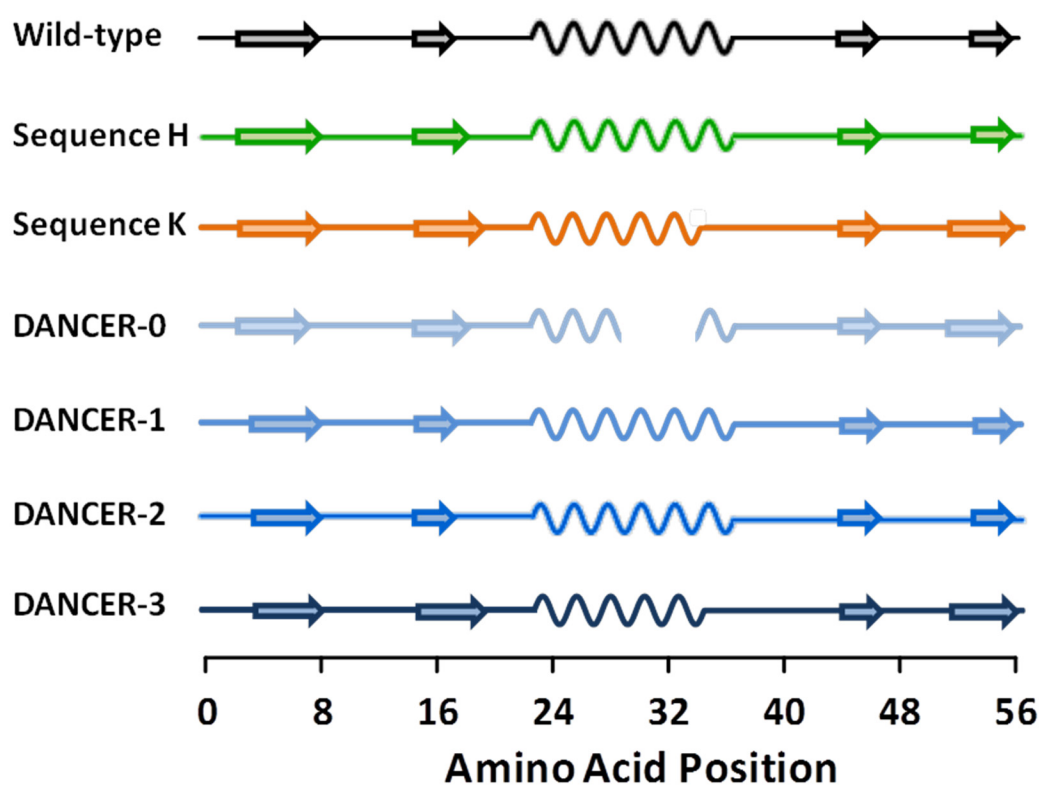


Figure 7.11. Secondary structure of test set proteins by NMR chemical shift index analysis. Figure provided courtesy of *Adam M Damry*.

To determine whether the test set proteins adopt the G β 1 fold, chemical shift indices were measured to identify the arrangement of secondary structural elements across their respective sequences [Figure 7.11]. Chemical shift assignment data was

performed at 25 °C. NMR data was processed with the NMRPipe software package [Delaglio *et al.*, 1995] and spectra were analyzed with NMRViewJ (One Moon Scientific) [Johnson and Blevins, 1994]. Backbone and side-chain chemical shift assignments were obtained from the standard suite of 3D triple resonance experiments, including HSQC, HNCO, HNCACB and CBCA(CO)NH spectra. Analysis of chemical shift indices [Wishart *et al.*, 1992] indicates that the secondary structure arrangement of sequence H, K, DANCER-0, DANCER-1, DANCER-2, and DANCER-3 are consistent with that of the WT sequence.

To determine whether the DANCER proteins adopt either a single conformation or multiple conformations in solution, ^1H - ^{15}N -HSQC spectra were collected [Figure 7.12]. This revealed that three of the four dynamic candidate proteins (DANCER-1, DANCER-2, and DANCER-3) show doubling of peaks, suggesting that they form two distinct conformational states in solution. The ^1H - ^{15}N -HSQC spectrum of DANCER-0 exhibited peak broadening, indicative of dynamics on a faster timescale. Interestingly, full backbone assignment of DANCER-0 could not be completed due to significant peak broadening and loss of signal for specific residues belonging to the α -helix. Furthermore, the calculated energy of the $-t$ state conformation relative to its minimum energy state, serving as the energetic proxy of the exchange activation barrier of DANCER-0 ($\Delta E^\ddagger = 7.8 \text{ kcal}\cdot\text{mol}^{-1}$), is the lowest energy reported for the set of DANCER proteins. Conversely, the ^1H - ^{15}N -HSQC spectrum of sequences H and K [Figure 7.13] show only a single set of peaks, similar to that of the WT sequence, indicating that they form a single conformation in solution.

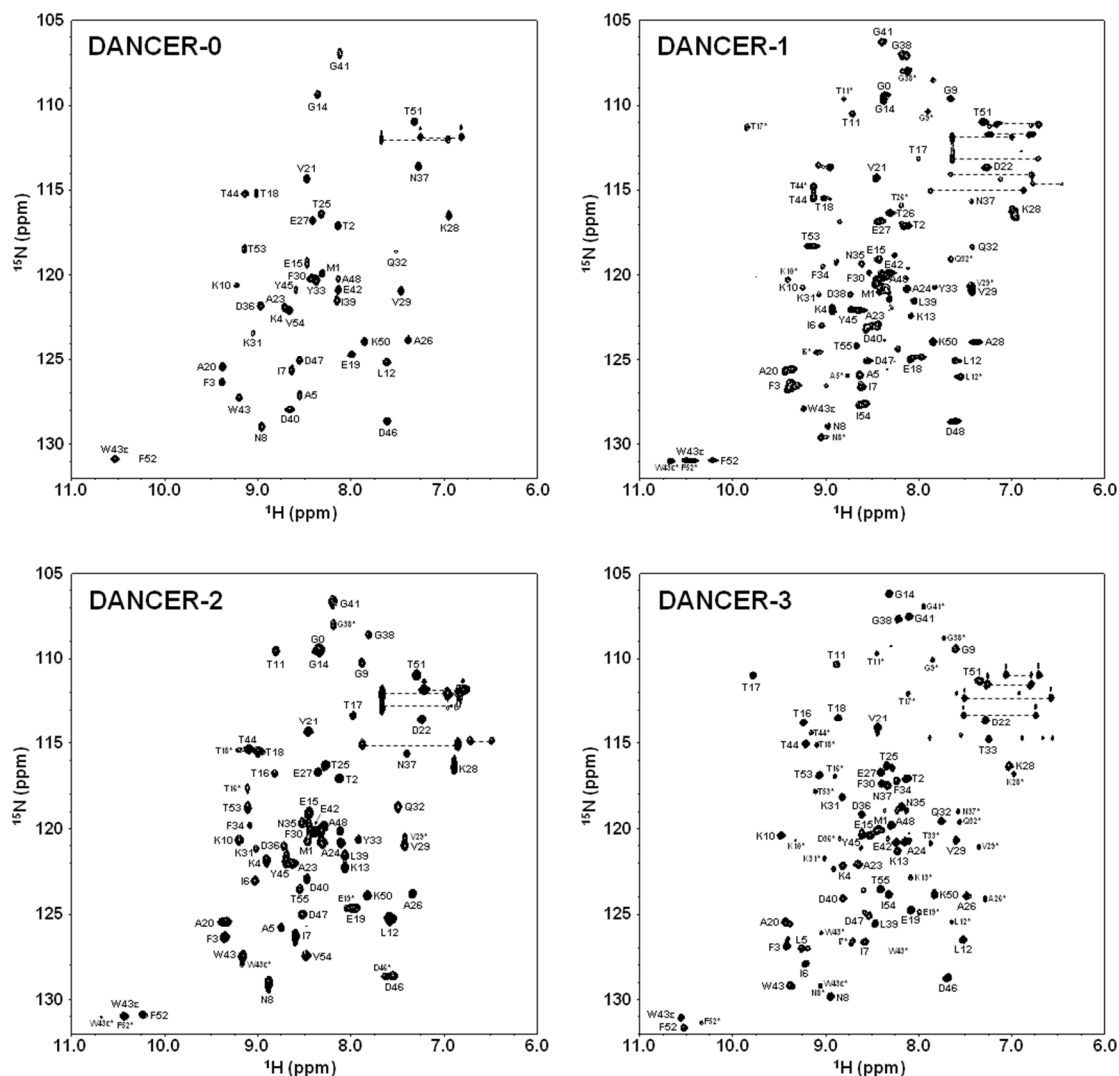


Figure 7.12. ^1H - ^{15}N -HSQC NMR spectroscopy of DANCER proteins. DANCER NMR spectra are provided courtesy of Adam M Damry.

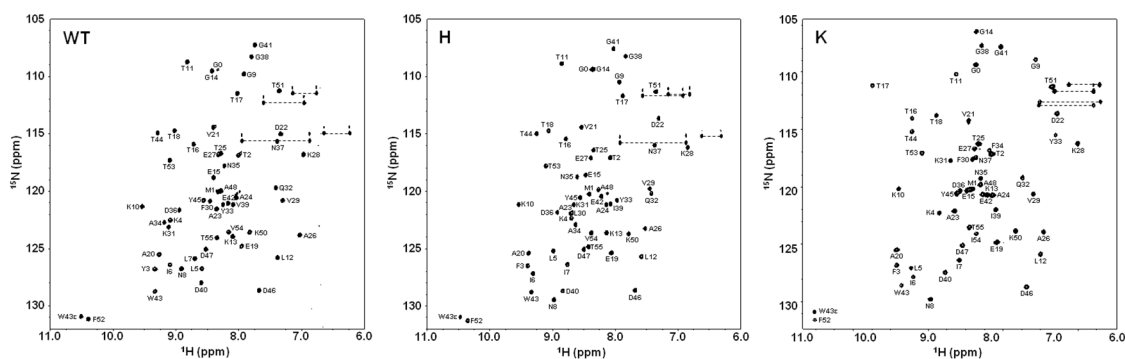


Figure 7.13. ^1H - ^{15}N -HSQC NMR spectroscopy of static G β 1 proteins. NMR spectra of the WT G β 1 protein along with mutant sequences H and K are provided courtesy of Adam M Damry.

While ^1H - ^{15}N -HSQC spectroscopy suggest that two DANCER-1, DANCER-2, and DANCER-3 proteins adopt two stable conformations, the timescale of their interconversion was not known. To confirm that DANCER-1, DANCER-2, and DANCER-3 undergo conformational exchange, ZZ-exchange spectroscopy experiments were performed at temperatures varying from 5 °C to 25 °C. Rates of exchange were determined by fitting 4-term relaxation and exchange curves using nonlinear least-squares regression, and thermodynamic parameters were determined by fitting exchange rates and temperatures to the Eyring equation. These experiments confirmed that the DANCER-1, DANCER-2, and DANCER-3 undergo conformational exchange, and enabled measurement of the kinetic and thermodynamic exchange parameters for DANCER-1 and DANCER-3 reported in Table 7.6. Unfortunately, the faster timescale of exchange (< ms) from the minor state to the major state and their relative population for DANCER-2 prohibits the measurement of its kinetic and thermodynamic exchange parameters by ZZ-exchange. Nonetheless, these three of the four DANCER proteins are dynamic and exchange between two distinct conformations on the millisecond timescale.

Table 7.6. Kinetic parameters of DANCER proteins determined by ZZ-exchange NMR spectroscopy

Sequence	$K_{\text{minor}} (\text{s}^{-1})$	$K_{\text{major}} (\text{s}^{-1})$	$K_{\text{minor/major}}$	$\Delta G^\ddagger (\text{kcal}\cdot\text{mol}^{-1})$
DANCER-1	60.7	79.1	1.30	14.4
DANCER-3	3.02	17.4	5.76	16.2

All parameters are observed at 15 °C

Exchange peaks observed for DANCER-2 by ZZ-exchange NMR spectroscopy

Preliminary data for the exchange parameters [Table 7.6] reveal that DANCER-1 interconverts between its major and minor species more rapidly than DANCER-3, reflected by its lower experimentally observed activation barrier ($0.6 - 1.8 \text{ kcal}\cdot\text{mol}^{-1}$), depending on the direction of its exchange. This result is in agreement with the energy difference calculated by *meta*-MSD between the $-t$ state and their respective minimum energy state conformations, which serves as the energetic proxy for the exchange activation barrier [Table 7.3], with the predicted barrier of DANCER-1 ($\Delta E^\ddagger = 8.4 \text{ kcal}\cdot\text{mol}^{-1}$) being lower than that of DANCER-3 ($\Delta E^\ddagger = 13.7 \text{ kcal}\cdot\text{mol}^{-1}$).

It is important to note that although the set of DANCER proteins do exhibit dynamic exchange as predicted, the atomic configuration of their conformational states had not yet been determined. For the *meta*-MSD framework to be truly predictive the atomic structure of the test set sequences must match the intended designed state conformations. Nonetheless, the behaviour of the seven test set G β 1 sequences can be correctly identified as static (WT, H, and K) and dynamic (DANCER-0, DANCER-1, DANCER-2, and DANCER-3).

7.6. Validation of DANCER Conformational States

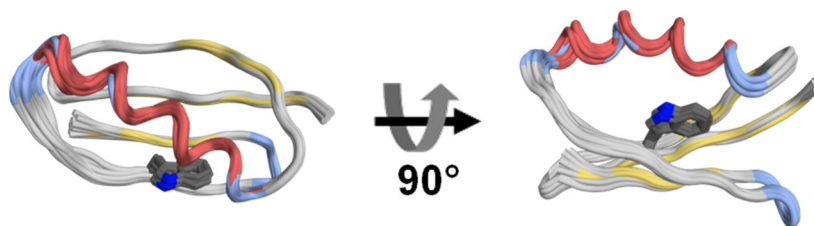
To confirm that the ensemble models used in the design of dynamic exchange accurately resemble the true conformations of the DANCER proteins and that the design objective of exchange between the $-g(-) \leftrightarrow -t \leftrightarrow -g(+)$ state conformations was achieved, *Adam M Damry* and *Dr. Natalie K Goto* collected NH-TOCSY, CCH-TOCSY and HCCH-TOCSY spectra to facilitate side-chain assignments and defined distance

restraints, by recording ^{15}N -edited and ^{13}C -edited HSQC-NOESY spectra of the DANCER proteins to produce atomic resolution data. Due to the difficulty presented, specifically discerning distance restraints for each of the major and minor state species, when attempting to solve solution NMR structures of proteins that exist and spontaneously interconvert between two conformations on the same timescale as the NMR experiments, DANCER-2 was the only mutant for which a solution structure could be solved. The structure of the major and minor states of the DANCER-2 protein could be successfully modelled with solution NMR data because this mutant uniquely populates its major state by approximately a 10-to-1 ratio. This allowed *Adam M Damry* and *Dr. Natalie K Goto* to differentiate between ^1H - ^{15}N -HSQC-NOESY signal contributions arising from the major state having strong signals from the aspartic acid residues at positions 45 and 46, and the minor state populations and solve a structure of the major species [Figure 7.14.A] and a model of the minor species [Figure 7.14.B].

NMR structure solutions were generated using TALOS+ [Shen *et al.*, 2009] to determine secondary structure propensities and backbone dihedral restraints for G β 1 variants on the basis of measured chemical shifts for ^1H , ^{15}N , $^{13}\text{C}_\alpha$, and $^{13}\text{C}_\beta$ chemical shifts. The structure of the major state of DANCER-2 were then calculated using the CYANA 2.1 program [Herrmann *et al.*, 2002] using the dihedral restraints and chemical shifts with ^1H - ^1H NOEs as distance restraints, resulting in a total of 417 unique and non-redundant distance restrains. These restraints were used in several rounds of manual refinement that resulted in NMR ensembles comprised of the 10 to 20 lowest energy conformers. Key to the solution of the minor state was the observation that the major

state structure of DANCER-2 did not satisfy a network of NOEs to and from the sidechain of tryptophan 43 to the core of the protein, suggesting a different Trp 43 conformation exists for the minor species.

A. DANCER-2 major species solution-NMR ensemble



B. DANCER-2 minor species solution-NMR model

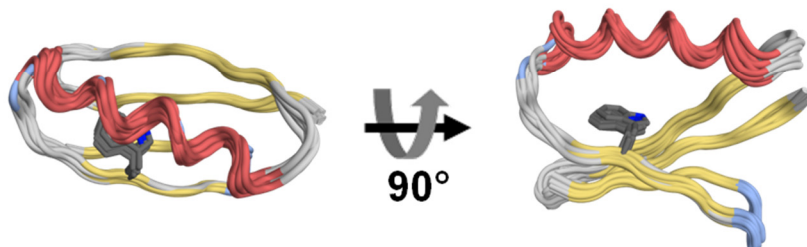


Figure 7.14. Solution NMR ensembles of DANCER-2 protein. Structures of the major species (A) and models of the minor species (B) of DANCER-2 were rendered. Colors indicate secondary structures: α -helix (red), β -strand (yellow), loop (blue), and undefined structure (white). The conformations of the tryptophan residue at position 43 are shown with the major and minor species adopting the $-g(+)$ and $-g(-)$ states, respectively. Coordinates for the DANCER-2 protein were provided courtesy of *Adam M Damry*.

Consistent with its chemical shift index [Figure 7.11], the solution NMR ensemble of the DANCER-2 major species displays secondary structures in an arrangement consistent with the G β 1 fold [Figure 7.14.A]. The ensemble has backbone and heavy

atom RMSD values [Table 7.7] of 2.00 ± 0.07 Å and 3.02 ± 0.08 Å with respect to the crystallographic structure of WT Gβ1 (1PGA) [Gallagher *et al.*, 1994], and 1.85 ± 0.08 Å and 2.76 ± 0.07 Å with respect to a previously published solution NMR structure of WT Gβ1 (1GB1) [Gronenborn *et al.*, 1991]. The structural diversity of the solution NMR DANCER-2 major state ensemble, as measured by the average pair-wise RMSD between its members, was measured at 0.64 ± 0.17 Å and 1.25 ± 0.17 Å, for backbone and heavy atoms, respectively. Despite the fact that the NMR solutions structures for the major species of DANCER-2 adopt the Gβ1 fold, having similar core packing, overall topology, and secondary structures, these coordinate RMSD results indicate considerable divergence from the designed structure. The cause of these discrepancies originates predominantly from the packing of the third β-strand of the Gβ1 fold against the α-helix. The use of structural comparison methodologies that reduce the reliance of comparison on explicit atomic coordinates, such as TM-score [Xu and Zhang, 2010; Zhang and Skolnick, 2004], would likely result in an analysis indicating structural significant similarity between the design models and the NMR solution structure data.

Table 7.7. RMSD coordinate comparisons of solution NMR ensembles

Sequence	Backbone RMSD (Å)			Heavy atom RMSD (Å)		
	Ensemble Diversity	1PGA	1GB1	Ensemble Diversity	1PGA	1GB1
Static Gβ1 Proteins						
H	0.51 ± 0.11	1.13 ± 0.06	1.39 ± 0.09	1.08 ± 0.10	1.97 ± 0.05	2.06 ± 0.08
K	0.57 ± 0.15	1.85 ± 0.10	2.28 ± 0.13	1.12 ± 0.17	2.88 ± 0.10	3.13 ± 0.12
DANCER-2						
Major species	0.64 ± 0.17	2.00 ± 0.07	1.85 ± 0.08	1.25 ± 0.17	3.02 ± 0.08	2.76 ± 0.07
Minor Species	1.01 ± 0.20	2.13 ± 0.11	2.21 ± 0.18	1.73 ± 0.19	3.06 ± 0.07	2.91 ± 0.13

Importantly, χ_1 and χ_2 dihedral angles of the tryptophan 43 residue across the ensemble are found at $73.9 \pm 0.6^\circ$ and $-70.3 \pm 2.9^\circ$, respectively [Table 7.8]. Thus, the solution NMR structure demonstrates that the major species of DANCER-2 occupies the $-g(+)$ state conformation. It is important to note however that this structure does not accommodate all observed tryptophan 43 indole ^{15}N -NOE correlations [Figure 7.15]. Omission of the strong tryptophan 43 indole ^{15}N -NOE correlations with residues 46 and 47, which enforce the exclusive generation of the major species during CYANA calculations, allows the tryptophan 43 residue to satisfy alternate ^{15}N -NOE correlations observed from core positions in the G β 1 scaffold.

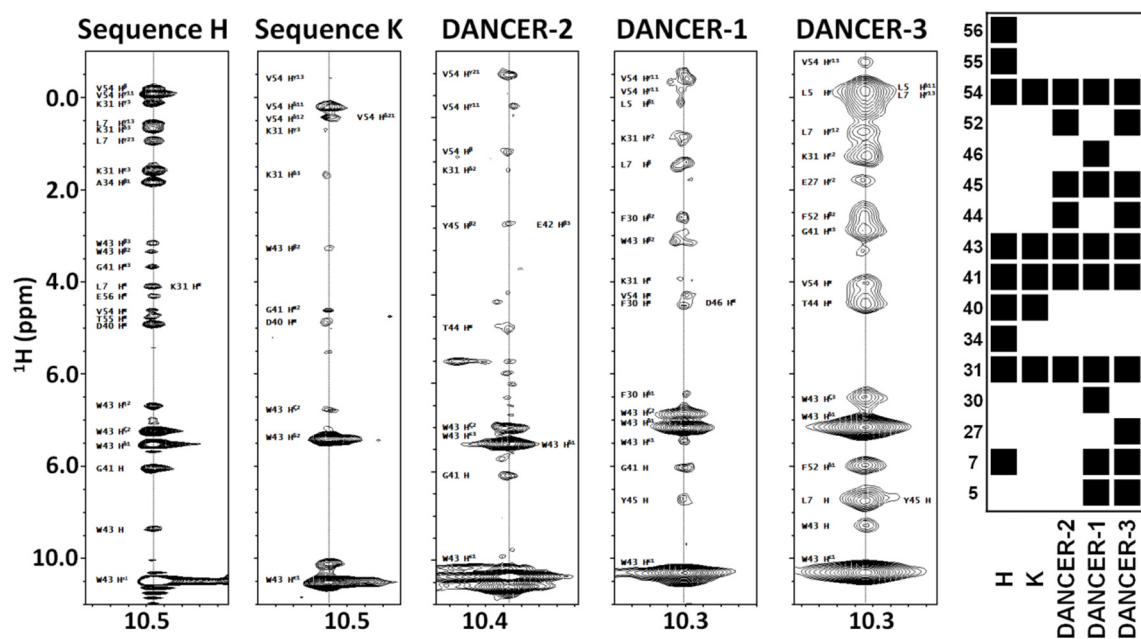


Figure 7.15. Test set tryptophan 43 ^{15}N -NOE correlations. ^{15}N -NOE correlation strips of tryptophan 43 for sequences H, K, DANCER-1, DANCER-2, and DANCER-3, with residues in their respective folds. The far right plot indicates residues (black squares) that show ^{15}N -NOE correlations with the indole N_H proton.

By omitting the strong ^{15}N -NOE correlations between the tryptophan 43 indole N_H proton with residues 46 and 47, *Adam M Damry* and *Dr. Natalie K Goto* were able to generate a model for the structure of the minor species of DANCER-2 which satisfied the remaining signals in the NOE network [Figure 7.14.B]. This model ensemble has backbone and heavy atom RMSD values [Table 7.7] of $2.13 \pm 0.11 \text{ \AA}$ and $3.06 \pm 0.07 \text{ \AA}$ with respect to the crystallographic structure of WT G β 1 (1PGA) [Gallagher *et al.*, 1994], and $2.21 \pm 0.18 \text{ \AA}$ and $2.91 \pm 0.13 \text{ \AA}$ with respect to a previously published solution NMR structure of WT G β 1 (1GB1) [Gronenborn *et al.*, 1991]. The structural diversity of the model DANCER-2 minor species ensemble was measured at $1.01 \pm 0.20 \text{ \AA}$ and $1.73 \pm 0.19 \text{ \AA}$, for backbone and heavy atoms, respectively. Importantly, χ_1 and χ_2 dihedral angles of the tryptophan 43 residue across the ensemble are found at $-95.0 \pm 1.3^\circ$ and $-110.0 \pm 2.4^\circ$, respectively [Table 7.8]. Thus, the model solution NMR ensemble demonstrates that the minor species of DANCER-2 occupies the $-g(-)$ state conformation. Thus, the position of the tryptophan 43 residue in the minor species model satisfies the remaining observed ^{15}N -NOE correlations [Figure 7.15]. Again, coordinate RMSD metrics indicate similar discrepancies between design models and the NMR solutions structures for the minor species of DANCER-2 as observed for comparisons of the design models against the NMR solution structures for the major species of DANCER-2.

Table 7.8. Ensemble tryptophan conformations for solution NMR structures and models of sequences H, K, and the major and minor species of DANCER-2

Sequence	State	Tryptophan χ_1 Angle (°)		Tryptophan χ_2 Angle (°)	
Static Gβ1 Proteins					
H	+g(−)	−84.1 ± 3.5	[−88.5 : −76.7]	80.4 ± 4.1	[74.2 : 85.7]
K	−g(+)	50.7 ± 3.1	[46.5 : 54.0]	−88.3 ± 1.8	[−91.9 : −86.0]
DANCER-2					
Major species	−g(+)	73.9 ± 0.6	[72.7 : 75.6]	−70.3 ± 2.9	[−74.9 : −65.0]
Minor species	−g(−)	−95.0 ± 1.3	[−97.9 : −93.6]	−110.0 ± 2.4	[−112.2 : −105.4]

The ensemble average ± standard deviation [minimum : maximum] values for dihedral angles are reported

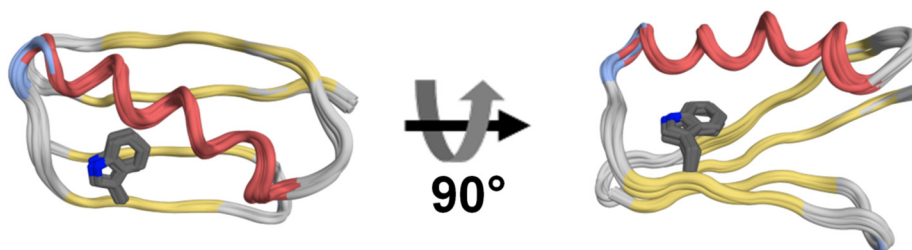
While solution NMR structures and models allow for atomic-level characterization of the major and minor species of the DANCER-2 protein, the structures of the major and minor species of DANCER-1 and DANCER-3 cannot be modeled as it is not possible to deconvolute and differentiate peaks arising from both the major and minor species in NH-TOCSY, CCH-TOCSY, HCCH-TOCSY, and ^1H - ^{15}N -HSQC-NOESY spectra. Instead, characterization of the tryptophan 43 conformation in the major and minor species of DANCER-1 and DANCER-3 must be completed by comparing their ^1H - ^{15}N -HSQC and ^1H - ^{15}N -HSQC-NOESY spectra with alternate protein sequences. Visual inspection shows considerable similarity between the ^1H - ^{15}N -HSQC spectra of DANCER-1 and DANCER-2 [Figure 7.12]. Furthermore, their sequences differ by a single mutation at position 54 [Table 7.4] suggesting that their major and minor species may also be similar. This is further supported by DANCER-1 tryptophan 43 indole ^{15}N -NOE correlations throughout the G β 1 fold that closely match the pattern of tryptophan 43 indole ^{15}N -NOE correlations obtained with DANCER-2 [Figure 7.15].

Unlike DANCER-1, DANCER-3 presents a ^1H - ^{15}N -HSQC spectrum that is observably dissimilar to that of DANCER-2 [Figure 7.12]. An alternate G β 1 variant, sequence K, differs from DANCER-3 by a single mutation at position 39 and has a ^1H - ^{15}N -HSQC spectrum [Figure 7.13] that is visually similar to that of the major species of DANCER-3 [Figure 7.12]. In order to determine the exchanging tryptophan conformations in DANCER-3, Adam M Damry and Dr. Natalie K Goto solved the solution structure of the mutant G β 1 sequence K predicted to be static with its tryptophan occupying the $-g(+)$ state conformation using the *meta*-MSD framework [Table 7.4].

The solution NMR ensemble of sequence K, build from 507 unique distance restraints, has backbone and heavy atom RMSD values [Table 7.7] of 1.85 ± 0.10 Å and 2.88 ± 0.10 Å with respect to the crystallographic structure of WT G β 1 (1PGA) [Gallagher *et al.*, 1994], and 2.28 ± 0.13 Å and 3.13 ± 0.12 Å with respect to the previously published solution NMR structure of WT G β 1 (1GB1) [Gronenborn *et al.*, 1991]. The structural diversity of the ensemble was measured at 0.57 ± 0.15 Å and 1.12 ± 0.17 Å, for backbone and heavy atoms, respectively. The ensemble shows that the tryptophan residue occupies a single conformation, correctly predicted by *meta*-MSD as the $-g(+)$ state conformation, with χ_1 and χ_2 dihedrals of $50.7 \pm 3.1^\circ$ and $-88.3 \pm 1.8^\circ$, respectively [Table 7.8]. As expected, analysis of sequence K's ^{15}N -NOE spectrum shows a loss of correlations made from the tryptophan 43 side chain indole with core G β 1 residues (residues 44, 45, 46, and 52) reflecting its solvent exposed orientation as the tryptophan is no longer integrated into the protein core [Figure 7.15]. Given the similarity between the ^1H - ^{15}N -HSQC spectra of sequence K and DANCER-3, and the collective similarity of

the ^{15}N -NOE correlations observed for DANCER-1, DANCER-2, and DANCER-3 proteins, it is likely that the major species of DANCER-3 adopts the $-g(+)$ state conformation and that its minor species adopts the $-g(-)$ state conformation.

A. Sequence H solution-NMR ensemble



B. Sequence K solution-NMR ensemble

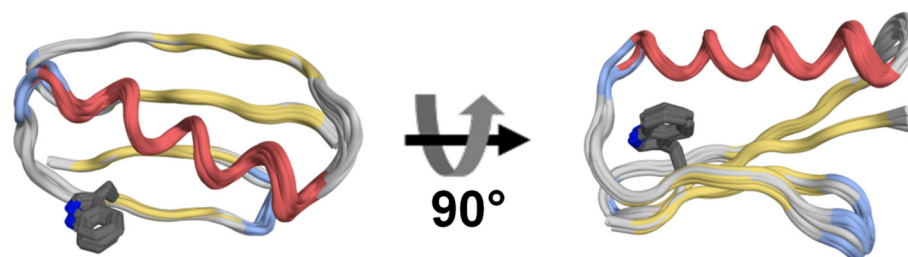


Figure 7.16. Solution NMR ensembles of static test set proteins. Structures of sequence H (A) and sequence K (B) were rendered. Colors indicate secondary structures: α -helix (red), β -strand (yellow), loop (blue), and undefined structure (white). The conformations of the tryptophan residue at position 43 are shown with sequences H and K adopting the $+g(-)$ and $-g(+)$ states, respectively. Coordinates for sequences H and K were provided courtesy of *Adam M Damry*.

While comparison to DANCER-2 and sequence K give good evidence that the major species of DANCER-1 and DANCER-3 adopt the intended $-g(+)$ conformation, ambiguity exists as to whether the minor species of DANCER-1 and DANCER-3 adopt the intended $-g(-)$ state conformation or the alternate and unintended $+g(-)$ state

conformation. To conclusively determine that the major and minor species of the DANCER proteins all adopt their intended respective $-g(+)$ and $-g(-)$ state conformations, the tryptophan 43 indole ^{15}N -NOE correlations of the DANCER proteins can be collectively compared with the pair of static designed sequences H and K [Figure 7.16]. Sequence H was chosen for comparison because it was designed and experimentally determined by solution NMR (build using 753 unique distance restraints) to exist in the single $+g(-)$ state conformation [Figure 7.16], similar to that of the WT structure, with tryptophan 43 χ_1 and χ_2 dihedrals of $-84.1 \pm 3.5^\circ$ and $74.2 \pm 4.1^\circ$, respectively [Table 7.8]. Careful examination of the sequence H tryptophan 43 indole ^{15}N -NOE spectrum reveals the presence of two correlations with residues 55 and 56 at the C-terminal of G β 1 that are absent in the spectra of sequence K as well as DANCER-1, DANCER-2, and DANCER-3 proteins. Additionally, new correlations observed between the tryptophan indole and residues 44, 45, 46, and 52 in the core of DANCER-1, DANCER-2, and DANCER-3 are not present in sequences H and K. Together, this set of results [Figure 7.17] strongly suggest that the minor species of the DANCER proteins is $-g(-)$, as design by our *meta*-MSD calculations. Thus, application of this new *meta*-MSD framework has enabled the rational design of a specific conformational exchange into the rigid G β 1 scaffold.

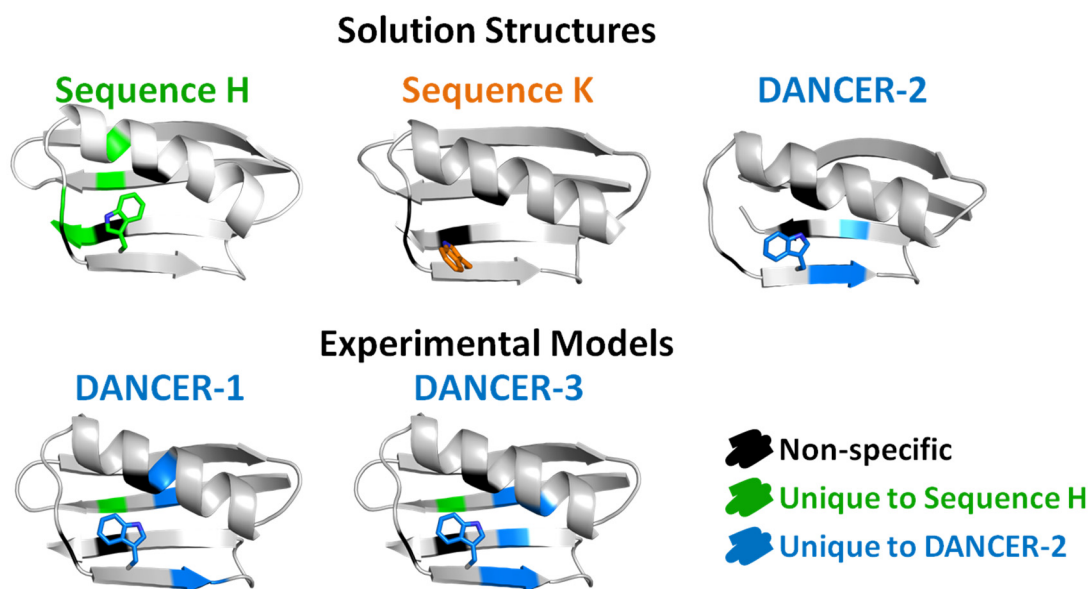


Figure 7.17. Tryptophan 43 orientation relative to its ^{15}N -NOE correlations for the minor species of DANCER proteins. Correlations unique to sequence H (green) are lost in DANCER-1, DANCER-2, and DANCER-3. Correlations corresponding to the minor species of DANCER-2, not found in sequence H or K, are shared with DANCER-1 and DANCER-3. These results indicate that the DANCER minor species is the $-g(-)$ state conformation. This figure was provided courtesy of Adam M Damry.

7.7. Comparison of DANCER *meta*-Multistate Design Models and Solution NMR Structures

Having experimentally validated the protein dynamics predictions made using the *meta*-MSD framework, it is important that the degree of quantitative accuracy and quality of their design models be assessed. The use of geometric constraints to partition the set of all designed sub-states into their respective sub-states is central to the methodology. While the end states of the dynamic exchange can be experimentally observed and readily defined *in silico* as the $-g(+)$ and $-g(-)$ state conformations, the

structure of the transition state controlling the kinetics of exchange and its design model proxy, the $-t$ state conformation, has not been experimentally observed and validated. While it is unlikely that the geometry $-t$ conformation represents the exact atomic configuration of the transition state, it is plausible that the trajectory of exchange visits protein configurations sampling the $-t$ state conformation. The ability of the *meta*-MSD framework to rank sequences according to their rate of exchange or predict the activation energy would serve as substantial evidence in support of the model. Unfortunately, the kinetics of exchange could only be experimentally determined for two DANCER proteins, a sample size too small to permit a quantitative analysis. However, from a qualitative perspective, the *meta*-MSD calculation was able to correctly rank-order DANCER-0 ($\Delta E^\ddagger = 7.8 \text{ kcal}\cdot\text{mol}^{-1}$), DANCER-1 ($\Delta E^\ddagger = 8.4 \text{ kcal}\cdot\text{mol}^{-1}$), and DANCER-3 ($\Delta E^\ddagger = 13.7 \text{ kcal}\cdot\text{mol}^{-1}$) sequences in order of their increasing exchange timescales [Table 7.4 and Table 7.6].

The second energetic parameter that must be correctly calculated by the *meta*-MSD framework is the equilibrium energy (ΔE_{EQ}) between the major and minor species. For the model to be accurate, the equilibrium energy value should be capable of correctly reporting on the identity and relative populations of the major and minor species. While it was possible to rationally design sequences exhibiting exchange between the desired $-g(-) \leftrightarrow -t \leftrightarrow -g(+)$ state conformations the calculation methodology did not consistently assign the correct conformation of the major state nor could it predict their populations with quantitative accuracy [Table 7.4 and Table 7.6]. Specifically, the major state of DANCER-1, DANCER-2, and DANCER-3 proteins were

predicted to occupy the $-g(+)$, $-g(-)$, and $-g(+)$ state conformations, respectively, while experiment suggests that they all occupy the $-g(+)$ state conformation. Nonetheless, the *meta*-MSD framework was able to correctly distinguish between sequences that are (DANCER proteins) and are not (sequences H and K) dynamic.

The assessment of protein dynamics behaviour required that *meta*-MSD could correctly evaluate both the exchange equilibrium (ΔE_{EQ}) and barrier ($\Delta E^\ddagger$) energies for the searched sequence space. The success of the predictions can be attributed to the calculations ability to correctly partition designed sub-state configurations into their respective states. While it is promising that the *meta*-MSD framework proved qualitatively predictive, examination of model structural features and their comparison against atomic resolution data of the experimentally characterized test set proteins would allow for an evaluation of the model's structural accuracy. With this in mind, χ_1 and χ_2 dihedral angles of the tryptophan 43 residue were calculated for the stable sub-state design models of the DANCER-2 protein as well as the static sequences H and K [Table 7.9].

The set of stable design sub-state models were assigned to state conformations and used to predict each sequences exchange equilibrium (ΔE_{EQ}) and barrier ($\Delta E^\ddagger$) energies. Figure 7.18 depicts a comparison between the tryptophan χ_1 and χ_2 dihedral angles belonging to the solution NMR structures [Table 7.8], stable design sub-state models [Table 7.9], and set of all sub-states belonging to the mega ensemble [Table 7.2] for sequences H, K, and the major and minor states of the DANCER-2 protein.

Table 7.9. Tryptophan conformations found in model *meta*-MSD sub-states

State	Sub-states (n)	Tryptophan χ_1 Angle (°)			Tryptophan χ_2 Angle (°)		
H							
+g(−)	50	−71.1 ± 3.9	[−79.1 : −57.2]		86.0 ± 16.1	[71.3 : 115.5]	
+t	2	170.5 ± 0.4	[−189.2 : 170.2]		83.7 ± 13.5	[74.1 : 93.2]	
+g(+)	29	66.6 ± 5.0	[55.3 : 72.2]		9.0 ± 8.7	[0.1 : 46.9]	
−g(−)	5	−73.0 ± 8.7	[−87.2 : −64.8]		−91.5 ± 8.9	[−103.7 : −79.3]	
−t	89	−169.9 ± 8.3	[−160.6 : 163.6]		−93.9 ± 5.4	[−112.2 : −79.7]	
−g(+)	51	68.2 ± 4.2	[53.1 : 71.2]		−69.5 ± 20.9	[−90.0 : −2.4]	
K							
+g(−)	4	−65.7 ± 6.5	[−73.5 : −57.5]		34.8 ± 27.8	[15.8 : 74.8]	
+t	91	175.4 ± 6.1	[−163.6 : 167.3]		87.3 ± 16.1	[7.3 : 102.3]	
+g(+)	17	57.9 ± 9.1	[44.8 : 74.9]		43.9 ± 34.5	[1.8 : 104.8]	
−g(−)	3	−60.1 ± 0.3	[−60.4 : −59.8]		−99.4 ± 5.7	[−104.8 : −93.4]	
−t	1	−167.1			−93.6		
−g(+)	18	59.9 ± 7.3	[53.1 : 71.2]		−44.1 ± 25.1	[−82.0 : −23.6]	
DANCER-0							
+g(−)	5	−64.3 ± 10.8	[−76.2 : −56.3]		81.1 ± 7.7	[73.2 : 90.7]	
+t	274	173.4 ± 5.7	[−163.6 : 162.8]		88.1 ± 13.0	[3.0 : 102.4]	
+g(+)	12	58.9 ± 7.1	[46.8 : 69.5]		28.2 ± 25.4	[1.9 : 76.7]	
−g(−)	20	−68.1 ± 9.3	[−86.3 : −55.5]		−93.1 ± 6.7	[−103.7 : −83.5]	
−t	22	178.3 ± 11.9	[−162.7 : 165.6]		−94.8 ± 12.8	[−114.6 : −78.5]	
−g(+)	41	68.3 ± 3.7	[53.3 : 71.0]		−73.1 ± 13.8	[−83.1 : −23.6]	
DANCER-1							
+g(−)	10	−66.0 ± 9.1	[−76.2 : −53.9]		61.8 ± 22.0	[27.3 : 85.2]	
+t	669	172.8 ± 4.8	[−163.6 : 162.8]		88.4 ± 11.8	[3.0 : 102.4]	
+g(+)	14	58.5 ± 8.4	[46.6 : 68.6]		48.6 ± 38.5	[8.8 : 94.0]	
−g(−)	21	−68.9 ± 12.1	[−87.7 : −42.7]		−91.6 ± 7.9	[−105.5 : −83.3]	
−t	10	−178.5 ± 11.7	[−162.9 : 167.4]		−92.4 ± 13.1	[−115.8 : −79.8]	
−g(+)	12	69.6 ± 1.2	[67.9 : 71.0]		−73.2 ± 14.0	[−78.1 : −29.0]	
DANCER-2							
+g(−)	4	−55.8 ± 1.3	[−56.5 : −53.8]		76.6 ± 5.6	[73.2 : 84.9]	
+t	438	172.7 ± 5.0	[−163.6 : 162.8]		88.8 ± 12.1	[3.0 : 102.4]	
+g(+)	14	57.4 ± 8.1	[46.6 : 70.3]		37.9 ± 32.6	[1.9 : 94.0]	
−g(−)	15	−67.7 ± 8.6	[−79.4 : −57.5]		−92.5 ± 7.6	[−103.7 : −83.5]	
−t	18	176.8 ± 11.5	[−162.9 : 165.6]		−93.8 ± 13.8	[−114.6 : −78.5]	
−g(+)	15	69.0 ± 1.0	[67.5 : 70.2]		−73.0 ± 12.4	[−81.4 : −29.0]	
DANCER-3							
+g(−)	9	−61.6 ± 6.4	[−73.5 : −54.2]		36.5 ± 23.4	[15.8 : 74.8]	
+t	88	175.0 ± 6.2	[−163.6 : 167.3]		87.1 ± 16.6	[7.3 : 102.4]	
+g(+)	20	58.7 ± 9.8	[44.8 : 69.9]		37.4 ± 30.1	[8.8 : 82.9]	
−g(−)	4	−66.2 ± 8.5	[−77.9 : −59.8]		−81.6 ± 26.6	[−104.8 : −43.6]	
−t	1	−167.1			−93.6		
−g(+)	14	58.7 ± 6.9	[53.1 : 71.0]		−35.2 ± 18.2	[−78.0 : −24.9]	

The ensemble average ± standard deviation [minimum : maximum] values for dihedral angles are reported

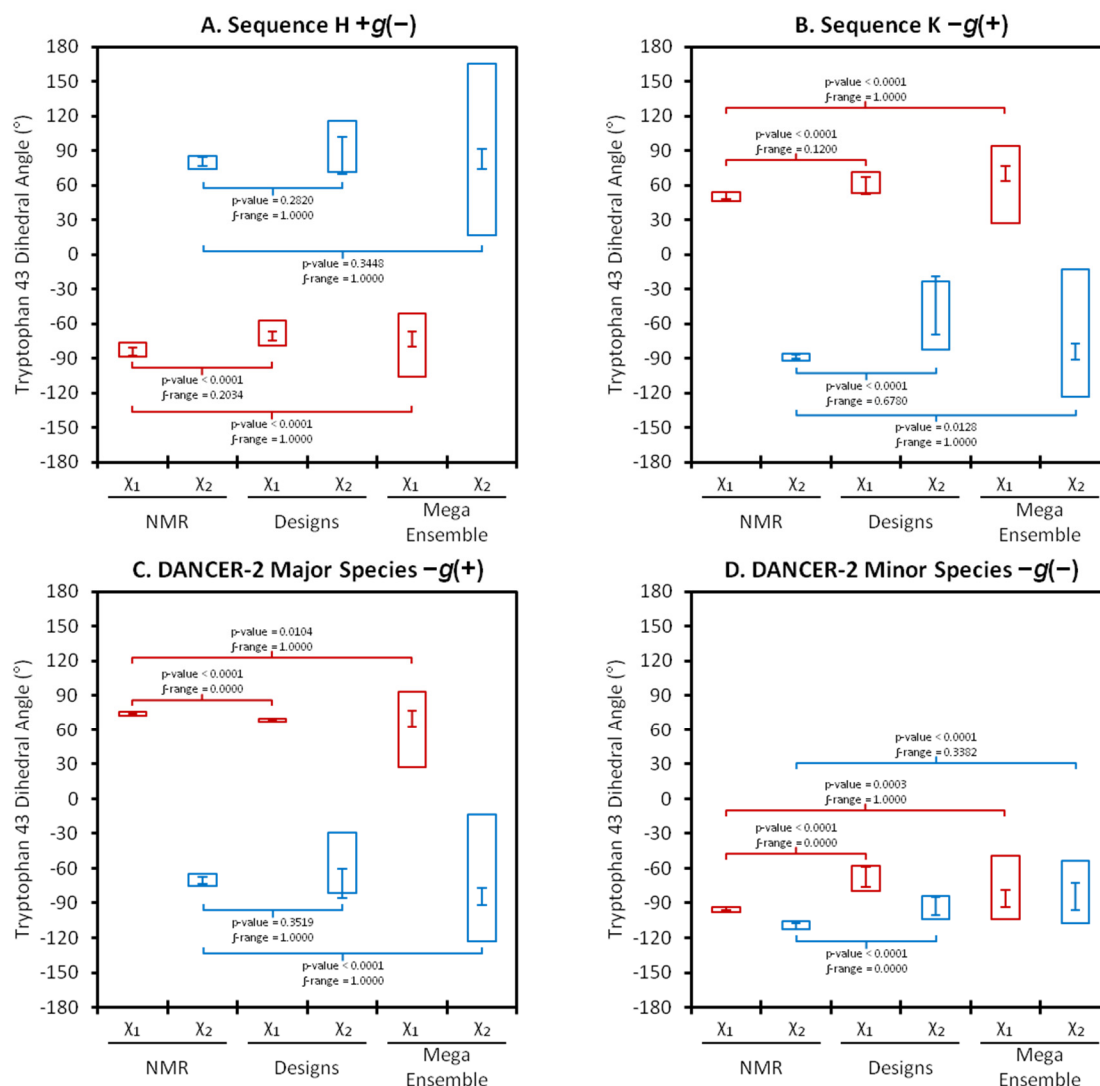


Figure 7.18. Comparative analysis between calculated and experimental tryptophan sub-state geometries. Ensemble tryptophan geometries from the solution NMR ensembles (NMR) are compared against those in the *meta*-MSD sub-states (Designs) and the set of all templates in the mega ensemble (Mega Ensembles). The population average, standard deviation (lines), and range (bar) are reported for χ_1 (red) and χ_2 (blue) dihedral angles. Results including the statistical analysis by an unpaired two-tailed T-test (p-value) and the shared dihedral space between the range of geometries occupied (f-range) between calculated and experimental dihedral angles are reported.

A comparative analysis [Figure 7.18] of the tryptophan conformations found in the solution NMR ensembles against those found in all 12,648 templates of the mega

ensemble provides the basis for a determination as to whether it is possible for the ensemble model space to allow for the prediction of the experimentally observed conformations. Further comparison of tryptophan conformations belonging to the solution NMR ensembles against the design sub-state conformations [Table 7.9] generated by *meta*-MSD identifies whether the stable rotameric states used to define the dynamic energy profile are similar. Two metrics are evaluated throughout this analysis: the first metric is the p-value calculated for a two-tailed, unpaired, t-test, with values approaching 0.0 indicating statistically significant geometry differences between the compared tryptophan populations, and the second metric is the calculated area ($\int$ -range) of the solution NMR encompassed by the design sub-state and the mega ensemble tryptophan conformations, with values of 0.0 indicating no coverage.

Analysis of the solution NMR tryptophan populations reveals that the majority of conformations occupy statistically significantly different (alpha-value < 0.1) χ_1 and χ_2 dihedral angle geometries when compared against the tryptophan populations belonging to the design sub-states and the mega ensemble across all three sequences (H, K, and the major and minor species of DANCER-2). The exceptions being the χ_2 angles for sequence H (p-value_{designs} = 0.2820 and p-value_{mega ensemble} = 0.3448) and the minor species of DANCER-2 (p-value_{mega ensemble} = 0.3382). This result strongly demonstrates that the population of tryptophan rotamers solved in the experimentally produced solution NMR ensembles vary considerably from the rotamer space of the Dunbrack library [Dunbrack and Cohen, 1997] and the mega ensemble produced from the application of Backrub motions [Davis *et al.*, 2006] and perturbed minimizations [Davey

and Chica, 2014] with the AMBER99 force field [Wang *et al.*, 2000]. Nonetheless, comparative analysis of the range of tryptophan conformations demonstrates that many of the solution NMR rotamers can be entirely found within the space of the design sub-states and mega ensemble ($\mathcal{J}$ -range = 1.0). Tryptophan populations that occupy an entirely distinct χ_1 or χ_2 dihedral angle space ($\mathcal{J}$ -range = 0.0) from the solution NMR ensembles include the rotamer populations found in DANCER-2 design sub-states for both the major and minor species. These results suggest that although the *meta*-MSD framework can correctly partition and evaluate state conformations of the tryptophan 43 residue, the accurate placement of its sidechain geometry with respect to experimental solution NMR ensembles could not be accomplished.

One contributing factor resulting in the inaccurate side chain placement of the tryptophan 43 residue could be that the backbone templates belonging to the mega ensemble do not resemble the local structure space of the DANCER proteins. To examine this possibility, the backbone and heavy atom coordinate RMSD between the solution NMR structures of sequences H, K, and the major and minor species of DANCER-2 and the minimum energy model, set of designed stable models, and mega ensemble design sub-states [Table 7.10]. Across the set of compared structures and sequences, backbone RMSD values range from 0.93 to 2.66 Å while the heavy atom RMSD values range from 1.69 to 3.51 Å. When compared against RMSD values between the set of solution NMR sequences and the 1PGA [Gallagher *et al.*, 1994] crystallographic structure (RMSD_{backbone} values of 1.13 to 2.13 Å, RMSD_{heavy atom} values of 1.97 to 3.06 Å) and 1GB1 [Gronenborn *et al.*, 1991] NMR average structure

(RMSD_{backbone} values of 1.39 to 2.28 Å, RMSD_{heavy atom} values of 2.06 to 3.13 Å), the RMSD values of the design models are as similar [Table 7.7]. Although all three sequences adopt the Gβ1 fold, considerable observable structural differences exist [Figure 7.14 and Figure 7.16].

Table 7.10. RMSD coordinate comparisons of solution NMR ensembles against sub-state design models calculated by *meta*-MSD

Sequence	Design Model			
	Min. Model	Designs Models		All Models
Backbone RMSD (Å)				
H	1.26	1.26 ± 0.07	[1.17 : 1.49]	1.26 ± 0.09 [0.93 : 1.90]
K	1.92	1.93 ± 0.06	[1.85 : 2.07]	1.97 ± 0.10 [1.67 : 2.66]
Major species DANCER-2	2.12	2.13 ± 0.08	[1.83 : 2.21]	2.01 ± 0.12 [1.57 : 2.53]
Minor species DANCER-2	2.26	2.26 ± 0.04	[2.12 : 2.32]	2.08 ± 0.12 [1.58 : 2.66]
Heavy atom RMSD (Å)				
H	2.01	1.99 ± 0.03	[1.92 : 2.06]	2.01 ± 0.11 [1.69 : 2.62]
K	2.89	2.84 ± 0.08	[2.71 : 2.96]	2.87 ± 0.11 [2.57 : 3.47]
Major species DANCER-2	3.34	3.30 ± 0.11	[2.90 : 3.39]	3.02 ± 0.15 [2.47 : 3.45]
Minor species DANCER-2	3.23	3.22 ± 0.07	[3.01 : 3.38]	3.01 ± 0.13 [2.56 : 3.51]

The ensemble average ± standard deviation [minimum : maximum] values for RMSD values are reported

Examination of the differences between solution NMR structures for sequences H, K, and DANCER-2 against the set of mega ensemble and design model structures has demonstrated the extent of accuracy afforded by the *meta*-MSD framework. While the prediction of dynamic from non-dynamic sequences was accomplished, the detailed

atomic modelling of the test set proteins demonstrates several limitations to the methodology. These limitations are present at two principle steps in the calculation procedure: the first is the backbone template generation procedure, and the second is the evaluation of design sub-states. The purpose of backbone ensembles and their use in MSD is that they should afford improved calculation accuracy over traditional single-state methodologies. However the creation of adequate backbone templates enabling the evaluation of any sequence in the context of the folded state remains a persistent challenge [Davey and Chica, 2014].

The fact that the DANCER proteins incorporate mutations to approximately 10% of their sequence space exacerbates this issue. The second challenge involving the evaluation of design sub-states arises as a result of the potential energy model employed during sequence optimization. The set of tryptophan residue states considered occupy a variety of configurations having variable solvent accessible surface areas. The incorporation and consideration of solvent interactions is complicated by poorly predictive implicit models [Lazaridis and Karplus, 1999; Street and Mayo, 1998] and the result is that particular structural features, including the compactness of ensembles [Davey *et al.*, 2015] are unphysically favored. The ability to correctly model solvent interactions will be critical to future rational design of protein dynamics as fast solvent motions are strongly coupled with both slow and fast protein motions [Lewandowski *et al.*, 2015]. Despite the shortcomings of the structural models, this work represents the first successful rational design of protein dynamics.

7.8. Implications of Mega Ensemble Model Space on *meta*-Multistate Design Prediction Outcomes

The ability of the *meta*-MSD framework to correctly predict the DANCER-1, DANCER-2, and DANCER-3 dynamic G β 1 proteins is dependent on the structure space and quality of the backbone templates used to model design sub-states during calculation. Alternative CPD frameworks such as the traditional SSD approach are unlikely to prove suitable for the task because they do not consider the stability of sequences across an ensemble of templates. The idea that the designed sub-states can be correctly partitioned and predict the relative stabilities of states across the energy landscape of the searched sequence space is central to the method. Thus, it is important to identify which templates found in the mega ensemble enabled the prediction of dynamic sequences. By restricting the *meta*-MSD calculation to subsets of the mega ensemble, their contribution to predictions can be dissected from the model without altering the framework. By restricting the calculation to ensemble members of specific state conformations, it is possible to examine the prediction outcomes of smaller multistate ensembles, and determine whether the large structure space and calculation framework were necessary.

The energetic profile of the design states by MSA across the 195 stable sequences was conducted using nine ensemble structure spaces of varying size [Figure 7.19]. The first three ensemble structure spaces, referred to as inter-state ensembles, are assembled according to their ensemble generation strategy [Figure 7.6] from the set of all eight seed structures [Table 7.1]. Specifically, the Seeds inter-state ensemble is an

ensemble space consisting of the eight minimized seed structures [Figure 7.19: Seeds], the PertMin inter-state ensemble consists of 248 backbone templates collected from the Seeds ensemble and the set of 30 PertMin structures generated from each of the eight seed structures [Figure 7.19: PertMin], and the Backrub inter-state ensemble with 408 members consists of the eight structures of the Seeds ensemble and the set of 50 energy minimized Backrub structures generated from each of the seeds structures [Figure 7.19: Backrub]. The importance of the step-wise ensemble procedure, specifically the three methodologies used here, can be determined by comparing MSA results across these inter-state ensembles with the MSA across the total structure space.

The other six ensemble structure spaces, referred to as intra-state ensembles, are collected from templates across the total structure space regardless of the origin of their ensemble generation procedure. Instead, all minimized seed, Backrub, and PertMin structures are grouped into their respective states based on the orientation of the tryptophan 43 residue present prior to *meta*-MSD. The result is six intra-state ensembles named according to their tryptophan state geometry: $+g(-)$ with 6073 structures, $+t$ with 32 structures, $+g(+)$ with 1458 structures, $-g(-)$ with 341 structures, $-t$ with 94 structures, and $-g(+)$ with 4650 structures.

Three metrics are reported for MSA with each inter- and intra-state ensemble. The first is the sequence count (n), detailing the number of the 195 sequences that can be evaluated as stable ($E < 0 \text{ kcal}\cdot\text{mol}^{-1}$) in each state. The second, sequence energy RMSD

(kcal·mol⁻¹), evaluates the state root-mean-square increase in energy per sequence evaluated as stable by each ensemble relative to the *meta*-MSD values.

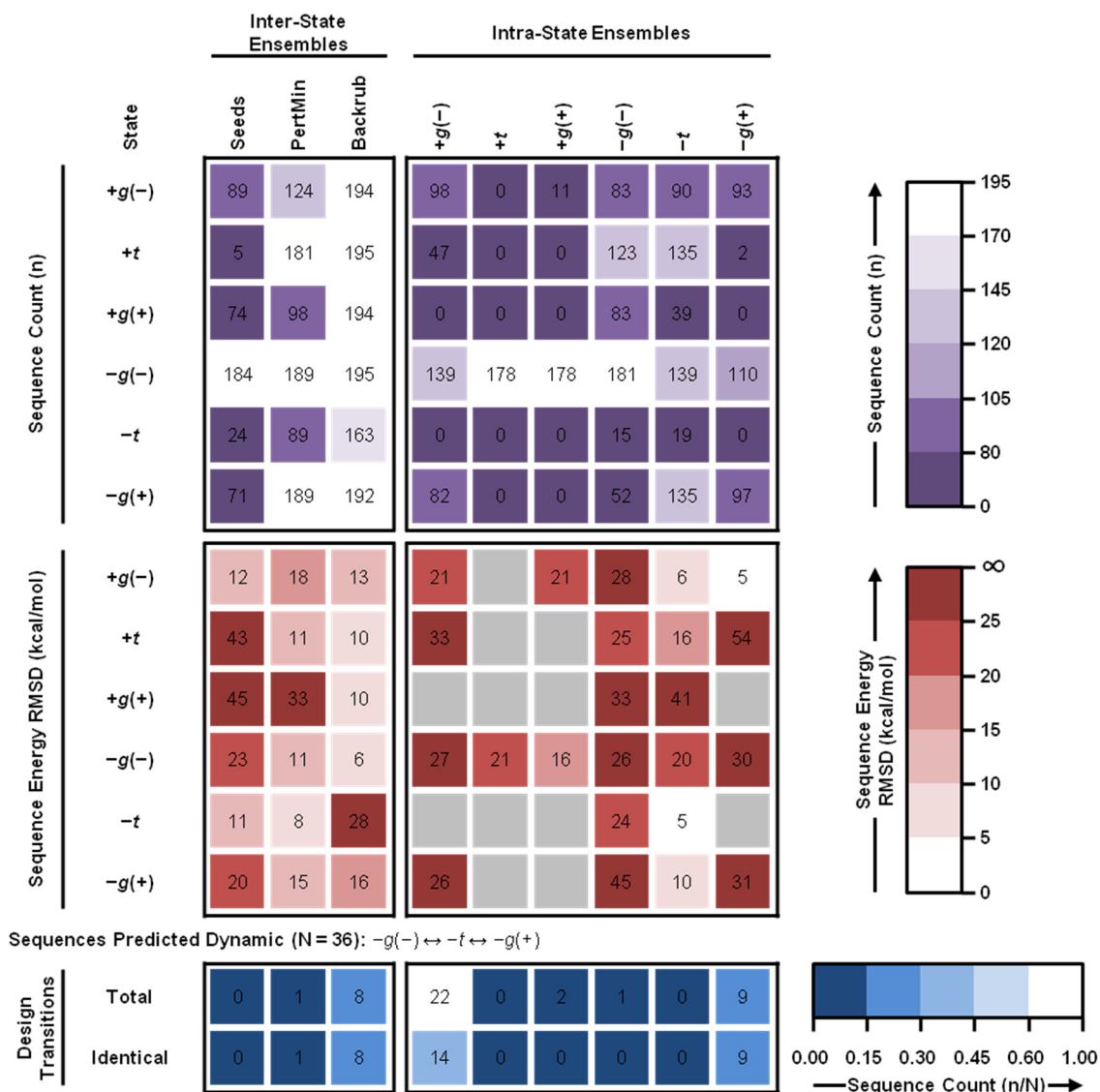


Figure 7.19. Ensemble structure space and *meta*-MSD predictions. Predictions by *meta*-MSD across the set of 195 predicted stable sequences, using subsets of structures collected from the mega ensemble and grouped into respective inter- and intra-state ensembles. Evaluated metrics include a count of state conformations that can be evaluated as stable ($E < 0$ kcal·mol⁻¹), sequence state RMSD energy relative to their energy evaluated using the set of mega ensemble structures, and the number of

sequences whose energy profiles correspond to the objective dynamic exchange, separated into any predicted dynamic sequence (total) and those 36 identified by the mega ensemble to be dynamic (identical). Metrics for predictions from *meta*-MSD using the inter- and intra-state ensembles are colored dark if the properties deviate substantially from those of the mega ensemble.

The last metric, the number of sequences predicted dynamic, counts the number of sequences (n) that satisfy the objective dynamic exchange $-g(-) \leftrightarrow -t \leftrightarrow -g(+)$. The results of *meta*-MSD using the restricted set inter- and intra-state ensembles is reported with two metrics, total and identical. The first metric (total) reports on the total number of sequences predicted to be dynamic independent of their predicted dynamic assignment provided by *meta*-MSD using the total ensemble structure space. The second metric (identical) reports on the number of sequences predicted to be dynamic that are common to those predicted by *meta*-MSD ($N = 36$) using the total ensemble structure space.

The analysis shown in Figure 7.19 demonstrates that not a single inter- or intra-state ensemble can adequately approximate results obtained using *meta*-MSD across the mega ensemble. Comparing the number of sequences that can be stably evaluated in the context of all states allows for a quantitative assessment of the ensemble's utility at evaluating the potential energy landscape. All inter-state ensembles, regardless of their size or the methodology used to create them, can evaluate more of the conformational landscape than most of the intra-state ensembles. This result highlights the importance of the seed template generation step [Davey and Chica, 2015] toward enabling the evaluation of the set of state conformations. Conversely, many of the state

conformations cannot be evaluated for any of the 195 sequences using the intra-state ensembles.

Examining the state sequence energy RMSD of the ensemble predictions allows for an assessment of their ability to quantitatively evaluate the quality of the sub-state designs. Regardless of the ensemble used, all evaluate sequence state energies between 5 to 45 kcal·mol⁻¹ higher than the mega ensemble. Importantly, even if one state can be evaluated with similar energetics with respect to the mega ensemble, for example the evaluation of the +g(-) state using the -g(+) intra-state ensemble (RMSD_E = 5 kcal·mol⁻¹), the remaining states cannot be reasonably evaluated (RMSD_E > 30 kcal·mol⁻¹). This result demonstrates that no one restricted ensemble space would allow for a reliable evaluation of the relative energy between states for any given sequence.

To conclusively demonstrate that the restricted structure space of inter-state and intra-state ensembles could not be used to successfully predict protein dynamics, the energy profiles of the states were parsed to identify sequences exhibiting the required energy profile to be classified as dynamic. Of the 195 sequences, 36 were predicted by *meta*-MSD using the mega ensemble to be dynamic (N). The majority of inter- and intra-state ensembles could not predict any sequence to be dynamic (n_{total} = 0). Those that did predict sequences to be dynamic identify less than half of the 36 predicted sequences (n_{identical}). For example, the intra-state ensemble comprised of templates with the +g(-) tryptophan conformation predicted that 22 sequences (n_{total}) satisfied the required dynamic profile while only 14 of these same sequences were predicted by *meta*-MSD using the total ensemble structure space (n_{identical}). If the assignment of sequence

dynamics by *meta*-MSD using the total ensemble structure space is to be considered correct, then this result indicates that *meta*-MSD using the $+g(-)$ intra-state ensemble includes 8 false positives (false positives = $n_{\text{total}} - n_{\text{identical}}$) and 22 false negatives (false negatives = $N - n_{\text{identical}}$). These results demonstrate that step-wise ensemble generation protocol was essential to the success of the *meta*-MSD predictions, and that it is unlikely that SSD using a single backbone template, or MSD using a reduced number of backbone templates, could successfully predict G β 1 dynamics.

7.9. Conclusions on the Application of *meta*-Multistate Design

The dynamic behaviours of proteins pose a series of central and fundamental questions to physical biology. Unfortunately, the current understanding of the origin and functional implication of protein dynamics has been limited to the characterization of naturally occurring proteins and a small number of their mutant sequences. With this in mind, I sought to develop a framework for the rational design of protein dynamics into the static scaffold of G β 1. The eventual goal and utility of such a strategy would enable the design of sequence specific perturbations inducing controlled dynamic motions in any given protein fold.

In this chapter, I report on what is to the best of my knowledge, the first successful protein engineering effort to rationally design protein sequences exhibiting a specific dynamic exchange into a globular protein fold. The computational design of protein dynamics required the development of a new multistate framework, referred to as

meta-multistate design (*meta*-MSD), which builds upon current MSD methodologies by considering assignment of design state post calculation by means of a MSA. Use of this *meta*-MSD approach, enabled the *in silico* search for dynamic sequences across a limited sequence space (1,296 total sequences) across a vast ensemble space (12,648 total structures) using the G β 1 scaffold. Sequences were searched specifically for those having energetic profiles corresponding to a previously unobserved exchange of the tryptophan 43 sidechain between two non-native conformations. The result was the creation of four new proteins referred to as DANCER-0, DANCER-1, DANCER-2, and DANCER-3, named for their behaviour as Dynamic Native Conformation ExchangeR (DANCER) proteins. The four DANCER proteins were predicted to exist in equilibrium and exchange between the two non-native conformations of Trp43 with various barrier heights (7-12 kcal·mol⁻¹). Study by ¹⁵N-HSQC and ZZ-exchange NMR spectroscopy demonstrated that DANCER-1, DANCER-2, and DANCER-3 proteins were dynamic, with DANCER-1 and DANCER-3 exchanging between the two conformations on the 10 to 100 millisecond time scales, respectively. NMR solution structures of a dynamic sequence and additional non-dynamic G β 1 variants suggest good agreement between calculation and experiment.

The ability to rationally engineer dynamic motions at will into any protein scaffold and to reliably predict and control the dynamic behaviour of existing motions promises to enable its consideration in the future study and design of complex protein functions, such as catalysis and association, with atomic level detail and energetic precision.

CHAPTER 8

Conclusions

Having introduced the topic of CPD and its respective single-state and multistate theoretical frameworks, a pair of backbone template preparation methodologies were developed and discussed. The first methodology, referred to as the PertMin ensemble generation procedure, was evaluated using MSD calculations involving the prediction and recapitulation of G β 1 WT and mutant sequence stabilities. These calculations were benchmarked by comparison against MSD calculations conducted using various alternate ensembles and SSD. The second methodology, referred to as the ROM procedure, allowed for the production of backbone templates biasing the favorably predicted sequence space of SSD calculations. The PertMin methodology was applied to the positive MSD of enzyme-substrate multi-specificity and the negative MSD of protein stability, while the ROM and PertMin methodologies were applied using a new CPD framework, referred to as *meta*-MSD, to the design of non-native protein dynamics into the G β 1 scaffold. In the following chapter, I will discuss, criticize, and summarize the key findings of the application of these multistate frameworks, the methods used to generate the ensemble spaces that were considered, and elaborate on the theory and future of MSD.

8.1. Applications

A central theme is evident when comparing the quality of predictions afforded by CPD using either the SSD or MSD strategies. This theme, that the quality of predictions made using MSD consistently outperform those predictions made using SSD methodologies, is observed for each of the three applied computational design projects reported in this thesis.

In the applied positive design of SMYD2 substrate specificity we were able to conclude, in a collaboration with the *Couture* laboratory, that SMYD2 methylates the sequence recognition motif [LFM]₋₁-K*-[AFYMSHRK]₊₁-[LYK]₊₂ with position -1 and +2 exhibiting the highest degree of selectivity. Use of the defined recognition motif allowed for the application of a combination of different bioinformatics and biochemical approaches to identify four previously unidentified substrates of SMYD2. A comparison of the high throughput SPOT assay showed that positive MSD using a PertMin backbone ensemble could correctly predict the substrate sequence space recognized by SMYD2. Conversely, when the same prediction was attempted using SSD, the recognition motif was poorly recapitulated. This was especially true for SSD calculations attempting to recapitulate the broad specificity observed at the +1 position reported for the SPOT peptide array recognition motif.

In the applied negative design of protein stability a similar trend is also observed, except that in this specific project three calculation frameworks, SSD, positive MSD, and negative MSD, were examined. The project involved the recapitulation of a training set

of Gβ1 mutant sequence stabilities by various CPD calculations. The use of the training set provided a means to inform and identify which structural models and calculation frameworks gave the best agreement between calculation and experiment. The best correlations between calculation and experiment were obtained using negative MSD employing ensemble models representing both the native and non-native protein fold. The quality of the stability correlations by negative MSD and positive MSD were shown to be variable based on the ensemble model used in each calculation. It was determined that when calculating training set sequence stabilities by positive MSD alone, the best ensemble models were those corresponding to the native protein fold. However, regardless of the template or ensemble model used SSD or positive MSD, negative MSD with the appropriate ensemble states provided the best correlations.

These results are consistent with the notion that improved modelling of the structure space and incorporation of additional protein states, improves upon the accuracy of predictions made by conventional SSD calculations. However, the results highlight an important caveat – that the identification of predictive and physical native and non-native models for use in multiple conformation MSD remains a challenge. Nonetheless, the protocol of employing a training set of sequences to inform not just the evaluation of sequence stability but also the choice and behaviour of ensemble models should aid future protein engineering projects that attempt to consider protein properties and functions for which limited structural information is available for.

The last applied protein engineering project presented in this thesis involved the design of Non-native And Conformational ExchangeR, or DANCER, proteins from the

G β 1 scaffold. In this protein dynamics project, a mega ensemble was constructed to facilitate the evaluation of protein sequences across alternate folded states. Sequences were searched and ranked by their ability to stabilize subsets of structure space corresponding to the conformations of the desired dynamic motion. The challenge associated with the design of the dynamic motion is that the energetic profile of each sequence must be evaluated across a dynamic transition that has not yet been experimentally observed. In this way, the design of protein dynamics shares a common challenge with the prediction of protein stability using negative design, as both implementations of MSD require structural models for states that have not been experimentally observed. The difference here is that the production and selection of ensemble model states cannot be guided using a training set of dynamic sequences with the specific conformational exchange as none are currently known to exist.

Instead, the ensemble structure space originating from the crystallographic structure of the G β 1 fold must be manipulated to represent the intended set of configurations as best as possible. To accomplish this, the ROM template and PertMin ensemble generation strategies were employed to provide a mega ensemble capable of portioning and evaluating design conformations. The major difference with the positive and negative MSD frameworks employed previously, and the multistate framework employed in the design of dynamics, is that the assignment of state for each designed structure was determined post MSD. This *meta*-MSD framework allowed for the independent prediction of protein stability, equilibrium, and activation barrier heights via a MSA across sequence designs. Here, MSD is employed to optimize rotamer

configurations with the objective of stabilizing their respective templates while the objective design function, the equilibrium energy and activation barrier, are effectively ignored. While this approach has the benefit avoiding the complications of *Pareto front* optimizations [Grigoryan *et al.*, 2009; Suarez *et al.*, 2008] Grigoryan, the search of the sequence space does not guarantee the solution of a dynamic sequence, as the framework does not actively consider the objective during sequence optimization.

Nonetheless, use of the mega ensemble space did prove essential to the prediction of the DANCER proteins. A key finding was that reduction of the ensemble structure space considered in MSD calculation substantially hindered the ability of *meta*-MSD to identify sequences that could populate and adopt the energetic profile of DANCER proteins. These results are again consistent with the statement that improved modelling of the structure space and the incorporation of additional protein states, improves upon the accuracy of predictions made by conventional SSD calculations.

Future protein engineering and rational design efforts will likely push the boundaries of accurate protein property predictions to encompass all possible conformational and chemical design elements into a single, unified, multistate calculation. To assemble this sort of calculation framework substantially more information will be required detailing the protein sequence and structure relationship. Because many protein conformational and chemical states have not been observed, there is a need to construct MSD protocols that incorporate multiple elements of training set data to inform both the evaluation of sequences and the modelling of structure space. The design of protein stability for example could incorporate additional

training set features in a regression protocol similar to that employed in our negative MSD of protein stability. For example, properties related to the ensemble model solvent accessible surface area, such as a sequence's change in heat capacity upon thermal denaturation or the cooperativity of chemical denaturant binding during unfolding [Myers *et al.*, 1995], could be incorporated alongside stability metrics during the evaluation of ensemble model pairs. By evaluating a pool of positive and negative state ensemble candidates in the context of their ability to predict both the energetic stability of a fold and a structural property of the fold, a more physical and reliable pair of ensembles may be identified. For the time being, however, these three positive, negative and *meta*-MSD projects represent the first examples of their application to the design of substrate discovery, stability prediction, and the design of protein dynamics.

8.2. Methods

In this thesis, two methods were developed and tested for the purposes of ensemble and backbone template generation for use in CPD. The ensemble method PertMin, involving coordinate perturbations followed by energy minimization, was developed and its use tested in the prediction of G β 1 sequence stabilities by positive MSD. The method allowed for the generation of native backbone ensembles that could preferentially enrich and bin MSD predictions for stable sequences (on-target). Use of the PertMin ensemble in MSD, along with conventional ensemble generation strategies

employing MD and Backrub simulation ensembles, resulted in improved accuracy of protein stability predictions compared to SSD with the crystal structure.

While the PertMin ensemble generation protocol is easy to implement, the diversity of the ensemble members produced tend to be tightly restricted about the geometry of the input backbone template. The generation of useful and predictive ensemble structures is an important requirement for a MSD calculation to be successful. While PertMin has the benefit of creating precise and narrow structural diversity about an input template, if the original starting structure is unsuitable for CPD predictions, it is not likely that a structure generated from it using PertMin would result in a predictive ensemble. For example, while the crystal structure of G β 1 allowed of the creation of a PertMin ensemble capable of enriching and binning sequence stability predictions with a majority of stable sequences, a small number of incorrectly ranked stable sequences were consistently rejected (false negative prediction). This issue of false negative predictions demonstrates that even with detailed available structural information, the evaluation of sequence perturbations to folded state is difficult. A similar result was observed for specific substitutions to the SMYD2 recognition motif. To circumvent PertMin's limitation of low diversity structure generation, the method can be applied in series after an ensemble generation strategy of high energy and high diversity. Such a strategy was employed with success for the design of protein dynamics presented in this thesis. By using the Backrub ensemble generation protocol, and subjecting each resulting Backrub structure to PertMin, a highly diverse but on-target low energy mega ensemble could be constructed.

The second methodology developed in this thesis, referred to as ROM, exploits the bias that the geometry of the template enforces on a design calculation by evaluating sequence stabilities in the context manipulated templates. Templates prepared with ROM, involving the rotamer optimization and energy minimization of alternative sequences onto a backbone template, were used in SSD to improve or detract from calculation accuracies depending on the sequence present during the ROM procedure. The method effectively demonstrated the existence of a bias in CPD that arises from variations in side chain contacts made by rotamers in the input template used. This rotamer bias causes CPD calculations to favor rotamer configurations that are similar to those present when the template was prepared.

Because CPD calculations exhibit this rotamer bias, the effect can be exploited to manipulate template models and improve the accuracy of protein stability predictions. And, because implementation of the ROM procedure requires only standard single-state CPD protocols and energy minimization, it is easily implemented. In an ideal design circumstance, prior knowledge of the sequence space could accompany rational design efforts through the use of ROM templates that were biased for a specific property or feature of a desired sequence. For example, if SSD using a specific backbone template cannot successfully enrich or bin its sequence predictions for desired sequences, the template can be manipulated provided that the identity of at least one sequence satisfying the design objective exists.

The ROM method could also potentially be applied to iterative protein design strategies. Traditional iterative protein design strategies involve a cycle of three steps

where sequences are first predicted, then experimentally characterized, and the information is applied to refine modeling and design predictions [Privett *et al.*, 2012]. The last step in this iterative procedure is time consuming, requires substantial analysis, and does not ensure that changes to the model will result in an improvement to CPD predictions. The ROM template generation procedure could be applied in place of the model refinement step, whereby successive generations of predictions are guided by the sequence identified as the best candidate from the previous step. Such a procedure could even be automated, provided that at least one suitable sequence can be found in the first round of the procedure, and those successive generations of the procedure produce different top ranked sequences.

8.3. Theories

In general, the application of the positive, negative, and *meta*-multistate design calculation frameworks to the prediction of multi-specificity, stability, and protein dynamics, improved upon the accuracy of predictions made using the traditional SSD calculation framework. For example, when attempting to predict the stabilities of 18 G β 1 mutant sequences in chapter six, calculations performed using SSD, positive MSD, and negative MSD using non-native backbone ensembles gave correlations (R^2) to experimental stability values of 0.12, 0.42, and 0.82, respectively. The ability to successively improve calculation accuracy by expanding on the size and quality of structure space models presents a promising avenue toward the development of improve calculation strategies.

From a practical standpoint, calculations must be limited to structure space models based on experimental data, or previously validated proxies of structure space, such as the non-native G β 1 models tested in chapter six. While calculation models that compare computational and experimental data allow us to test our knowledge of protein folding mechanics and to engineer useful proteins, there is an inherent limitation to the methodology. This limitation arises due to the fact that conclusions made by probing protein behaviours through sequence perturbations can be biased based on the protein systems studied.

Although the MSD applications presented in this thesis were performed for many protein systems, including: G β 1, SMYD2, *Staphylococcus aureus* nuclease, bacteriophage f1 gene V protein, bacteriophage T4 lysozyme, and barley seed chymotrypsin inhibitor 2; the methodological development presented in chapters three and four of this thesis were benchmarked using a small data set of 85 sequence variants for a single protein fold, G β 1. This bias is particularly evident, for example, when attempting to identify which of the ensembles evaluated in chapter three were on- or off-target with respect to the G β 1 protein fold. The conclusion as to which ensemble features were important for successful MSD was based on the ability of each ensemble to recapitulate sequence stabilities for desired members of the data set. And, while it is true that particular ensemble features, such as low deviation from the original crystallographic structure, can suggest as to whether a particular ensemble can successfully enrich predictions, it is also possible that alternative ensemble structures, not adhering to the parameters of an on-target on ensemble, might also predict for desirable sequences not considered in the

data set. This is particularly evident when considering data presented in chapter four, demonstrating that CPD exhibits considerable difficulty at attempting to identify desirable sequences that diverge significantly from the WT sequence of the model template.

Thus, it is important to expand upon the areas of structure space that are required for calculation of various sequences having different behaviours and properties. To accomplish this, a multistate calculation could be performed across a vast sequence space data set while considering all possible chemical and conformational states. Templates approximating an infinite ensemble structure space would then be divided into subsets of structures in order to identify those critical to aspects of prediction accuracy. By dissecting the useful and predictive states from the redundant or non-predictive structures, a more complete understanding of protein design could be realized. The ideal calculation would investigate prediction outcomes as a function of all possible ensemble configurations.

8.4. Summary

The set of CPD calculations in this thesis serve as additional examples on how to apply and implement a multistate framework in the design of complex protein properties and functions. Importantly, the methodology and the success of these calculations relied on the generation of suitable ensemble states capable of representing these protein properties and behaviours. The goal of this thesis was to improve upon the understanding of the relationship between the ensembles used in

MSD and their utility to predict protein properties. And despite the success of the methods and applications presented, considerable challenges convolute the long term objective of CPD to engineer any property, into any protein, at will, with perfect accuracy.

Previously, I discussed that the ideal MSD calculation would incorporate all possible chemical and conformational states allowing for the consideration of any protein property or function in calculation. However, because the landscape of a protein is conformationally vast, such an ensemble cannot be considered. In this thesis, I explored alternative ensemble representations, and tested their application in the design of multi-specificity, sequence stability, and dynamics. These studies demonstrated that by comparing design predictions with experiment, and using those comparisons to guide our ensemble structure generation strategies, we were able to predict increasingly complex protein properties. By tackling the design of alternative protein properties and functions in this manner, new previously unattainable protein design goals can be realized.

REFERNECES

Abu-Farha M, Lambert JP, Al-Madhoun AS, Elisma F, Skerjanc IS and Figeys D (2008) The tale of two domains - proteomics and genomics analysis of smyd2, a new histone methyltransferase. *Molecular & Cellular Proteomics*, 7(3):560-572.

Abu-Farha M, Lanouette S, Elisma F, Tremblay V, Butson J, Figeys D and Couture JF (2011) Proteomic analyses of the smyd family interactomes identify hsp90 as a novel target for smyd2. *Journal of Molecular Cell Biology*, 3(5):301-308.

Ali MH, Taylor CM, Grigoryan G, Allen KN, Imperiali B and Keating AE (2005) Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. *Structure*, 13(2):225-234.

Alland L, David G, Hong SL, Potes J, Muhle R, Lee HC, Hou H, Chen K and DePinho RA (2002) Identification of mammalian sds3 as an integral component of the sin3/histone deacetylase corepressor complex. *Molecular and Cellular Biology*, 22(8):2743-2750.

Allen BD and Mayo SL (2006) Dramatic performance enhancements for the faster optimization algorithm. *Journal of Computational Chemistry*, 27(10):1071-1075.

Allen BD and Mayo SL (2010) An efficient algorithm for multistate protein design based on faster. *Journal of Computational Chemistry*, 31(5):904-916.

Allen BD, Nisthal A and Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46):19838-19843.

Allert M, Rizk SS, Looger LL and Hellinga HW (2004) Computational design of receptors for an organophosphate surrogate of the nerve agent soman. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):7907-7912.

Alvizo O, Mittal S, Mayo SL and Schiffer CA (2012) Structural, kinetic, and thermodynamic studies of specificity designed hiv-1 protease. *Protein Science*, 21(7):1029-1041.

Ambroggio XI and Kuhlman B (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society*, 128(4):1154-1161.

Andersen CAF, Palmer AG, Brunak S and Rost B (2002) Continuum secondary structure captures protein flexibility. *Structure*, 10(2):175-184.

Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Stoddard BL and Baker D (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656-659.

Ayer DE, Lawrence QA and Eisenman RN (1995) Mad-max transcriptional repression is mediated by ternary complex-formation with mammalian homologs of yeast repressor sin3. *Cell*, 80(5):767-776.

Babor M, Mandell DJ and Kortemme T (2011) Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody herceptin-her2 interface. *Protein Science*, 20(6):1082-1089.

Bajaj M and Blundell T (1984) Evolution and the tertiary structure of proteins. *Annual Review of Biophysics and Bioengineering*, 13(1):453-492.

Barros MC, Katayama MLH, Brentani H, Abreu APS, Barbosa EM, Oliveira CT, Goes JCS, Brentani MM and Folgueira MAAK (2010) Gene trio signatures as molecular markers to predict response to doxorubicin cyclophosphamide neoadjuvant chemotherapy in breast cancer patients. *Brazilian Journal of Medical and Biological Research*, 43(12):1225-1231.

Bhabha G, Lee J, Ekiert DC, Gam J, Wilson IA, Dyson HJ, Benkovic SJ and Wright PE (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, 332(6026):234-238.

Blais A, Tsikitis M, Acosta-Alvear D, Sharan R, Kluger Y and Dynlacht BD (2005) An initial blueprint for myogenic differentiation. *Genes & Development*, 19(5):553-569.

Boas FE and Harbury PB (2008) Design of protein-ligand binding based on the molecular-mechanics energy model. *Journal of Molecular Biology*, 380(2):415-424.

Boas FE and Harbury PB (2007) Potential energy functions for protein design. *Current Opinion in Structural Biology*, 17(2):199-204.

Bolon DN, Grant RA, Baker TA and Sauer RT (2005) Specificity versus stability in computational protein design. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12724-12729.

Bolon DN and Mayo SL (2001) Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America*, 98(25):14274-14279.

Bond SD, Leimkuhler BJ and Laird BB (1999) The nose-poincare method for constant temperature molecular dynamics. *Journal of Computational Physics*, 151(1):114-134.

Borgo B and Havranek JJ (2012) Automated selection of stabilizing mutations in designed and natural proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 109(5):1494-1499.

Boucher CA, Winchester CL, Hamilton GM, Winter AD, Johnson KJ and Bailey MES (2000) Structure, mapping and expression of the human gene encoding the homeodomain protein, six2. *Gene*, 247(1-2):145-151.

Bouvignies G, Vallurupalli P, Hansen DF, Correia BE, Lange O, Bah A, Vernon RM, Dahlquist FW, Baker D and Kay LE (2011) Solution structure of a minor and transiently formed state of a t4 lysozyme mutant. *Nature*, 477(7362):111-134.

Brown MA, Sims RJ, Gottlieb PD and Tucker PW (2006) Identification and characterization of smyd2: A split set/mynd domain-containing histone h3 lysine 36-specific methyltransferase that interacts with the sin3 histone deacetylase complex. *Molecular Cancer*, 5(1):26-36.

Busch MSA, Lopes A, Mignon D and Simonson T (2008) Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *Journal of Computational Chemistry*, 29(7):1092-1102.

Butterworth S, Lamzin VL, Wigley DB, Derrick JP and Wilson KS (1997) Anisotropic structure of protein g igg-binding domain iii at 1.1 angstrom resolution. *Structure Deposited to the Protein Data Bank*, PDBID: 2IGD.

Calhoun JR, Kono H, Lahr S, Wang W, DeGrado WF and Saven JG (2003) Computational design and characterization of a monomeric helical dinuclear metalloprotein. *Journal of Molecular Biology*, 334(5):1101-1115.

Chandler D (2005) Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640-647.

Chen TS and Keating AE (2012) Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods. *Protein Science*, 21(7):949-963.

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS and Richardson DC (2010) Molprobity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(D):12-21.

Chica RA, Doucet N and Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: Combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology*, 16(4):378-384.

Chica RA, Moore MM, Allen BD and Mayo SL (2010) Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47):20257-20262.

Cho HS, Hayami S, Toyokawa G, Maejima K, Yamane Y, Suzuki T, Dohmae N, Kogure M, Kang D, Neal DE, Ponder BAJ, Yamaue H, Nakamura Y and Hamamoto R (2012) Rb1 methylation by smyd2 enhances cell cycle progression through an increase of rb1 phosphorylation. *Neoplasia*, 14(6):476-486.

Choi EJ, Guntas G and Kuhlman B (2009) Future challenges of computational protein design. 18:367-385. *"Protein engineering and design"*, Park SJ and Cochran JR. Boca Raton, Florida, United States of America.

Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV and Mann M (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325(5942):834-840.

Clayden J, Greeves N, Warren S and Wothers P (2007) The chemistry of life. 49:1353-1356. *"Organic chemistry"*, Clayden J, Greeves, N., Warren, S., Wothers, P. New York, New York, United States of America.

Coutsias EA, Seok C, Jacobson MP and Dill KA (2004) A kinematic view of loop closure. *Journal of Computational Chemistry*, 25(4):510-528.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004) Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188-1190.

Dahiyat BI and Mayo SL (1997) Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10172-10177.

Daniel RM, Smith JC, Ferrand M, Hery S, Dunn R and Finney JL (1998) Enzyme activity below the dynamical transition at 220 k. *Biophysical Journal*, 75(5):2504-2507.

Dantas G, Kuhlman B, Callender D, Wong M and Baker D (2003) A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, 332(2):449-460.

Davey JA (2011a) The addition of hydrogens and input perturbation sensitivity. MSc Chemistry Thesis: *"On the energy minimization of large molecules"*, 2(4):46-51. Carleton University, Ottawa, Ontario, Canada.

Davey JA (2011b) The minimization algorithm and input perturbation sensitivity. MSc Chemistry Thesis: *"On the energy minimization of large molecules"*, 2(6):54-57. Carleton University, Ottawa, Ontario, Canada.

Davey JA (2011c) Perturbation size and input perturbation sensitivity. MSc Chemistry Thesis: *"On the energy minimization of large molecules"*, 2(2):41-43. Carleton University, Ottawa, Ontario, Canada.

Davey JA and Chica RA (2014) Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins-Structure Function and Bioinformatics*, 82(5):771-784.

Davey JA and Chica RA (2012) Multistate approaches in computational protein design. *Protein Science*, 21(9):1241-1252.

Davey JA and Chica RA (2015) Optimization of rotamers prior to template minimization improves stability predictions made by computational protein design. *Protein Science*, 24(4):545-560.

Davey JA, Damry AM, Euler CK, Goto NK and Chica RA (2015) Prediction of stable globular proteins using negative design with non-native backbone ensembles. *Structure*, 23(11):2011-2021.

Davis IW, Arendall WB, Richardson DC and Richardson JS (2006) The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, 14(2):265-274.

de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A and Hulo N (2006) Scanprosite: Detection of prosite signature matches and prerule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(W):362-365.

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J and Bax A (1995) Nmrpipe - a multidimensional spectral processing system based on unix pipes. *Journal of Biomolecular Nmr*, 6(3):277-293.

Derrick JP and Wigley DB (1994) The third igg-binding domain from streptococcal protein-g - an analysis by x-ray crystallography of the structure alone and in a complex with fab. *Journal of Molecular Biology*, 243(5):906-918.

- Desjarlais JR and Handel TM (1999) Side-chain and backbone flexibility in protein core design. *Journal of Molecular Biology*, 290(1):305-318.
- Desmet J, Demaeyer M, Hazes B and Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539-542.
- Desmet J, Spriet J and Lasters I (2002) Fast and accurate side-chain topology and energy refinement (faster) as a new method for protein structure optimization. *Proteins-Structure Function and Genetics*, 48(1):31-43.
- Diehl F, Brown MA, van Amerongen MJ, Novoyatleva T, Wietelmann A, Harriss J, Ferrazzi F, Bottger T, Harvey RP, Tucker PW and Engel FB (2010) Cardiac deletion of smyd2 is dispensable for mouse heart development. *Plos One*, 5(3):e9748.
- Dill KA (1999) Polymer principles and protein folding. *Protein Science*, 8(6):1166-1180.
- Dill KA, Bromberg S, Yue KZ, Fiebig KM, Yee DP, Thomas PD and Chan HS (1995) Principles of protein-folding - a perspective from simple exact models. *Protein Science*, 4(4):561-602.
- Ding L, Chen K, Santini PA, Shi ZS and Kallenbach NR (2003) The pentapeptide ggagg has pii conformation. *Journal of the American Chemical Society*, 125(27):8092-8093.
- Donlin LT, Andresen C, Just S, Rudensky E, Pappas CT, Kruger M, Jacobs EY, Unger A, Zieseniss A, Dobenecker MW, Voelkel T, Chait BT, Gregorio CC, Rottbauer W, Tarakhovsky A and Linke WA (2012) Smyd2 controls cytoplasmic lysine methylation of hsp90 and myofilament organization. *Genes & Development*, 26(2):114-119.
- Doster W (2008) The dynamical transition of proteins, concepts and misconceptions. *European Biophysics Journal with Biophysics Letters*, 37(5):591-602.
- Doster W, Cusack S and Petry W (1989) Dynamical transition of myoglobin revealed by inelastic neutron-scattering. *Nature*, 337(6209):754-756.
- Dunbrack RL and Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, 6(8):1661-1681.
- Dunbrack RL and Karplus M (1993) Backbone-dependent rotamer library for proteins - application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543-574.
- Dyson HJ and Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197-208.

Fenimore PW, Frauenfelder H, Magazu S, McMahon BH, Mezei F, Migliardo F, Young RD and Stroe I (2013) Concepts and problems in protein dynamics. *Chemical Physics*, 424(2-6).

Ferguson AD, Larsen NA, Howard T, Pollard H, Green I, Grande C, Cheung T, Garcia-Arenas R, Cowen S, Wu JQ, Godin R, Chen HW and Keen N (2011) Structural basis of substrate methylation and inhibition of smy2. *Structure*, 19(9):1262-1273.

Frauenfelder H, Chen G, Berendzen J, Fenimore PW, Jansson H, McMahon BH, Stroe IR, Swenson J and Young RD (2009) A unified model of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5129-5134.

Frauenfelder H, Sligar SG and Wolynes PG (1991) The energy landscapes and motions of proteins. *Science*, 254(5038):1598-1603.

Frey KM, Georgiev I, Donald BR and Anderson AC (2010) Predicting resistance mutations using protein design algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13707-13712.

Friedland GD and Kortemme T (2010) Designing ensembles in conformational and sequence space to characterize and engineer proteins. *Current Opinion in Structural Biology*, 20(3):377-384.

Friedland GD, Linares AJ, Smith CA and Kortemme T (2008) A simple model of backbone flexibility improves modeling of side-chain conformational variability. *Journal of Molecular Biology*, 380(4):757-774.

Gaussian 09 revision e.01 (2009.09 Revision E.01) Gaussian, Inc. Wallingford, Connecticut, United States of America.

Gainza P, Nisonoff HM and Donald BR (2016) Algorithms for protein design. *Current Opinion in Structural Biology*, 39(April 2016):16-26.

Gainza P, Roberts KE and Donald BR (2012) Protein design using continuous rotamers. *Plos Computational Biology*, 8(1):e1002335.

Gallagher T, Alexander P, Bryan P and Gilliland GL (1994) Two crystal-structures of the b1 immunoglobulin-binding domain of streptococcal protein-g and comparison with nmr. *Biochemistry*, 33(15):4721-4729.

Gasteiger J and Marsili M (1980) Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, 36(22):3219-3228.

Gordon DB, Hom GK, Mayo SL and Pierce NA (2003) Exact rotamer optimization for protein design. *Journal of Computational Chemistry*, 24(2):232-243.

Gordon DB, Marshall SA and Mayo SL (1999) Energy functions for protein design. *Current Opinion in Structural Biology*, 9(4):509-513.

Greenwald J and Riek R (2010) Biology of amyloid: Structure, function, and regulation. *Structure*, 18(10):1244-1260.

Grigoryan G (2013) Absolute free energies of biomolecules from unperturbed ensembles. *Journal of Computational Chemistry*, 34(31):2726-2741.

Grigoryan G and Keating AE (2006) Structure-based prediction of bzip partnering specificity. *J Mol Biol*, 355(5):1125-1142.

Grigoryan G, Ochoa A and Keating AE (2007) Computing van der waals energies in the context of the rotamer approximation. *Proteins: Structure, Function, and Bioinformatics*, 68(4):863-878.

Grigoryan G, Reinke AW and Keating AE (2009) Design of protein-interaction specificity gives selective bzip-binding peptides. *Nature*, 458(7240):859-852.

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT and Clore GM (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein-g. *Science*, 253(5020):657-661.

Guo JJ and Zhou HX (2016) Protein allostery and conformational dynamics. *Chemical Reviews*, 116(11):6503-6515.

Havranek JJ and Harbury PB (2003) Automated design of specificity in molecular recognition. *Nature Structural Biology*, 10(1):45-52.

Heanue TA, Reshef R, Davis RJ, Mardon G, Oliver G, Tomarev S, Lassar AB and Tabin CJ (1999) Synergistic regulation of vertebrate muscle development by *dach2*, *eya2*, and *six1*, homologs of genes required for drosophila eye formation. *Genes & Development*, 13(24):3231-3243.

Heath AP, Kavraki LE and Clementi C (2007) From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins-Structure Function and Bioinformatics*, 68(3):646-661.

Henikoff S and Henikoff JG (1992) Amino-acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915-10919.

Henzler-Wildman K and Kern D (2007) Dynamic personalities of proteins. *Nature*, 450(7172):964-972.

Herrmann T, Guntert P and Wuthrich K (2002) Protein nmr structure determination with automated noe assignment using the new software candid and the torsion angle dynamics algorithm dyana. *Journal of Molecular Biology*, 319(1):209-227.

Hilpert K, Winkler DFH and Hancock REW (2007) Peptide arrays on cellulose support: Spot synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. *Nature Protocols*, 2(6):1333-1349.

Howell SC, Inampudi KK, Bean DP and Wilson CJ (2014) Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases. *Structure*, 22(2):218-229.

Huang DW, Sherman BT and Lempicki RA (2009) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1-13.

Huang J, Perez-Burgos L, Placek BJ, Sengupta R, Richter M, Dorsey JA, Kubicek S, Opravil S, Jenuwein T and Berger SL (2006) Repression of p53 activity by smyd2-mediated methylation. *Nature*, 444(7119):629-632.

Huang J, Sengupta R, Espejo AB, Lee MG, Dorsey JA, Richter M, Opravil S, Shiekhataar R, Bedford MT, Jenuwein T and Berger SL (2007a) P53 is regulated by the lysine demethylase Lsd1. *Nature*, 449(7158):105-U180.

Huang PS, Love JJ and Mayo SL (2007b) A de novo designed protein-protein interface. *Protein Science*, 16(12):2770-2774.

Humphris EL and Kortemme T (2008) Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure*, 16(12):1777-1788.

Hutchinson EG and Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Science*, 3(12):2207-2216.

Hynes TR and Fox RO (1991) The crystal-structure of staphylococcal nuclease refined at 1.7 Å resolution. *Proteins-Structure Function and Genetics*, 10(2):92-105.

Itzhaki LS, Otzen DE and Fersht AR (1995) The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods - evidence

for a nucleation-condensation mechanism for protein-folding. *Journal of Molecular Biology*, 254(2):260-288.

Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T and Kuhlman B (2016) Design of structurally distinct proteins using strategies inspired by evolution. *Science*, 352(6286):687-690.

Jansson H, Bergman R and Swenson J (2011) Role of solvent for the dynamics and the glass transition of proteins. *Journal of Physical Chemistry B*, 115(14):4099-4109.

Jee L, Byeon I, Louis JM and Gronenborn AM (2007) Solution structure gb1 a34f mutant. *Structure Deposited to the Protein Data Bank*, PDBID: 2RMM.

Jha RK, Leaver-Fay A, Yin SY, Wu YB, Butterfoss GL, Szyperski T, Dokholyan NV and Kuhlman B (2010) Computational design of a pak1 binding protein. *Journal of Molecular Biology*, 400(2):257-270.

Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL and Baker D (2008) De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387-1391.

Jiang YY, Trescott L, Holcomb J, Zhang X, Brunzelle J, Sirinupong N, Shi XB and Yang Z (2014) Structural insights into estrogen receptor alpha methylation by histone methyltransferase smyd2, a cellular event implicated in estrogen signaling regulation. *Journal of Molecular Biology*, 426(20):3413-3425.

Johnson B and Blevins R (1994) Nmr view: A computer program for the visualization and analysis of nmr data. *J Biomol NMR*, 4(5):603-614.

Jones DT (1994) De-novo protein design using pairwise potentials and a genetic algorithm. *Protein Science*, 3(4):567-574.

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R and Wu AY (2002) An efficient k-means clustering algorithm: Analysis and implementation. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881-892.

Kapp GT, Liu S, Stein A, Wong DT, Remenyi A, Yeh BJ, Fraser JS, Taunton J, Lim WA and Kortemme T (2012) Control of protein signaling using a computationally designed gtpase/gef orthogonal pair. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5277-5282.

Karanicolas J and Kuhlman B (2009) Computational design of affinity and specificity at protein-protein interfaces. *Curr Opin Struct Biol*, 19(4):458-463.

Kawamura S, Yoshigai E, Kuhara S and Tashiro K (2008) Smyd1 and smyd2 are expressed in muscle tissue in xenopus laevis. *Cytotechnology*, 57(2):161-168.

Keedy DA, Kenner LR, Warkentin M, Woldeyes RA, Hopkins JB, Thompson MC, Brewster AS, Van Benschoten AH, Baxter EL, Uervirojnangkoorn M, McPhillips SE, Song JH, Alonso-Mori R, Holton JM, Weis WI, Brunger AT, Soltis SM, Lemke H, Gonzalez A, Sauter NK, Cohen AE, van den Bedem H, Thorne RE and Fraser JS (2015) Mapping the conformational landscape of a dynamic enzyme by multitemperature and xfel crystallography. *Elife*, 4(1):e07574.

Kellogg EH, Leaver-Fay A and Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, 79(3):830-838.

Kemnitz CR and Loewen MJ (2007) "Amide resonance" correlates with a breadth of c-n rotation barriers. *Journal of the American Chemical Society*, 129(9):2521-2528.

Kirkpatrick S, Gelatt CD and Vecchi MP (1983) Optimization by simulated annealing. *Science*, 220(4598):671-680.

Klemm JD, Wozniak JA, Alber T and Goldenberg DP (1991) Correlation between mutational destabilization of phage-t4 lysozyme and increased unfolding rates. *Biochemistry*, 30(2):589-594.

Knab JR, Chen JY, He YF and Markelz AG (2007) Terahertz measurements of protein relaxational dynamics. *Proceedings of the Ieee*, 95(8):1605-1610.

Komatsu S, Imoto I, Tsuda H, Kozaki K, Muramatsu T, Shimada Y, Aiko S, Yoshizumi Y, Ichikawa D, Otsuji E and Inazawa J (2009) Overexpression of smyd2 relates to tumor cell proliferation and malignant outcome of esophageal squamous cell carcinoma. *Carcinogenesis*, 30(7):1139-1146.

Kortemme T and Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14116-14121.

Krieger E, Nabuurs SB and Vriend G (2003) Homology modeling. 25:507-521. *"Methods of biochemical analysis: Structural bioinformatics"*, Bourne PE and Weissig H. Hoboken, New Jersey, United States of America.

Kudithipudi S, Dhayalan A, Kebede AF and Jeltsch A (2012) The set8 h4k20 protein lysine methyltransferase has a long recognition sequence covering seven amino acid residues. *Biochimie*, 94(11):2212-2218.

Kudithipudi S, Lungu C, Rathert P, Happel N and Jeltsch A (2014) Substrate specificity analysis and novel substrates of the protein lysine methyltransferase nsd1. *Chemistry & Biology*, 21(2):226-237.

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

Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H and Sarai A (2006) Protherm and pronit: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research*, 34(D):204-206.

Kuszewski J, Gronenborn AM and Clore GM (1999) Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. *Journal of the American Chemical Society*, 121(10):2337-2338.

Labute P (2009) Protonate3d: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins-Structure Function and Bioinformatics*, 75(1):187-205.

Lacroix E, Viguera AR and Serrano L (1998) Elucidating the folding problem of alpha-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of nmr parameters. *Journal of Molecular Biology*, 284(1):173-191.

Lanouette S, Davey JA, Elisma F, Ning ZB, Figeys D, Chica RA and Couture JF (2015) Discovery of substrates for a set domain lysine methyltransferase predicted by multistate computational protein design. *Structure*, 23(1):206-215.

Lanouette S, Mongeon V, Figeys D and Couture JF (2014) The functional diversity of protein lysine methylation. *Molecular Systems Biology*, 10(4):724-749.

Larson SM, England JL, Desjarlais JR and Pande VS (2002) Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. *Protein Science*, 11(12):2804-2813.

Lassila JK (2010) Conformational diversity and computational enzyme design. *Current Opinion in Chemical Biology*, 14(5):676-682.

Lauck F, Smith CA, Friedland GF, Humphris EL and Kortemme T (2010) Rosettabackrub-a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Research*, 38(W):569-575.

Lazaridis T and Karplus M (1999) Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics*, 35(2):133-152.

Lee GM and Craik CS (2009) Trapping moving targets with small molecules. *Science*, 324(5924):213-215.

Lennard-Jones JE (1924) On the determination of molecular fields - ii from the equation of state of a gas. *Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character*, 106(738):463-477.

Lewandowski JR, Halse ME, Blackledge M and Emsley L (2015) Direct observation of hierarchical protein dynamics. *Science*, 348(6234):578-581.

Lian LY, Derrick JP, Sutcliffe MJ, Yang JC and Roberts GCK (1992) Determination of the solution structures of domain-ii and domain-iii of protein-g from streptococcus by h-1 nuclear-magnetic-resonance. *Journal of Molecular Biology*, 228(4):1219-1234.

Lilien RH, Stevens BW, Anderson AC and Donald BR (2005) A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *Journal of Computational Biology*, 12(6):740-761.

Lovell SC, Davis IW, Adrendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS and Richardson DC (2003) Structure validation by c alpha geometry: Phi,psi and c beta deviation. *Proteins-Structure Function and Genetics*, 50(3):437-450.

Mackereil AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D and Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102(18):3586-3616.

Malakauskas SM and Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology*, 5(6):470-475.

Mandell DJ and Kortemme T (2009) Computer-aided design of functional protein interactions. *Nat Chem Biol*, 5(11):797-807.

Marvin JS and Hellinga HW (2001) Conversion of a maltose receptor into a zinc biosensor by computational design. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):4955-4960.

Mayo SL, Olafson BD and Goddard WA (1990) Dreiding - a generic force-field for molecular simulations. *Journal of Physical Chemistry*, 94(26):8897-8909.

McMurray J (2004) Biomolecules: Amino acids, peptides, and proteins. 26(1):988-989. "Organic chemistry", McMurray J. Belmont, California, United States of America.

Mcphalen CA and James MNG (1987) Crystal and molecular-structure of the serine proteinase-inhibitor ci-2 from barley-seeds. *Biochemistry*, 26(1):261-269.

Meng WL, Lyle N, Luan BW, Raleigh DP and Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6):2123-2128.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087-1092.

Mittag T and Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. *Current Opinion in Structural Biology*, 17(1):3-14.

Molecular operating environment (2011.13) Chemical Computing Group, Inc. Montreal, Québec, Canada.

Molecular operating environment (2012.14) Chemical Computing Group, Inc. Montreal, Québec, Canada.

Murphy GS, Mills JL, Miley MJ, Machius M, Szyperski T and Kuhlman B (2012) Increasing sequence diversity with flexible backbone protein design: The complete redesign of a protein hydrophobic core. *Structure*, 20(6):1086-1096.

Myers JK, Pace CN and Scholtz JM (1995) Denaturant m-values and heat-capacity changes - relation to changes in accessible surface-areas of protein unfolding. *Protein Science*, 4(10):2138-2148.

Nash SG (2000) A survey of truncated-newton methods. *Journal of Computational and Applied Mathematics*, 124(1-2):45-59.

Nauli S, Kuhlman B and Baker D (2001) Computer-based redesign of a protein folding pathway. *Nature Structural Biology*, 8(7):602-605.

Newman JRS and Keating AE (2003) Comprehensive identification of human bzip interactions with coiled-coil arrays. *Science*, 300(5628):2097-2101.

Ollikainen N, Smith CA, Fraser JS and Kortemme T (2013) Flexible backbone sampling methods to model and design protein alternative conformations. *Methods in Protein Design*, 523(1):61-85.

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

Parak F and Formanek H (1971) Study on vibration and crystal lattice fault of temperature factor in myoglobin by comparative mossbauer absorption measurements with x-ray structure data. *Acta Crystallographica Section A*, 27(A):573-578.

Patrick AN, Cabrera JH, Smith AL, Chen XJS, Ford HL and Zhao R (2013) Structure-function analyses of the human six1-eya2 complex reveal insights into metastasis and bor syndrome. *Nature Structural & Molecular Biology*, 20(4):447-453.

Piao L, Kang D, Suzuki T, Masuda A, Dohmae N, Nakamura Y and Hamamoto R (2014) The histone methyltransferase smyd2 methylates parp1 and promotes poly (adp-ribosyl)ation activity in cancer cells. *Neoplasia*, 16(3):157-264.

Pierce NA and Winfree E (2002) Protein design is np-hard. *Protein Engineering*, 15(10):779-782.

Pokala N and Handel TM (2005) Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology*, 347(1):203-227.

Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN and Mayo SL (2012) Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America*, 109(10):3790-3795.

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu XW, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJR, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF and Snyder M (2005) Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068):679-684.

Ramachandran GN, Ramakrishnan C and Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95-99.

Rasmussen BF, Stock AM, Ringe D and Petsko GA (1992) Crystalline ribonuclease-a loses function below the dynamic transition at 220-k. *Nature*, 357(6377):423-424.

Rathert P, Dhayalan A, Murakami M, Zhang X, Tamas R, Jurkowska R, Komatsu Y, Shinkai Y, Cheng XD and Jeltsch A (2008) Protein lysine methyltransferase g9a acts on non-histone targets. *Nature Chemical Biology*, 4(6):344-346.

Renfrew PD, Choi EJ, Bonneau R and Kuhlman B (2012) Incorporation of noncanonical amino acids into rosetta and use in computational protein-peptide interface design. *Plos One*, 7(3):e32637.

Robustelli P, Cavalli A and Vendruscolo M (2008) Determination of protein structures in the solid state from nmr chemical shifts. *Structure*, 16(12):1764-1769.

Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS and Baker D (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190-194.

Saddic LA, West LE, Aslanian A, Yates JR, Rubin SM, Gozani O and Sage J (2010) Methylation of the retinoblastoma tumor suppressor by smyd2. *Journal of Biological Chemistry*, 285(48):37733-37740.

Saio T, Ogura K, Yokochi M, Kobashigawa Y and Inagaki F (2009) Two-point anchoring of a lanthanide-binding peptide to a target protein enhances the paramagnetic anisotropic effect. *Journal of Biomolecular Nmr*, 44(3):157-166.

Sakamoto LHT, de Andrade RV, Felipe MSS, Motoyama AB and Silva FP (2014) Smyd2 is highly expressed in pediatric acute lymphoblastic leukemia and constitutes a bad prognostic factor. *Leukemia Research*, 38(4):496-502.

Sammond DW, Eletr ZM, Purbeck C, Kimple RJ, Siderovski DP and Kuhlman B (2007) Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. *J Mol Biol*, 371(5):1392-1404.

Sandberg WS and Terwilliger TC (1991) Energetics of repacking a protein interior. *Proceedings of the National Academy of Sciences of the United States of America*, 88(5):1706-1710.

Saunders CT and Baker D (2005) Recapitulation of protein family divergence using flexible backbone protein design. *Journal of Molecular Biology*, 346(2):631-644.

Schneider M, Fu XR and Keating AE (2009) X-ray vs. Nmr structures as templates for computational protein design. *Proteins: Structure, Function, and Bioinformatics*, 77(1):97-110.

Schubert HL, Blumenthal RM and Cheng XD (2003) Many paths to methyltransfer: A chronicle of convergence. *Trends in Biochemical Sciences*, 28(6):329-335.

Schultz HP (1989) Topological organic-chemistry. 1. Graph-theory and topological indexes of alkanes. *Journal of Chemical Information and Computer Sciences*, 29(3):227-228.

Schultz HP, Schultz EB and Schultz TP (1990) Topological organic-chemistry. 2. Graph-theory, matrix determinants and eigenvalues, and topological indexes of alkanes. *Journal of Chemical Information and Computer Sciences*, 30(1):27-29.

Sese B, Barrero MJ, Fabregat MC, Sander V and Belmonte JCI (2013) Smyd2 is induced during cell differentiation and participates in early development. *International Journal of Developmental Biology*, 57(5):357-364.

Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA and Mayo SL (2007) Full-sequence computational design and solution structure of a thermostable protein variant. *Journal of Molecular Biology*, 372(1):1-6.

Shen Y, Delaglio F, Cornilescu G and Bax A (2009) Talos plus : A hybrid method for predicting protein backbone torsion angles from nmr chemical shifts. *Journal of Biomolecular Nmr*, 44(4):213-223.

Shifman JM and Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23):13274-13279.

Shortle D, Stites WE and Meeker AK (1990) Contributions of the large hydrophobic amino-acids to the stability of staphylococcal nuclease. *Biochemistry*, 29(35):8033-8041.

Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE and Baker D (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*, 329(5989):309-313.

Sirinupong N, Brunzelle J, Doko E and Yang Z (2011) Structural insights into the autoinhibition and posttranslational activation of histone methyltransferase smyd3. *Journal of Molecular Biology*, 406(1):149-159.

Sirinupong N, Brunzelle J, Ye J, Pirzada A, Nico L and Yang Z (2010) Crystal structure of cardiac-specific histone methyltransferase smyd1 reveals unusual active site architecture. *Journal of Biological Chemistry*, 285(52):40635-40644.

Skinner MM, Zhang H, Leschnitzer DH, Guan Y, Bellamy H, Sweet RM, Gray CW, Konings RNH, Wang AHJ and Terwilliger TC (1994) Structure of the gene-v protein of bacteriophage-f1 determined by multiwavelength x-ray-diffraction on the selenomethionyl protein. *Proceedings of the National Academy of Sciences of the United States of America*, 91(6):2071-2075.

Smith CA and Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4):742-756.

Smith CA and Kortemme T (2011) Predicting the tolerated sequences for proteins and protein interfaces using rosettabackrub flexible backbone design. *Plos One*, 6(7):e20451.

Smith CA and Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic pdz domains. *Journal of Molecular Biology*, 402(2):460-474.

Steed PM, Tansey MG, Zalevsky J, Zhukovsky EA, Desjarlais JR, Szymkowski DE, Abbott C, Carmichael D, Chan C, Cherry L, Cheung P, Chirino AJ, Chung HH, Doberstein SK, Eivazi A, Filikov AV, Gao SX, Hubert RS, Hwang M, Hyun L, Kashi S, Kim A, Kim E, Kung J, Martinez SP, Muchhal US, Nguyen DHT, O'Brien C, O'Keefe D, Singer K, Vafa O, Vielmetter J, Yoder SC and Dahiyat BI (2003) Inactivation of tnf signaling by rationally designed dominant-negative tnf variants. *Science*, 301(5641):1895-1898.

Stewart JJP (2007) Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12):1173-1213.

Stranges PB and Kuhlman B (2013) A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, 22(1):74-82.

Street AG and Mayo SL (1999) Computational protein design. *Structure*, 7(5):105-109.

Street AG and Mayo SL (1998) Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design*, 3(4):253-258.

Strop P, Marinescu AM and Mayo SL (2000) Structure of a protein g helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Science*, 9(7):1391-1394.

Sturgeon JB and Laird BB (2000) Symplectic algorithm for constant-pressure molecular dynamics using a nose-poincare thermostat. *Journal of Chemical Physics*, 112(8):3474-3482.

Suarez M, Tortosa P, Carrera J and Jaramillo A (2008) Pareto optimization in computational protein design with multiple objectives. *Journal of Computational Chemistry*, 29(16):2704-2711.

- Summa CM, Rosenblatt MM, Hong JK, Lear JD and DeGrado WF (2002) Computational de novo design, and characterization of an a(2)b(2) diiron protein. *Journal of Molecular Biology*, 321(5):923-938.
- Tarek M and Tobias DJ (2002) Role of protein-water hydrogen bond dynamics in the protein dynamical transition. *Physical Review Letters*, 88(13):138101.
- Tomlinson JH, Green VL, Baker PJ and Williamson MP (2010) Structural origins of pH-dependent chemical shifts in the b1 domain of protein g. *Proteins-Structure Function and Bioinformatics*, 78(14):3000-3016.
- Traore S, Roberts KE, Allouche D, Donald BR, Andre I, Schiex T and Barbe S (2016) Fast search algorithms for computational protein design. *Journal of Computational Chemistry*, 37(12):1048-1058.
- Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins .V. Conformation of a system of 3 linked peptide units. *Biopolymers*, 6(10):1425-1436.
- Vitkup D, Ringe D, Petsko GA and Karplus M (2000) Solvent mobility and the protein 'glass' transition. *Nature Structural Biology*, 7(1):34-38.
- Voelkel T, Andresen C, Unger A, Just S, Rottbauer W and Linke WA (2013) Lysine methyltransferase smyd2 regulates hsp90-mediated protection of the sarcomeric titin springs and cardiac function. *Biochimica Et Biophysica Acta-Molecular Cell Research*, 1833(4):812-822.
- Voet D and Voet JG (2004) Three-dimensional structure of proteins. 8(3):245-258. "Biochemistry", Voet D and Voet JG. Hoboken, New Jersey, United States of America.
- Voigt CA, Gordon DB and Mayo SL (2000) Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*, 299(3):789-803.
- Wang C, Schueler-Furman O and Baker D (2005) Improved side-chain modeling for protein-protein docking. *Protein Science*, 14(5):1328-1339.
- Wang JM, Cieplak P and Kollman PA (2000) How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049-1074.
- Weaver LH and Matthews BW (1987) Structure of bacteriophage-t4 lysozyme refined at 1.7 a resolution. *Journal of Molecular Biology*, 193(1):189-199.

Weik M and Colletier JP (2010) Temperature-dependent macromolecular x-ray crystallography. *Acta Crystallographica Section D-Biological Crystallography*, 66(437-446).

Wen X, Tannukit S and Paine ML (2008) Tifp11 interacts with mdeah9, an rna helicase involved in spliceosome disassembly. *International Journal of Molecular Sciences*, 9(11):2105-2113.

Williams CI and Feher M (2008) The effect of numerical error on the reproducibility of molecular geometry optimizations. *Journal of Computer-Aided Molecular Design*, 22(1):39-51.

Wilton DJ, Tunnicliffe RB, Kamatari YO, Akasaka K and Williamson MP (2008) Pressure-induced changes in the solution structure of the gb1 domain of protein g. *Proteins-Structure Function and Bioinformatics*, 71(3):1432-1440.

Winkler DFH and Campbell WD (2008) Methods in molecular biology: Peptide-based drug design. 4:47-70. "*The spot technique: Synthesis and screening of peptide microarrays on cellulose membranes*", Otvos L. New York, United States of America.

Wishart D, Sykes B and Richards F (1992) The chemical shift index: A fast and simple method for the assignment of protein secondary structure through nmr spectroscopy. *Biochemistry*, 31(6):1647-1651.

Wylie BJ, Sperling LJ, Nieuwkoop AJ, Franks WT, Oldfield E and Rienstra CM (2011) Ultrahigh resolution protein structures using nmr chemical shift tensors. *Proceedings of the National Academy of Sciences of the United States of America*, 108(41):16974-16979.

Xu JR and Zhang Y (2010) How significant is a protein structure similarity with tm-score=0.5? *Bioinformatics*, 26(7):889-895.

Yanover C, Fromer M and Shifman JM (2007) Dead-end elimination for multistate protein design. *Journal of Computational Chemistry*, 28(13):2122-2129.

Yin S, Ding F and Dokholyan NV (2007a) Modeling backbone flexibility improves protein stability estimation. *Structure*, 15(12):1567-1576.

Yin SY, Ding F and Dokholyan NV (2007b) Eris: An automated estimator of protein stability. *Nature Methods*, 4(6):466-467.

Zhang X, Tanaka K, Yan JS, Li J, Peng DN, Jiang YY, Yang Z, Barton MC, Wen H and Shi XB (2013) Regulation of estrogen receptor α by histone methyltransferase smyd2-mediated

protein methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(43):17284-17289.

Zhang Y and Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins-Structure Function and Bioinformatics*, 57(4):702-710.

Zuber J, Rappaport AR, Luo W, Wang E, Chen C, Vaseva AV, Shi J, Weissmueller S, Fellmann C, Taylor MJ, Weissenboeck M, Graeber TG, Kogan SC, Vakoc CR and Lowe SW (2011) An integrated approach to dissecting oncogene addiction implicates a myb-coordinated self-renewal program as essential for leukemia maintenance. *Genes Dev*, 25(15):1628-1640.