University of Ottawa
Chemistry and Biomolecular Sciences
Faculty of Science

Université d'Ottawa
Chimie et Sciences Biomoléculaires
Faculté des Sciences

From protein sequence to motion to function: Towards the rational design of functional protein dynamics

Adam Damry

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements for the Ph.D. degree in Chemistry

Abstract

Protein dynamics are critical to the structure and function of proteins. However, due to the complexity they inherently bring to the protein design problem, dynamics historically have not been considered in computational protein design (CPD). Herein, we present meta-MSD, a new CPD methodology for the design of protein dynamics. We applied our methodology to the design of a novel mode of conformational exchange in Streptococcal protein G domain B1, producing dynamic variants we termed DANCERs. Predictions were validated by NMR characterization of selected DANCERs, confirming that our meta-MSD framework is suitable for the computational design of protein dynamics. We then performed a thorough NMR characterization of the sequence determinants of dynamics in one DANCER, isolating two mutations responsible for the novel dynamics this protein exhibits. The first, A34F, is responsible for destabilizing the highly stable native Gβ1 conformation, allowing the protein to sample other conformational states. The second, V39L mediates subtle interactions that stabilize the designed conformational trajectory in the context of the A34F mutation. Together, these results highlight the role of protein plasticity in the development of dynamics and the need for highly accurate computational tools to approach similar design problems. Finally, we present an NMR-based characterization of structural dynamics in a family of related red fluorescent proteins (RFPs) and pinpoint regions of the RFP structure where dynamics correlate to RFP brightness. This overview of the RFP dynamics-function relationship will be used in future projects to perform a computation design of functional dynamics in RFPs.

Acknowledgements

The work I present in this thesis would not have been possible without the contributions of several key individuals, towards whom I would like to express my gratitude.

To my supervisor, Dr. Roberto A. Chica, I am grateful for your supervision throughout my years as a graduate student. I would not be where I am now without your support. You have always pushed me to greater heights, while still allowing me the space to develop into an independent scientist. I am thankful for the trust you invested in me, and for the confidence you held in me. Moving forward, you can be sure that I will hold that to heart, and that I will not bring shame to the lab as an alumnus.

To Dr. Natalie K. Goto... I cannot imagine what my graduate years would have been like without your mentorship and guiding hand aiding me along. Be it when I needed a sanity check for some results or just someone to listen to me gushing about something or another that had me either thrilled or upset, you were always there when I needed you, and I could not be more thankful for it.

To my colleagues James A. Davey, Marc M. Mayer, Serena E. Hunt, and Dr. Aron Broom, it has been a pleasure and an honor to work alongside you on the projects outlined herein. Your contributions made these projects what they are, and I would never have been able to complete this substantial body of work without each and every minute you offered me.

To Dr. Glenn A. Facey, Dr. Eric Ye, and Dr. Peter Pallister, I would like to extend my thanks for all the help provided with the NMR spectrometers that truly were the workhorse instruments for my work. Whenever I had one problem or another, or needed a new pulse sequence, I knew that I could always count on you.

To the other members of the Chica lab, with a special mention to my dear friend Matthew G. Eason with whom I've spent countless hours bouncing ideas off one another, I want to express my gratitude for the companionship you provided. You all truly made the lab group feel like a family to me.

And finally, to Sarah B. Damry, my dear mother, none of this at all would have been possible without your support every step of the way. I could not have asked for a better pillar.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CD – Circular Dichroism

C_m – Chemical denaturation midpoint of unfolding

CPD – Computational Protein Design

CPMG – Carr-Purcell-Meiboom-Gill

DANCER – Dynamic and Native Conformational ExchangeR

EC – Exctinction Coefficient

EDTA – Ethylenediaminetetraacetic acid

FASTER – Fast and Accurate Side-chain Topology and Energy Refinement

FID – Free Induction Decay

FP – Fluorescent Protein

FRET – Förster Resonance Energy Transfer

Gβ1 – Streptococcal protein G domain β1

GFP – Green Fluorescent Protein

GMEC – Global Minimum Energy Configuration

HetNOE – Heteronuclear Nuclear Overhauser Effect

HSQC – Heteronuclear Single Quantum Coherence

IDP – Intrinsically Disordered Protein

IgG – Immunoglobulin G

IPTG – Isopropyl β-D-1-thiogalactopyranoside

LB – Luria-Bertani

MOE – Molecular Operating Environment

MRE – Mean Residue Ellipticity

MSD – Multistate Design

NERD – Non-Exchanging Rigid Design

NMR – Nuclear Magnetic Resonance

NOE – Nuclear Overhauser Effect

PDB – Protein Data Bank

PertMin – Coordinate Perturbation followed by energy Minimization

QY – Quantum Yield

R_1 – Longitudinal Relaxation Rate

R_2 – Transverse Relaxation Rate

RFP – Red Fluorescent Protein

RMSD – Root Mean Square Deviation

SEC – Size-Exclusion Chromatography

SSD – Single-state Design

T_1 – Longitudinal Relaxation Time

T_2 – Transverse Relaxation Time

$\boldsymbol{\tau}_c$ – Correlation Time

$\boldsymbol{\tau}_c^{app}$ – Apparent Correlation Time

T_m – Thermal denaturation midpoint of unfolding

UV – Ultraviolet

Chapter 1: Introduction

1.1 Protein structure and function

Proteins are large macromolecules responsible for complex processes that occur throughout biology. As a result of the vast functional wealth provided by the twenty proteinogenic amino acids, whose side-chains include non-polar and polar, acidic and basic, aliphatic and aromatic, and even nucleophilic groups, they are capable of carrying out as varied functions as catalysis of metabolic reactions,[1] stimulus response,[2,3] providing structure to cell components,[4,5] molecular transport,[6,7] and more. These functions are dictated by the protein's structure, a functionalized fold capable of accomplishing a specific biological function that arises from amino acid interactions with both other amino acids and with the surrounding medium. These structures are unique to each protein, and thus to each protein function. However, protein structure can be classified into a hierarchy of increasing complexity or order.

At the lowest complexity level, a protein's primary structure refers to the sequence of the amino acid chain from which it is composed, in absence of any higher ordering or three-dimensional structure. Interactions between local amino acids which create repeating structures, such as α-helices, β-sheets, and turns, form the protein's secondary structure. At a higher complexity still, interactions such as hydrogen bonds, disulfide bonds, and solvation define the spatial relationship of secondary structure elements, giving rise to a three-dimensional structure often comprising the necessary elements for protein function such as binding and catalytic sites. At the highest level of complexity, quaternary structure refers to the superstructure formed when several separate protein chains, each possessing a tertiary structure, interact together to form a single protein complex, thus defining the spatial relationship of tertiary structure elements. As protein function is the result of a precise arrangement of amino acids at a defined functional site,

which may be distant in absentia of higher-order structure, protein structure and protein function are intrinsically and tightly linked to one another. Thus, understanding the nature of this relationship remains a longstanding goal of structural biology.

Historically, approaches to the analysis of structure-function relationships in proteins have relied either on global fold similarities or on local motif similarities,[8-11] as it has been shown that structural similarity is more tightly associated to functional similarity than is sequence similarity.[12] However, these approaches are not foolproof, as many similar folds are found in proteins possessing different functions.[13] These observations suggest that protein function is therefore not only a question of the global protein fold, but also of finer interactions occurring within the protein.[14] Notably, though proteins are often studied as static structures that can be frozen out through techniques like X-ray crystallography,[15] many are in fact highly dynamic systems. Though complex in nature, structural protein dynamics have been shown to influence and even determine protein function by causing fluctuations in protein structure over a time component that are lost when handling static protein structures.[16-18] Thus, it is critical to consider the contribution of dynamics when studying the protein structure-function relationship.

1.1.1 Protein dynamics

Proteins are highly variable, dynamic entities exhibiting motion on length scales ranging from the sub-angstrom to nm scales and timescales ranging from picoseconds to seconds. Smaller amplitude motions tend to be correlated to faster timescales and an increasingly harmonic nature, whereas larger amplitude motions tend to be slower and anharmonic in nature.[19] However, protein motions are as diverse as the proteins in which they are found. It is therefore difficult to classify protein dynamics as many proteins are involved in several different dynamic regimes simultaneously, exhibiting observable motions that are a convoluted aggregate of these different

microscopic motions, and similar motions may occur on different timescales in different proteins. Local organization of protein structure plays an important role in determining the timescale of dynamic motions, where a structural element in a highly ordered and energetically stable region of the protein will tend to exhibit slower motions than an equivalent structural element in a more disordered or less stable region. Thus, we cannot define rigid timescale boundaries that must be respected for a given dynamic regime as even a simple motion such as a side-chain rotation may require a complex trajectory to accomplish that involves the rearrangement of other structural elements. Nonetheless, as amplitude is correlated to timescale of motion, dynamics involving larger structural elements tend to be slower. In spite of these complexities, protein dynamics can be broadly grouped by nature.[1,20,21]

At picosecond to short nanosecond timescales, dynamics are correlated primarily to local flexibility where fast bond vibration, rotation, and torsion give rise to larger motions involving side-chain or local structural elements. These local fluctuations are typically quasi-harmonic in nature, where stochastic motion allows protein elements to explore a single energy well of the protein energy landscape before returning to a single equilibrium conformation. The ruggedness of this energy landscape does not preclude rapid transitions between energy wells, such as exchange between side-chain rotameric configurations. However, the probability of encountering anharmonic motion where several energy wells are being explored over the course of a dynamic trajectory is proportional to both range and timescale of motion. As we move towards slower nanosecond timescale motions, dynamics therefore begin to involve the concerted motion of secondary structure elements such as loops. At the microsecond timescale and beyond, dynamics begin to involve collective motions of higher-order structural elements that are representative of

processes such as domain motion, ligand binding, and catalysis rather than independent motions of local structural elements.[20]

As dynamics occurring within a protein on different timescales correspond to concerted motions of different hierarchical elements, it is not uncommon for several dynamic regimes to be experienced at any given point in the protein. For example, examining a single side-chain may reveal fast nanosecond timescale harmonic motion, while examining the loop in which this side-chain is found may show that the entire loop is in motion at a slower timescale. The entire loop could in turn be part of a domain that is in conformational equilibrium at a timescale that is slower still. Complicating the nature of dynamics, these motions may occur independently from one another, or may be interlinked, where one mode of motion is promoted or inhibited by another. In a more complete picture of protein dynamics in reference to the protein energy landscape, dynamic motions can be therefore conceptualized as rapid fluctuations within one energy well concerted with slower transitions between energy wells that are themselves features at the bottom of larger wells corresponding to other even slower transitions (Fig. 1.1).

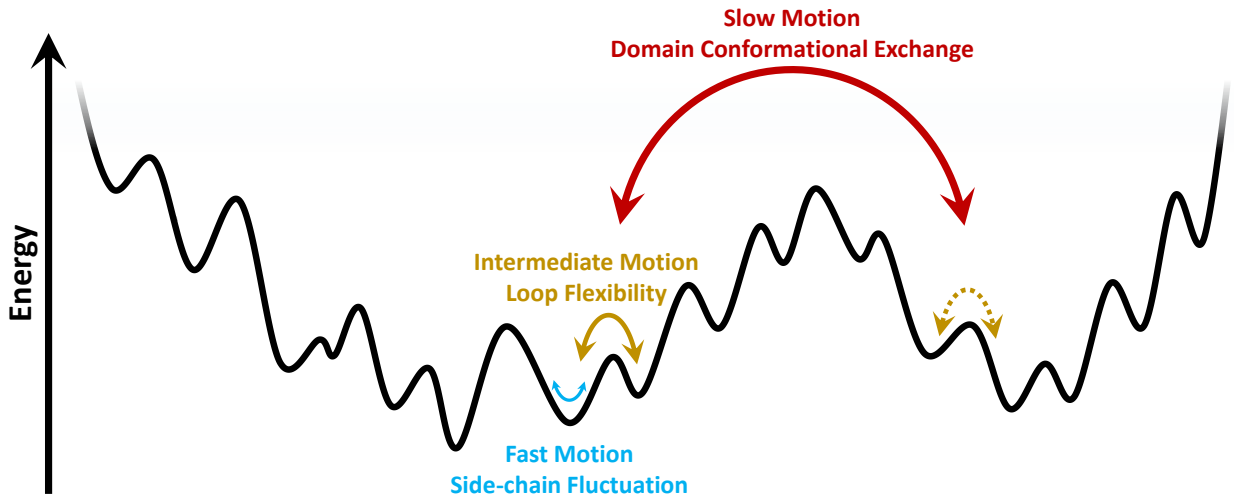


Figure 1.1. Timescale of motions as modeled on a hypothetical energy landscape. For an arbitrary hypothetical protein element, dynamics can be conceptualized as motions on the protein energy landscape, which relates protein configuration to internal energy. Several dynamic regimes can be experienced concurrently, corresponding to different transitions on the protein energy landscape. Fast quasi-harmonic motions such as the side-chain fluctuation shown here in blue correspond to the motions within an energy well. Slower anharmonic motions derive from transitions between energy wells, such as the intermediate timescale loop motion shown here in yellow. Several transitions of various hierarchy can be exhibited simultaneously, such as this loop flexibility occurring alongside the slow domain conformational exchange shown in red. These transitions may be independent, wherein the both transitions would still be observed irrespective of the other, or may be interlinked, wherein the probability for one motion to occur is affected by the other. In this hypothetical example, in the case where loop flexibility and domain conformational exchange are independent, the transition corresponding to loop flexibility will be accessible independent of domain conformational exchange as shown by the dotted yellow arrow. If this transition is inaccessible or slowed in one domain conformation, loop flexibility becomes dependent on domain conformational exchange.

Beyond the observations of the effects of protein motion on the global protein structure, dynamics have also been shown to play a significant role in protein function.[21-24] As protein function is, again, dependent on the precise geometry of active site residues in relation to the molecules they interact with, dynamics that modulate this geometry thus determine form and efficacy of function. This principle is the basis of the allosteric regulation of protein function, where the reversible binding of an allosteric effector at an allosteric site alters protein conformation at the active site in a manner that can either promote or inhibit its activity, or capacity to perform

its function. Though allostery is itself a dynamics-linked process,[25] this concept illustrates how dynamic conformational changes are capable of regulating protein function by forcing the active site into or out of a productive conformation. Further evidence for the functional role of dynamics in proteins comes from the study of intrinsically disordered proteins (IDPs). Though these proteins lack a defined three-dimensional structure under native conditions and are best described as a dynamic ensemble, they nonetheless possess discrete functions and are particularly important in crucial metabolic regulation functions.[26-29] This suggests that despite the disordered, dynamic structure of IDPs, their dynamics are capable of giving rise to functionally active poses.

However, protein dynamics are not linked to function solely through the dichotomy of functional versus non-functional poses. Several proteins, including enzymes, transporters, binding proteins, and more, rely on dynamic processes to accomplish their functions. In enzymes, multi-step catalytic processes may require different active site geometries to promote each step of the catalytic cycle, as well as substrate binding and product release.[30-32] As another example, many transporters are highly dynamic proteins whose dynamics facilitate the transport of target molecules through the cell membrane, and potentially against the concentration gradient.[33,34] Evolution-based studies have also shown that protein dynamics are key not only to existing protein functions, but also to the evolution of novel protein functions wherein dynamics promoting new function are expanded through evolution, followed by a reduction in dynamics not associated with this function that lead to inactive conformations.[18,35] Thus, it is widely accepted that protein dynamics play an intrinsic and necessary role in protein function. However, the form in which this is manifested is widely variable. In certain proteins, function is accompanied by wide-ranging motions that relocate entire domains (Fig. 1.2a), while in others, the effects of dynamics are far subtler (Fig. 1.2b), highlighting the system-dependence of dynamics-function relationships.

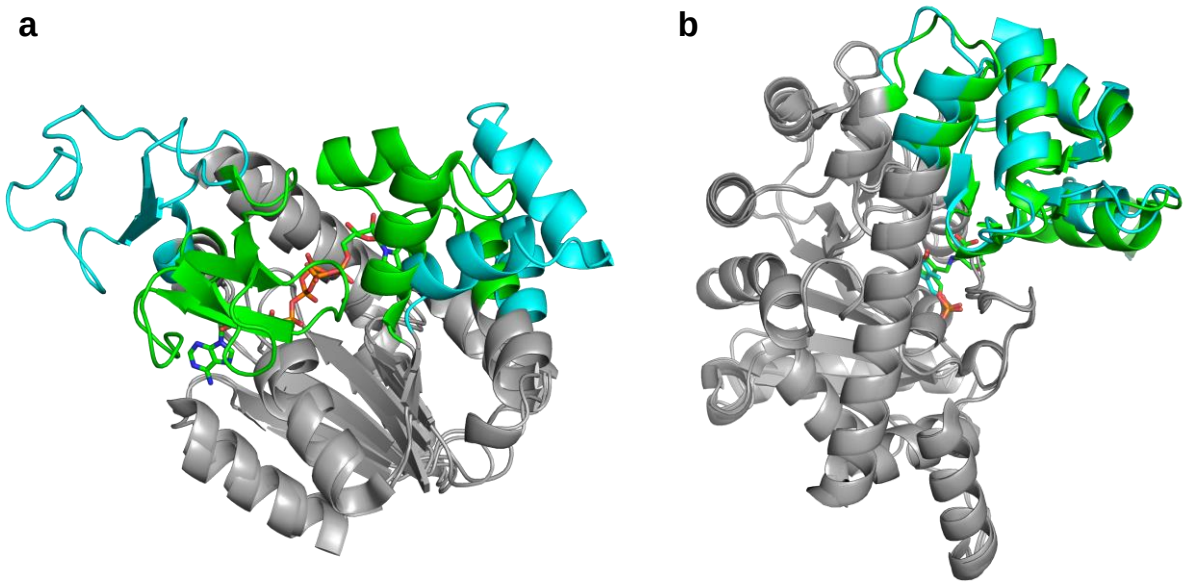


Figure 1.2. Protein dynamics mediate protein function. a, Cartoon representations of crystal structures of *E. coli* adenylate kinase with bound substrate analog shown as sticks (green, PDB ID: 1AKE[36]) and without substrate analog (cyan, PDB ID: 4AKE[37]) reveal a large conformational change that occurs upon substrate binding, highlighting how dynamics can cause substantial structural changes in proteins over their catalytic cycle. The CORE domain, which does not experience conformational exchange upon substrate binding, is shown in grey. **b,** Crystal structures of wild-type aspartate aminotransferase in the open conformation (green, PDB ID: 1ARS[38]) and a hexamutant in the closed conformation with bound substrate shown as sticks (cyan, PDB ID: 1X28[30]) reveal a small conformational change that occurs upon substrate binding, highlighting how small conformational changes can play a role in protein function. The static domain is shown in grey.

Given the proven link between dynamics and complex protein function, the rational protein design field has long been interested in including dynamic considerations in protein design. Despite the number of proteins for which dynamics-function relationships have been established through the use of techniques such as molecular dynamics simulations and NMR spectroscopy, our understanding of the general mechanisms underpinning these relationships remains poor, as dynamics are the result of incredibly complex and highly system-dependent concerted interactions within the protein structure. Thus, it is extremely difficult to accurately predict how mutations will affect dynamics beyond the local environment, and how dynamics might in turn affect function if

this relationship has not already been studied for the system in question. With the advent of highly accurate energy forcefields capable of replicating protein motion,[39-41] computational methods are poised to address the first amongst these obstacles. However, several others still remain for the rational design of dynamics and complex protein function to become feasible and widely applicable.

1.1.2 From protein sequence to dynamics to function

In rational protein design, the primary structure of proteins is modified in a specific and targeted manner to effect alterations to higher-order structural elements, and often in turn to effect a modification to protein function. The protein design problem is thus generally approached as the problem linking protein sequence to protein fold. The mechanisms linking these structural elements are relatively well understood in the simplified case of rigid, mostly static systems where the contribution of dynamics to function is negligible, as evidenced by the design of highly accurate *de novo* proteins by computational protein design methodologies.[42,43] The inclusion of dynamics into protein design methodologies however adds several additional layers of complexity to the protein design problem. The first is a matter of scale. Protein dynamics, by definition, cannot be represented by a single static protein structure. There is therefore a need to consider as many relevant states to the dynamic pathway as possible during designs. Therein lie additional complicating factors, as current computational protein design methodologies have not been built with the optimization of multiple conformational or chemical states in mind. Though there is no theoretical upper limit to the number of states that can be included in a design, it becomes difficult to define a fitness score encompassing more than one or two states in a manner that allows sequence discrimination. Many of the aforementioned relevant states to dynamic exchange are either unknown, present an undesignable high-energy transition state, or both. In addition, the links

between protein sequence and protein dynamics, between protein dynamics and protein function, and therefore between protein sequence and protein function in a dynamic system are poorly understood.[18]

The computational design of protein dynamics has thus far remained an untouchable problem. However, we seek here to bridge the gap between protein design and dynamics and to develop a computational protein design methodology capable of approaching the highly complex problem that is the design of dynamics. Using NMR validation to back our designs, we aim to design dynamics in the model protein Streptococcal protein G domain β1. We are also interested in studying the link between dynamics and function in red fluorescent proteins to better understand how dynamics at distal positions in these proteins affect dynamics at their chromophore (active site) and by extension their function, in anticipation of using these proteins as a second test case to benchmark our design methodologies.

1.2 Streptococcal protein G domain β1 (Gβ1)

Streptococcal protein G is a protein natively expressed in group C and G *Streptococcus*. In its native form, protein G is capable of tightly binding the Fc and Fab regions of IgG antibodies in a variety of different-sized complexes and demonstrates one of the broadest species and subclass specificities amongst known IgG binding proteins.[44,45] These binding properties have led to widespread use of protein G in antibody purification protocols, and early studies pinpointed a repeating subunit of 56 amino acids that formed the β1, β2, and β3 (in the case of group G *Streptococcus*) subdomains of protein G as the IgG Fc domain binding site.[46] Most widely studied amongst these is the β1 domain, or Gβ1, formed by four β-strands with an amphipathic α-helix crossing over the sheet these strands form, arranged in a βαβ hairpin-helix-hairpin structure (Fig. 1.3). Despite the protein's small size, it is fully globular in nature,[47] with a tightly-packed

hydrophobic core formed by one face of the β-sheet and the hydrophobic portion of the α-helix. Calorimetry of Gβ1 revealed an extreme thermal stability with a melting temperature (T_m) of 87 °C and reversible thermal denaturation,[48] making Gβ1 one of the most thermally stable well-characterized globular proteins at that time.[49] Though protein G was already widely used in immunology, the thermal properties and small size of its β1 domain, as well as an extreme solubility,[50,51] led it to become one of the most widely used model protein to study protein folding and to benchmark analytical techniques such as protein NMR.[52]

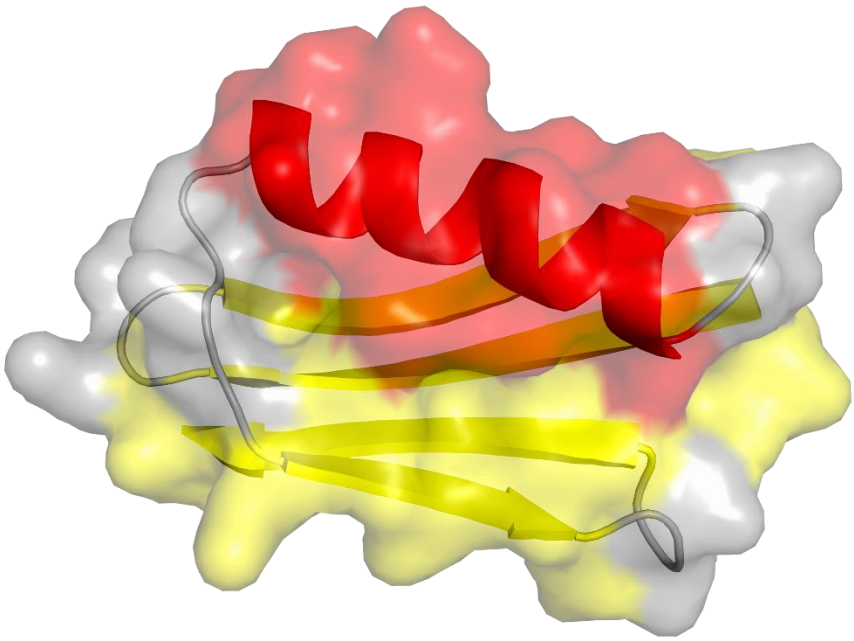


Figure 1.3. Crystal structure of Streptococcal protein G domain β1. A cartoon representation of the Gβ1 crystal structure (PDB ID: 1PGA)[53] is shown with the α-helix colored in red, β-strands in yellow, and loops in grey, showing the βαβ structure. A representation of the protein surface is overlaid, highlighting the well-packed, globular nature of Gβ1.

As we sought to use CPD to design novel dynamic properties into a protein, we chose Gβ1 as our model system of choice due to its native rigidity and suitability for NMR spectroscopy, one of the most informative techniques for atomic-resolution characterization of protein dynamics, where Gβ1 produces remarkably well-resolved spectra. On top of the experimental Gβ1 stability

and folding data available, current physics-based CPD forcefields and MSD methodologies were also benchmarked using Gβ1, resulting in an unparalleled wealth of both experimental and computational data to guide our designs.[54,55] Also important is the remarkable structural plasticity demonstrated by Gβ1. Despite its small size, high thermal stability, and the rigidity of its native structure, Gβ1 mutants have been shown to adopt a wide range of structures, including a dimer,[56] a domain-swapped dimer,[57] and a triple-helix bundle.[58] This is indicative of significant ruggedness in the Gβ1 energy landscape, and thus that a redesign of its dynamics should be feasible. Finally, Gβ1 also presents desirable sequence characteristics for our purposes. Its sequence contains no cysteine residues, which complicate designs through the need to consider the possibility of disulfide bridge formation, nor any proline residues that are invisible to standard NH-detected NMR experiments. Gβ1 also contains a single tryptophan residue that is spectroscopically distinct and will be used as a probe of dynamics.

Although Gβ1 does possess a function, that of IgG binding, this function is rarely recapitulated in protein structure studies using Gβ1 as a model as they focus instead on its stability and fold. There are therefore no reports of links between binding efficacy of Gβ1 and dynamics throughout its structure, in part due to the rigidity of the native Gβ1 structure. Likewise, we are interested in using Gβ1 to develop novel CPD methodologies for the design of dynamics, thus in using Gβ1 as a model protein that is highly responsive to standard dynamics-characterization experiments rather than for its function. However, the link between dynamics and function remains an incredibly important factor in the development of protein design methodologies, and we were interested in studying how to design function through dynamics once our work with Gβ1 had yielded a protocol for the computational design of dynamics. Rather than attempt to design the

binding-based function of Gβ1, we opted to instead move to a different system in which function and its link to dynamics could be more easily assayed, namely red fluorescent proteins.

1.3 Red fluorescent proteins (RFPs)

Red fluorescent proteins (RFPs) are genetically encodable fluorophores formed entirely from proteinogenic amino acids that emit light at wavelengths greater than ~570 nm. The first RFP to be cloned, DsRed, was isolated from a *Discosoma* coral species in 1999.[59] An obligate homotetramer possessing a β-barrel structure homologous to the structure of *Aequorea victoria*-derived green fluorescent proteins,[60] DsRed served as the parent to a future generation of engineered RFPs that would become widely used fluorescent probes. Since the discovery of DsRed, other natural RFPs have been discovered, such as eqFP578 and eqFP611, both isolated from *Entacmaea quadricolor*,[61] but natural RFPs found to date have all been oligomeric.[62,63] As monomeric FPs such as GFP are more desirable for imaging purposes than oligomers,[64] DsRed was engineered into a monomer, mRFP1, through sequential disruption of its protein-protein interfaces, followed by directed evolution to recover fluorescence in the resulting non-fluorescent monomer.[65] Further engineering efforts aimed at improving the spectral properties of mRFP1, in particular maturation and photostability, led to the creation of mCherry, amongst the most widely used RFPs today.[66] Though several variant RFPs have been developed since mCherry, in both the DsRed family and others, most intrinsically fluorescent monomeric RFPs share strong structural similarities. They form an 11-strand β-barrel at the center of which is found an α-helix supporting an intrinsically fluorescent chromophore formed by cyclization of a XYG tripeptide in an oxygen-dependent process termed maturation that generates a conjugated system comprising an acylimine group, an imidazolinone group, and a phenolate group (Fig. 1.4).[67-69]

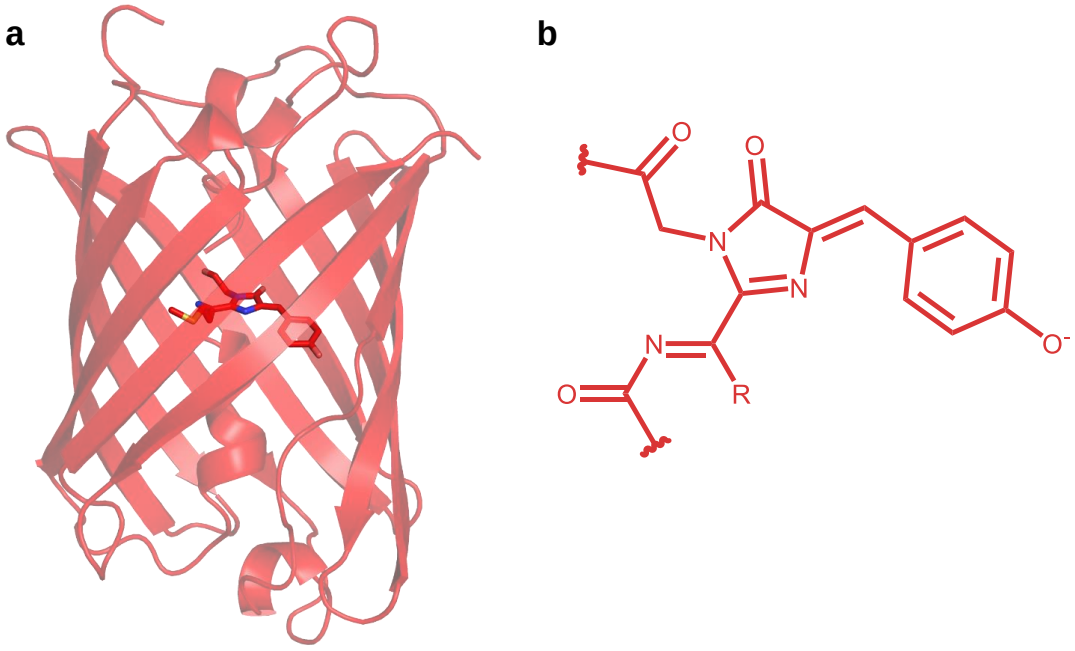


Figure 1.4. Structure of an archetypal RFP. a, The archetypal RFP β-barrel, taken from the mCherry crystal structure (PDB ID: 2H5Q)[70] is shown as a cartoon representation with the RFP chromophore shown as sticks within the barrel. **b,** The structure of the RFP chromophore is shown, demonstrating the conjugated system responsible for its fluorescence.

Since their discovery, FPs have become indispensable to biochemistry research for imaging applications *in vivo*. FP-based constructs have been used for as varied applications as protein tagging within cells and tissues,[71] to detect and study protein-protein and protein-ligand interactions through FRET pairing,[72,73] as signaling moieties in biosensors,[74-76] and more. Though FPs of varied wavelengths see use for all of these applications, RFPs possess certain advantages due to their longer excitation and emission wavelengths. As these wavelengths are lower in energy, they are less damaging to cells, and undergo a lower degree of scattering in tissue. In addition, the most absorbent biomolecules in complex tissues in the visible light spectrum, melanin and hemoglobin, are less absorbent in the RFP spectral range than at shorter wavelengths.[77] These beneficial properties are enhanced as emission and excitation wavelengths further increase, leading for a push towards the development of far-red fluorescent proteins.

Another critical FP property is brightness, as defined by the product of the chromophore's extinction coefficient (EC, ε) and quantum yield (QY, φ). The extinction coefficient, derived from the Beer-Lambert law, represents the chromophore's efficiency at absorbing incoming light, typically at its excitation wavelength, and is defined by Equation 1.1, where A is the absorbance of light at a given wavelength, c is the concentration of the chromophore, and l is the optical path length.

$$(Eq.\,1.1) \quad \varepsilon = \frac{A}{cl}$$

The quantum yield on the other hand represents the chromophore's capacity to reemit absorbed light through fluorescence,[78,79] defined by Equation 1.2.

$$(Eq.\,1.2) \quad \varphi = \frac{\#\ photons\ emitted}{\#\ photons\ absorbed}$$

Despite their desirable red-shifted emission and excitation wavelengths, RFPs tend to possess low brightness compared to their green and yellow counterparts, as evidenced by the archetypal RFP mCherry still possessing a brightness of less than half that of the widely used GFP, EGFP.[66,80] Though significant effort has been levied to engineer brighter red fluorescent proteins,[65,66,81-83] resulting in the creation of a few extremely bright RFP variants, such as mScarlet which possess a QY of 0.70,[83] far-red RFPs with emission wavelengths of > 630 nm still remain subpar.

Parsing through a database of FP properties reveals that the low brightness of far-red RFPs stems from their generally low QY rather than their EC.[84] As QY is dependent on the probability that an excited-state chromophore relaxes through fluorescence rather than alternative non-radiative relaxation pathways,[85] this suggests that non-radiative decay is promoted in RFPs relative to other FPs. As the energy gap between excited and ground states is smaller in RFPs than for FPs

emitting higher-energy shorter-wavelength light, RFPs could possess an increased probability of internal conversion from an excited state vibrational level to a matching vibrational level in the ground state (Fig. 1.5a). Though several factors such as chromophore planarity[70,86] and capacity to undergo *cis-trans* isomerization[77,87] are hypothesized to play a role in determining QY and brightness, there is significant evidence from both experimental and computational sources that rigidifying the RFP chromophore would restrict access to the vibrational states allowing non-radiative relaxation to occur, and thus increase the both QY and brightness of RFP (Fig. 1.5b).[88-90] By extension, RFPs therefore provide us with a system in which concrete dynamic properties, the rigidity of the chromophore, are linked to an easily measured protein function, fluorescence. As RFPs are also easily expressed, moderately small proteins of ~27 kDa size, this makes them a suitable system in which to design and study protein dynamics that are linked to a protein function.

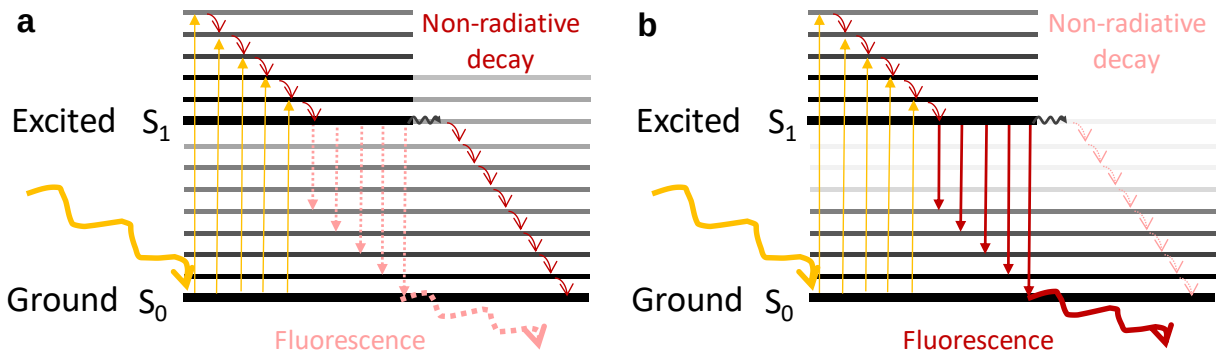


Figure 1.5. Jablonski diagram of RFP chromophore relaxation pathways. a. When the RFP chromophore is excited to its S_1 excited state, it can relax back to its S_0 ground state through either radiative (fluorescence) or non-radiative pathways. If internal conversion to an accessible S_0 vibrational level occurs, vibrational relaxation returns the chromophore to its ground state without fluorescence occurring. **b.** Inhibiting access to accessible S0 vibrational states through rigidification of the RFP chromophore reduces the probability of an internal conversion leading to non-radiative relaxation modes, and thus promotes fluorescence and increases QY.

1.4 Overview of principal techniques

Over the course of the projects presented herein, both Gβ1 and RFPs were used in studies of protein dynamics. Using Gβ1 as a model system, we developed a computational protein design (CPD) methodology for the design of dynamics, followed by a characterization of structure and dynamics in designed Gβ1 variants using protein NMR and circular dichroism. A thorough characterization of dynamics in mCherry-family RFPs was also performed using NMR, in anticipation of using the results for future designs of dynamics in RFPs. The principal techniques used throughout these projects, CPD, NMR, and CD, will be discussed herein.

1.4.1 Computational protein design (CPD)

Computational protein design is a computational methodology developed to identify protein sequences that can adopt and stabilize a specific protein structure.[91] Various CPD protocols have also since been developed to target and design various protein properties such as substrate specificity or binding affinity.[92,93] At its core however, the CPD methodology typically follows the same general steps for all of these applications. In general terms, sequences are first threaded onto a template structure defined by the design objective. Then, an energy calculation is computed to evaluate the sequence in the context of the design template. Finally, a sequence optimization step searches for sequences predicted to meet the design objective, thus the most energetically-favorable sequences in the context of the given design template, and outputs results as a rank-ordered list of sequences. Though CPD has yet to be applied to the design of protein dynamics, a prospective methodology for the design of dynamics would still follow these same basic steps. Thus, we will herein discuss the basic concepts governing the application of CPD to general protein design problems.

1.4.1.1 Potential energy functions

As CPD is a technique built around energy minimization and/or optimization, the heart of any CPD methodology is arguably its potential energy function. This energy function represents a relationship between a protein's structure and the energy intrinsic to that structure and is the primary tool by which sequences are evaluated and ranked. In practice, as it is not feasible to calculate the internal energy of a protein, we instead approximate this energy as that of arbitrarily referenced interaction energies between pairs of rotamers, which are idealized, low-internal energy side-chain configurations, and between rotamers and the backbone template. Several different energy functions exist, though they are generally classified in two categories. Statistical energy functions such as Rosetta[94-96] assign energies based on a empirically-derived statistical propensity for an interaction to occur, while physics-based energy functions such as PHOENIX[54,97] calculate the potential energy of interactions arising from discrete physical processes. As the designs presented in Chapter 2 of this thesis were performed using the PHOENIX protein design software, we will focus on the PHOENIX energy function. In this physics-based energy function, potential energy of a rotameric configuration (E_{total}) is represented as a sum of energy terms including van der Waals (vdW), electrostatics (elec), hydrogen bonding (H-bond), and solvation (solv), as shown in Equation 1.3.

$$(Eq.\,1.3)\quad E_{total} = E_{vdW} + E_{elec} + E_{H-bond} + E_{solv}$$

The van der Waals term (E_{vdw}) is calculated using a standard 12-6 Lennard-Jones potential, with atomic radii and well depth drawn from the Dreiding II force field,[98] as represented by Equation 1.4, where D_0 represents the energy well depth, α represents the van der Waals scaling factor, R represents interatomic distance, and R_0 represents the geometric mean of atomic radii.

$$(Eq.\,1.4)\quad E_{vdW} = D_0\left[\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right]$$

The electrostatics term (E_{elec}) is derived from Coulomb's law, and is represented by Equation 1.5, where q and q' represent the atomic charges separated by R, the interatomic distance, and ε represents the dielectric constant of the medium.

$$(Eq.\,1.5)\quad E_{elec} = \frac{q * q'}{\varepsilon R}$$

The hydrogen bond term ($E_{\text{H-bond}}$) is calculated using a geometry-dependent 12-6 potential energy term, as represented by Equation 1.6, where D_0, R, and R_0 again represent energy well depth, interatomic distance, and the geometric mean of atomic radii respectively. $F(\theta)$ is an angle-dependent correction factor that accounts for H-bond geometry.[55]

$$(Eq.\,1.6)\quad E_{H-Bond} = D_0\left[5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{6}\right]F(\theta)$$

Unlike other terms, which are calculated through discrete equations characterizing physical processes, the solvation term (E_{solv}) is not calculated using an explicit solvent model, which would be computationally demanding, but rather through a surface-area based model, where energetic benefits or penalties are calculated based on the solvent-exposed surface area of rotamers.[99,100] Following this model, burial of non-polar groups and exposure of polar groups is considered beneficial, while exposure of non-polar groups and burial of polar groups is penalized.

As the energies calculated by the energy function are arbitrarily referenced, their absolute magnitudes are difficult to interpret. Rather, sequence energies are compared to one another in CPD to rank order sequences following a probability of adopting the desired structure, where a

more energetically favorable sequence is seen as being more likely to adopt the designed structure than a higher-energy sequence on the same fold.

1.4.1.2 Sequence optimization algorithms

Despite the simplifications made in CPD energy functions to render complex energy calculations tractable for large sequence and structure datasets, it is still not feasible to test every possible configuration of rotamers on a given structure to find the global minimum energy configuration (GMEC) due to the enormity of combinatorial sequence space. Even considering only specific design positions often results in an intractable number of rotamer configurations to evaluate. Thus, CPD methodologies typically make use of a sequence optimization algorithm (or search algorithm) to converge towards the GMEC without having to test every sequence and rotamer configuration. Several algorithms exist, including both deterministic algorithms such as Dead-End-Elimination[101] that rigorously and reproducibly locate the GMEC and stochastic algorithms such as Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER)[102] that rely on probabilistic trajectories to converge rapidly, but do not guarantee the GMEC and may not provide reproducible solutions. As the rotamer optimization problem becomes exponentially more complex as the set of designed positions increases in size, and thus is classified as an NP-hard problem,[103] deterministic algorithms may not be computationally tractable to use for complex design problems. Thus, stochastic algorithms are widely used in CPD.

Despite not guaranteeing the GMEC, these stochastic algorithms rapidly converge towards a solution and thus do not significantly sample local sequence space, which is generally desirable for single-variable optimization problems like traditional positive and negative MSD, where a single fitness score is optimized. This very same rapid convergence however complicates potential approaches to the design of dynamics. Many dynamic trajectories cannot be easily condensed into

a single fitness score however, as the consideration of several states, in terms of both absolute and relative energies is required. However, if all relevant states are not considered simultaneously, the rapid convergence of these search algorithms makes it likely that sequences predicted to stabilize one state will be different from those predicted to stabilize a different state, potentially with little overlap between the sequence spaces searched in each case.

1.4.1.3 Single-state design (SSD)

The objective of a CPD calculation is to evaluate protein sequences in the context of a given protein structure and to return a fitness score for each sequence. A simple application of this concept is to conduct this search of sequence space using a single fixed backbone as template structure in a process termed single-state design (SSD). Beyond defining a sequence space to be searched and a backbone template on which to thread sequences, SSD also makes use of an energy function and search algorithm as described above, as well as a rotamer library that defines which rotamers will be evaluated during design calculations. As only a single protein structure is used, the fitness of a sequence in SSD is by definition the energy returned by the energy function for the evaluation of that sequence on the template (E_{seq}) in its minimum energy rotamer configuration, as defined by Equation 1.7, where E_{rc} represents the energy of a given rotamer configuration comprising *n* discrete rotamers (*r*) for a given sequence in the context of the backbone template, as defined by Equation 1.8. This energy is comprised of two terms; a one-body term (E_i) that arises from interactions between each rotamer *i* and the fixed backbone template in absence of other rotamers, and a two-body term (E_{ij}) that arises from interactions between each rotamer *i* and all other rotamers in absence of the fixed backbone template.

$$(Eq.\,1.7)\quad E_{seq} = \min(E_{rc})$$

$$(Eq.\,1.8)\quad E_{rc} = \sum_{i=1}^{n} E_i[r_i] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij}[r_i, r_j]$$

As the scope of this energy calculation scales combinatorially with the number of positions designed and thus rotamers evaluated,[104] it is generally not tractable to design all positions in a protein. Nonetheless, SSD is arguably amongst the least computationally demanding CPD methodologies, and has proven effective at predicting mutations that stabilize the input template structure.[91,105-107] However, the use of a single fixed backbone in SSD in combination with rigid rotamers leads to false negative predictions through the rejection of desirable sequences that would be accepted if the backbone geometry or rotamer configuration were slightly changed.[108,109] Various strategies such as flexible backbone algorithms,[110,111] iterative energy minimization,[112] and continuous rotamer optimization[113] have been implemented to reduce error introduced by this fixed backbone approximation. However, another flaw of SSD and SSD-like methodologies is that they preclude the evaluation of multiple protein states in a single design calculation. Proteins, however, are dynamic molecules, and for many proteins, a single state is insufficient to allow the design of a desired property.

1.4.1.4 Multistate design (MSD)

In cases where SSD has proven successful, design goals were tractable with the use of a single fixed backbone and an energy-minimizing scoring function. This approach however is not feasible any time we wish to consider multiple conformational or chemical states within the context of a single design problem. To address these fixed backbone-derived limitations that are intrinsic to SSD, multistate design (MSD) was developed by expanding the concepts driving SSD to utilize backbone ensembles for sequence evaluation rather than a single fixed backbone. The core concepts of MSD are identical to SSD, where an energy function is used to calculate the

energies of rotamer configurations threaded onto backbone templates, with a search algorithm guiding the search of sequence space. However, rather than score sequences on a single backbone, the process is parallelized to calculate energies for these sequences on each member of one or more backbone ensembles comprising several different backbone templates.[54,109,114,115] We thus obtain one energy corresponding to the minimum energy rotamer configuration for each sequence on each backbone template. These energies are then recombined using a scoring function into a single fitness score that can be evaluated in the context of the particular design objective and is the basis for sequence optimization by the search algorithm, thus sequences are optimized on the ensemble as a whole rather than on individual members. In the case of sequence stability, this fitness score is generally taken as the Boltzmann-weighted average of sequence energies on all members of the backbone ensemble. However, several different scoring functions can be used, depending on the design objective, contributing to the flexibility of MSD as a design tool (Fig. 1.6).

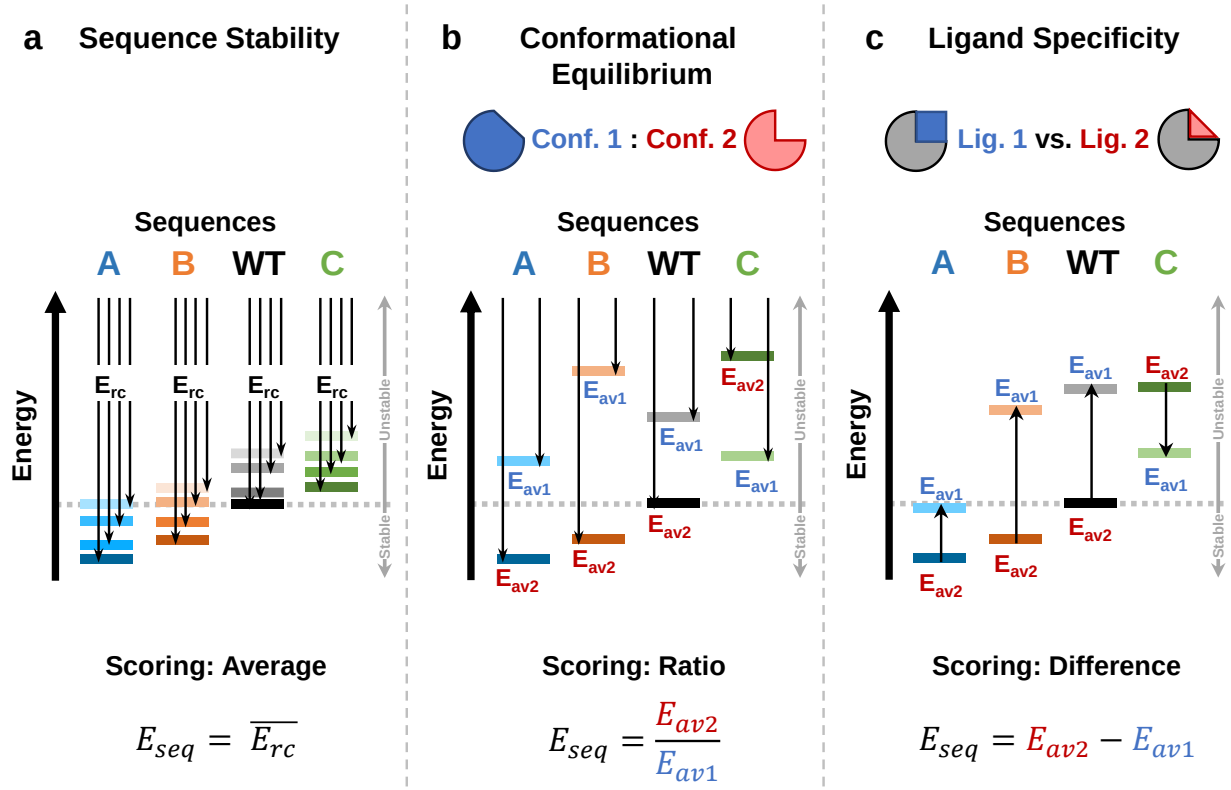


Figure 1.6. Examples of MSD and associated scoring functions. a, In a multistate design for sequence stability, sequences are scored on a backbone ensemble, and minimum energy rotamer configuration energies (E_{rc}) are combined in a Boltzmann-weighted average to give the fitness score for each sequence on the ensemble. **b,** In a multistate design for conformational equilibrium, sequences are scored on ensembles corresponding to one of two discrete protein conformations. Average energies on each ensemble (E_{av1} and E_{av2}) are combined ratiometrically to give the fitness score for each function. **c,** In a multistate design for ligand specificity, sequences are scored on ensembles corresponding to backbone structures bound to one of two ligands. Average energies on each ensemble (E_{av1} and E_{av2}) are combined differentially to give the fitness score for each function.

Current MSD methodologies can generally be categorized as either positive or negative MSD. Positive MSD relies on additive scoring functions such as energy sums and averages where all ensemble members contribute positively to the fitness score, and is useful in applications such as the design of stability where one absolute energy value is being optimized.[116] In negative MSD on the other hand, a differential scoring function is used, where certain ensemble members contribute positively to the fitness score and others contribute negatively, which is useful in

applications such as the redesign of substrate specificity where an energy differential between two competing conformational or chemical states is being optimized.[117]

Ultimately, whether a positive or negative MSD approach is used, MSD is capable of considering the energetic contributions of multiple protein conformational states towards a single design objective. As the design of protein dynamics by definition must consider protein motion or multiple conformational states, MSD is a promising technique with which to develop a CPD methodology for the prediction and design of protein dynamics. Two major challenges however still remain to be overcome. First, it is unclear how to model a backbone ensemble that is properly representative of a non-native dynamic regime, and second, as sequence optimization algorithms are guided by a single fitness score, stabilizing multiple conformational states along a desired dynamic trajectory will be difficult. Nonetheless, if these barriers to the computational design of dynamics can be overcome, the ability of MSD to consider a potentially unlimited number of conformational states would make it a powerful technique for designing and studying protein dynamics *in silico*. To reach this point however, we will need to develop a thorough understanding of protein dynamics *in vitro*, for which techniques such as nuclear magnetic resonance spectroscopy will be invaluable for their capacity to study protein dynamics in the solution state.

1.4.2 Nuclear magnetic resonance spectroscopy (NMR)

Nuclear magnetic resonance (NMR) spectroscopy is a technique that studies local magnetic effects around atomic nuclei, which are a function of their local electronic environment. Any nucleus with a non-zero nuclear spin quantum number (I) will be NMR active, however it is most common to use nuclei with an I of ½ such as ^{1}H, ^{13}C, or ^{15}N. As these nuclei possess a non-zero spin, or intrinsic angular momentum, they therefore exhibit the same properties as a rotating charged species and generate a magnetic dipole, whose magnitude is dependent on an intrinsic

property of the nucleus, its gyromagnetic ratio (γ).[118,119] When placed in an external magnetic field (B_o), these dipoles align themselves to the applied field, and in the case of I = ½ nuclei, adopt either a spin-aligned or spin-opposed orientation. These states are not of equivalent energies, separated by a theoretical energy gap defined by Equation 1.9, where $\hbar$ represents the reduced Planck's constant.

$$(Eq.\,1.9)\quad \Delta E = \gamma \hbar B_o$$

Thus, what is detected in NMR spectroscopy is the absorption of energy required to stimulate a transition from the lower energy spin-aligned state to the higher energy spin-opposed state. As the size of this energy difference depends both on γ and B_o, NMR spectroscopy becomes more sensitive when using nuclei with a higher gyromagnetic ratio such as ^{1}H and a higher external magnetic field.

Typically, the energy state transitions detected by NMR correspond to radiofrequency-range electromagnetic radiation. This frequency (v) corresponds to the rate of precession of the dipole in the external magnetic field, which is known in terms of angular momentum (rad s^{-1}) as the Larmor frequency ($\omega_o = 2\pi v$). Though ω_o depends primarily on γ and B_o, it is also dependent on the local movement of electrons near the nucleus that give rise to small magnetic fields and alter the effective magnetic field experienced by the nucleus. Therefore, the real absorption frequency of a given nucleus is instead represented by Equation 1.10, where B_{eff} is the effective magnetic field at the nucleus, and includes contributions from factors such as electron density, electrostatic environment, and bonding interactions.

$$(Eq.\,1.10)\quad \Delta E = \gamma \hbar B_{eff}$$

As these shielding effects are proportional to B_o, it is common to report them as shifts from a reference frequency to normalize for field strength, known as chemical shifts (δ), and as shielding effects are typically small compared to the contribution of the external magnetic field to B_{eff}, chemical shifts are typically reported in parts per million (ppm), defined by Equation 1.11, where ν_s and ν_{ref} correspond to the NMR frequencies of the sample nucleus and reference nucleus respectively.

$$(Eq.1.11) \quad \delta = \frac{\nu_s - \nu_{ref}}{\nu_{ref}} \times 10^6\, ppm$$

When working with proteins, 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) is typically employed as reference compound, with the nine equivalent methyl protons distinguishing the 0.0 ppm reference frequency. As the silicon atom's low electronegativity strongly shields these methyl protons, almost all protons found in organic molecules, including proteins, are found downfield (at a higher chemical shift) of this reference frequency.

In an NMR spectrometer, NMR signals are acquired through the detection of electromagnetic field oscillations. As the energy gap between spin-aligned and spin-opposed dipole orientations leads to a population imbalance between orientations, a bulk magnetization vector arises in the population that is aligned to B_o at equilibrium and that precesses around the B_o-defined axis with a frequency of ω_o. The application of a radiofrequency pulse perpendicular to the external magnetic field and matching the Larmor frequency of the nucleus in question causes the bulk magnetization vector to rotate perpendicular to the applied pulse. By tuning the length and power of this pulse, the bulk magnetization vector can be shifted into the transverse plane, where its precession can be detected by a receiver coil oriented perpendicularly to B_o as an oscillating electromagnetic field. A Fourier transform is then used to convert this time-domain

data, known as a free induction decay (FID), back to a frequency, again corresponding to the ω_o of the nucleus in question. In the case of a real sample, it is likely that more than one resonant frequency will be observed, resulting in a complex convolution of oscillations and giving rise to a complex FID, for which a Fourier transform becomes essential to decouple signals from one another.

In practice, it is also important to consider the effects of scalar couplings (or *J*-couplings) on magnetization, which are through-bond interactions between two nuclear spins. *J*-coupling interactions result in the partial or total transfer of magnetization from excited nuclei to adjacent nuclei. Coupling causes the splitting of NMR signals into multiplets that reveal information about molecular connectivity. However, this can be undesirable during chemical shift evolution and data acquisition of complex systems, leading to the inclusion of decoupling pulses in many NMR experiments. Nonetheless, coupling is vital to protein NMR, as the capacity to transfer magnetization from one atom to another through *J*-couplings and thus indirectly measure the chemical shift of several atoms in a single experiment is the basis of all multidimensional NMR. This capacity of NMR to observe several coupled atoms in a single spectrum and thus to draw correlations between bonded atoms is crucial to the study of complex systems such as proteins that would otherwise generate too many signals to deconvolve from one another.

Despite the atomistic resolution afforded by these correlation-based multidimensional NMR experiments, due to the small energy gap between the spin-aligned and spin-opposed dipole orientations, the bulk magnetization vector is small in magnitude, leading to NMR being inherently insensitive. For this reason, it is desirable to work with high-field magnets and high-gyromagnetic ratio isotopes. However, the sensitivity and resolution of an NMR experiment are determined not only by these parameters, but also by spin relaxation, which refers to the processes by which an

NMR signal decays over time. Relaxation is primarily composed of two components.[120] Longitudinal relaxation or spin-lattice relaxation (T_1) refers to the return of excited nuclei to their ground state, represented as a return of the bulk magnetization vector to a parallel alignment with the external magnetic field that is tied to a reduction in NMR signal intensity over time. Transverse relaxation or spin-spin relaxation (T_2) on the other hand refers to the dephasing of the bulk magnetization vector and is tied to both a reduction in NMR signal intensity and an increase in peak broadness over time. Both types of relaxation depend on dipole coupling and chemical shift anisotropy, thus on fluctuations in the local magnetic field generated by the motion of local spins in the external magnetic field. These fluctuations are in turn a function of the rate of molecular tumbling that is represented by the molecular rotational correlation time (τ_c). Importantly, as both T_1 and T_2 are a function of molecular motion, they are important observables in the study of protein dynamics by NMR, which will be discussed later. Since the oscillations responsible for relaxation occur at specific frequencies, as evidenced by excitation at the Larmor frequency ω_o, we can also describe molecular motion in an NMR sample using a spectral density function $J(\omega)$, which represents the power available to stimulate transitions at a given angular frequency ω, and is represented by Equation 1.12.

$$(Eq.\,1.12)\quad J(\omega) = \overline{B_{eff}^2}\ \frac{2\tau_c}{1 + (\omega\tau_c)^2}$$

For a large molecule with long τ_c like a protein, $\tau_c >> 1/\omega_o$ and thus spectral density is concentrated around the zero frequency point $J(0)$. Local magnetic fields oscillating at 0, ω_o, or $2\omega_o$ are capable of stimulating transitions from high energy to low energy states, and thus contribute to both T_1 and T_2 relaxation rates ($R_1 = T_1^{-1}$, $R_2 = T_2^{-1}$ respectively).[121] These contributions are defined by Equations 1.13 – 1.15 for homonuclear systems.

$$(Eq.1.13)\quad R_1 = K^2[J(0) + 3J(\omega_o) + 6J(2\omega_o)]$$

$$(Eq.1.14)\quad R_2 = K^2\left[\frac{5}{2}J(0) + \frac{9}{2}J(\omega_o) + 3J(2\omega_o)\right]$$

$$(Eq.1.15)\quad K = \frac{\gamma^2\hbar}{2r^3}$$

Equations 1.13 – 1.15 shed light onto the behavior of relaxation properties as molecular motion is varied. The R_1 dependency on the zero-frequency component $J(0)$ is weak, which is indicative of the promotion of T_1 by increased molecular motion. As T_1 relaxation depends on interactions between excited spins and lattice components to induce excited state relaxation, this suggests that increased molecular motion increases the probability of such interactions occurring and that T_1 relaxation is inefficient in large molecules like proteins. The R_2 dependency on $J(0)$ on the other hand is significantly enhanced in comparison to R_1 and is the dominant determinant of T_2 relaxation in proteins. As decreased τ_c only modestly increases $J(\omega_o)$ and $J(2\omega_o)$, due to increased spread of spectral density in faster regimes, T_2 relaxation is therefore fastest in large, slowly tumbling molecules, and is the primary factor responsible for loss of NMR signal and resolution proteins experience in comparison to small molecules. In addition, as experiments become more complex, longer, and go through more magnetization transfers, the effect of T_2 relaxation on the amount of signal that can be detected by the time it reaches the end of the pulse sequence is decreased even further. Thus, simpler protein NMR experiments such as ^{1}H-^{15}N HSQC spectra tend to be the most well resolved and sensitive.

1.4.2.1 The heteronuclear single quantum coherence spectrum (HSQC)

1D ^{1}H NMR, where each proton in a sample analyte gives rise to a characteristic chemical shift representative of its local environment, is the most frequently used form of NMR when

working with small molecules. As mentioned previously, the sheer number of protons in a protein gives rise to information-dense, poorly resolved spectra that are difficult if not impossible to parse. Thus, protein NMR experiments almost exclusively make use of multidimensional experiments that relay correlations between coupled nuclei, both reducing spectral overlap and yielding information about atomic connectivity in proteins. Though a wide variety of multidimensional NMR experiments designed for protein NMR exist, the most widely used is the ^{1}H-^{15}N heteronuclear single quantum coherence spectrum, or HSQC. In this spectrum, directly detected proton chemical shifts are correlated to that of directly-bonded ^{15}N nuclei, giving rise to a peak in a ^{1}H-^{15}N plane at the intersection of the ^{1}H and ^{15}N chemical shifts.[122] The resulting NH peaks arise primarily from the protein backbone, where each residue is expected to produce one peak, save proline. Additional peaks arising from the side-chain NH groups of tryptophan, asparagine, glutamine, and more uncommonly lysine, arginine, and histidine can also be detected in HSQCs.

As the pattern of peaks in an HSQC arises from the local environments of each NH pair, each protein gives rise to a unique spectrum. Due to the small number of *J*-coupling-mediated magnetization transfer steps needed to acquire chemical shifts for both ^{1}H and ^{15}N, it is also amongst the most sensitive multidimensional protein NMR experiments. The HSQC is therefore widely viewed as an excellent diagnostic spectrum. As chemical shifts are a function of local environment, any factors that alter this local environment can therefore detected in this spectrum, making it a useful tool in the assessment of protein structural integrity. When a protein is stably folded, differences in environment between the solvated surface and hydrophobic core as well as between different secondary structure elements give rise to a spectrum with a wide range of ^{1}H chemical shifts. In an unstable protein on the other hand, increased solvation and reduced secondary structure diversity causes a collapse of the HSQC spectrum over a narrow range of ^{1}H

chemical shifts that is characteristic of unfolded regions in proteins.[123] In a similar manner, events such as ligand binding, conformational switching, and protein-protein interactions change chemical shifts for nearby residues, allowing for identification of the regions of the protein involved in the event being studied.

Several dynamic processes can also be studied directly using HSQC and HSQC-derived spectra. Slow dynamics, which are generally correlated to larger-amplitude motions in proteins, cause alterations in the local chemical environment of nearby residues that sense the motion. If this exchange is slow enough, generally on a millisecond to second timescale,[124,125] peak doubling may be observed, where unique protein conformations each give rise to a distinct population of peaks. As motions accelerate, the peaks produced by the different conformers begin to coalesce into a broadened peak. This dynamic regime can also cause the disappearance of peaks from the spectrum if the broadening effect causes the peak intensity to drop to the noise level. Dynamics along these timescales however can be directly quantified in HSQCs using lineshape analysis, which relies on measurements of peak widths and shapes to derive kinetic parameters for the motions exhibited. Though faster timescale motions are also picked up in HSQCs, as rapid motions that slow T_2 relaxation lead to peaks of increased intensity, kinetic parameters at these timescales cannot be extracted directly from HSQC spectra.

1.4.2.2 Atom chemical shift assignments and associated spectra

Most NMR analyses of proteins rely on the characterization of local effects as reported by specific atomic correlations. However, in order to do this, it is crucial to be able to identify which atoms are giving rise to the effects observed. The process by which chemical shifts are assigned to atoms is thus critical to detailed protein NMR analyses. This process typically begins by assignment of backbone amide correlations, as backbone connectivity is generally less ambiguous

than side-chain connectivity due to the greater number of sequential correlations that can be linked along the backbone compared to a single side-chain. To establish connectivity using standard methods, two different 3D NMR spectra are generally used; a CBCA(CO)NH spectrum, which correlates the backbone NH group of a residue to the C^{α} and C^{β} atoms of the previous residue in the protein's primary sequence, and an HNCACB, which correlates the backbone NH group of a residue to both its own C^{α} and C^{β} atoms and to those of the preceding residue (Fig. 1.7).[126-128] Through shared C^{α} and C^{β} correlations amongst neighboring residues, as well as statistical distributions of H, N, C^{α}, and C^{β} chemical shifts by amino acid, we can establish backbone connectivity throughout the protein, and assign chemical shifts for these four atoms (Fig. 1.8).

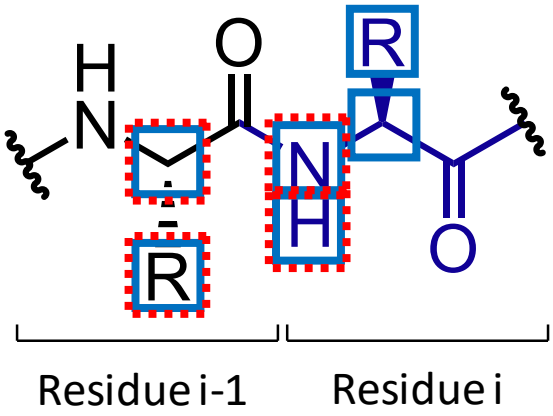


Figure 1.7. Correlated atoms in backbone assignment spectra. For a given backbone NH group, HNCACB correlations are shown as blue boxes, and CBCA(CO)NH correlations are shown as dotted red boxes. HNCACB correlations establish connectivity between adjacent residues, while CBCA(CO)NH correlations act as a control to identify which of the HNCACB peaks are inter-residue correlations and aid in the identification of overlapped or weak peaks.

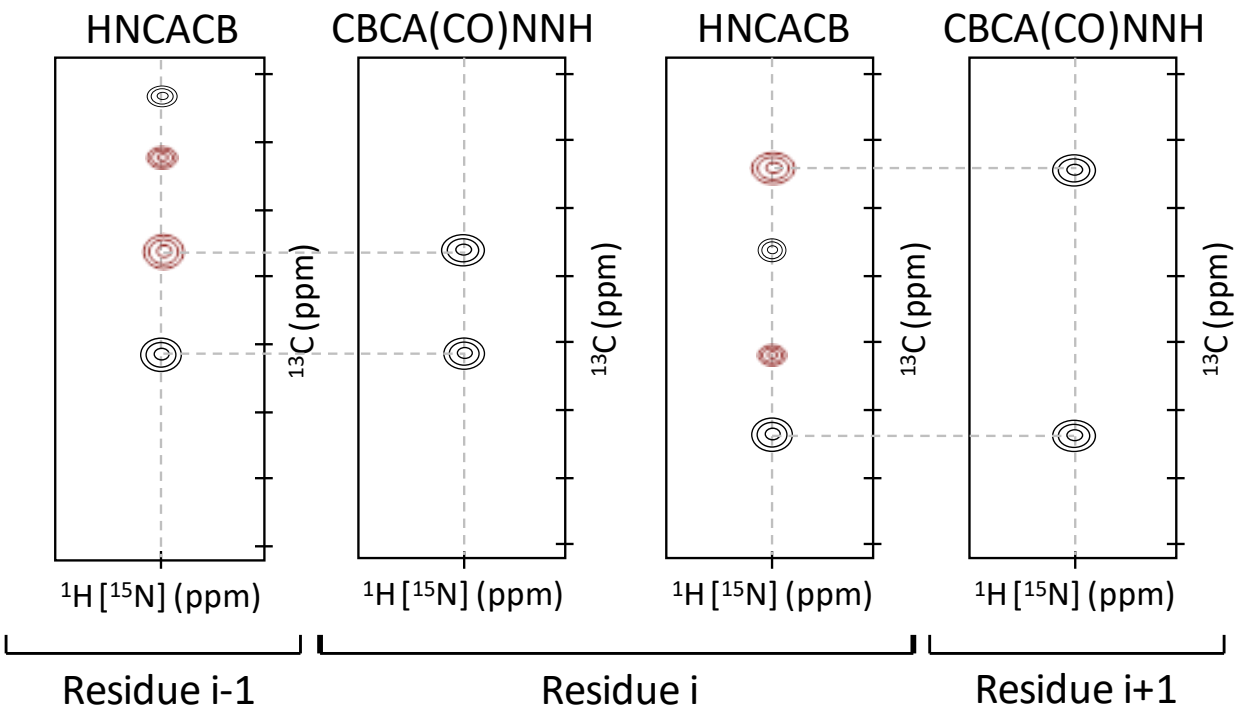


Figure 1.8. Residue connectivity as detected by backbone assignment spectra. For a given residue i, inter-residue correlations are identified in the CBCA(CO)NH spectrum and are matched to the C^{α} and C^{β} residues in the HNCACB spectrum of residue i−1. This proceeds iteratively down the backbone until all assignable residue have been linked. In a standard HNCACB spectrum, a negative intensity for C^{β} peaks is observed, as shown by a red coloration in this example. The strip plots shown herein display ^{13}C chemical shifts stemming from residue-dependent ^{1}H-^{15}N backbone amide correlations.

Once backbone NH, C^{α}, and C^{β} atoms have been assigned, other side-chain atoms can be assigned. The aliphatic portion of side-chains is typically assigned through a combination of COSY and TOCSY experiments.[129] In a COSY spectrum, side-chain CH groups are correlated only to their neighbors, while in a TOCSY, the correlation extends to the entire aliphatic portion of the side-chain. The use of several multidimensional spectra is standard, using combinations of C-, N-, and H- referenced spectra, such as CCH-COSY, HCCH-COSY, NH-TOCSY, CCH-TOCSY, and HCCH-TOCSY spectra in tandem, to deconvolute overlapped signals belonging to residues with similar chemical shifts, given the large number of CH groups typically found in

proteins. With N, H, C^{α} and C^{β} chemical shifts already known, and different amino acids giving rise to characteristic peak patterning, the remaining aliphatic chemical shifts can thereby be assigned through a similar chemical shift matching protocol down the side-chain as with backbone assignments. Aromatic residue chemical shifts on the other hand are not assignable through these COSY and TOCSY experiments due to drastically different chemical shifts and coupling constants in aromatic groups. Instead, HBCBCGCDHD and HBCBCGCDCEHE experiments can be used to assign aromatic side-chains, although these experiments are insensitive. Alternately, as was done in the projects presented in this thesis, if structural information about the protein is already known, NOE spectra can be used to assign aromatic atoms based on proximity to other assigned atoms.

1.4.2.3 Determining protein structures by NMR

Though we have so far primarily discussed NMR experiments that make use of through-bond correlations in proteins, there exist other experiments that make use of the Nuclear Overhauser Effect (NOE) to probe through-space correlations.[130,131] NOEs are the transfer of spin polarization from one population of spin-active nuclei to another through bond-independent dipole-dipole interactions. Though their relationship to internuclear distance is complex, with dependencies to spin diffusion, relaxation rate, and molecular motion, they are nonetheless semi-quantitative, and provide an estimate of distance from one atom to other proximal atoms given an r^{-6} dependence. So long as the majority of ^{1}H chemical shifts are assigned, these NOEs can be used to build a map of distance restraints that provides information on the 3D structure of the protein.

In parallel to NOEs, which provide distance restraints, NMR chemical shifts also provide information on dihedral angles in proteins. Though imprecise due to effects of the local chemical environment on chemical shifts, the statistical distribution of chemical shifts for each amino acid

is at least partially a function of its dihedral angles. Statistics-based prediction algorithms can thus predict dihedral angle ranges for protein φ, ψ, ω, and χ_1 dihedral angles.[132-134] With both torsional and distance restraints, the structure of a protein sequence can be determined by a fitting algorithm that attempts to find a physically reasonable structure that satisfies as many restraints as possible.[135] As this is a stochastic approach rather than a density-based method such as fitting a structure to an electron density map in crystallography, multiple possible solutions are generated by the fitting algorithm. NMR structures are therefore reported as ensembles, where ensemble diversity correlates to the number and quality of restraints available to fit that region of the structure.

1.4.2.4 Studying protein dynamics by NMR

Though NMR is commonly used to study protein interactions and solve protein structures, it is also capable of probing protein dynamics at an atomic level over a wide range of timescales, making it one of the most powerful techniques available for the *in vitro* study of protein dynamics.[124,125,136] Various NMR experiments have been developed that provide information on processes ranging from the second timescale to the picosecond timescale (Figure 1.9), though we will focus primarily on the techniques used in the projects presented herein.

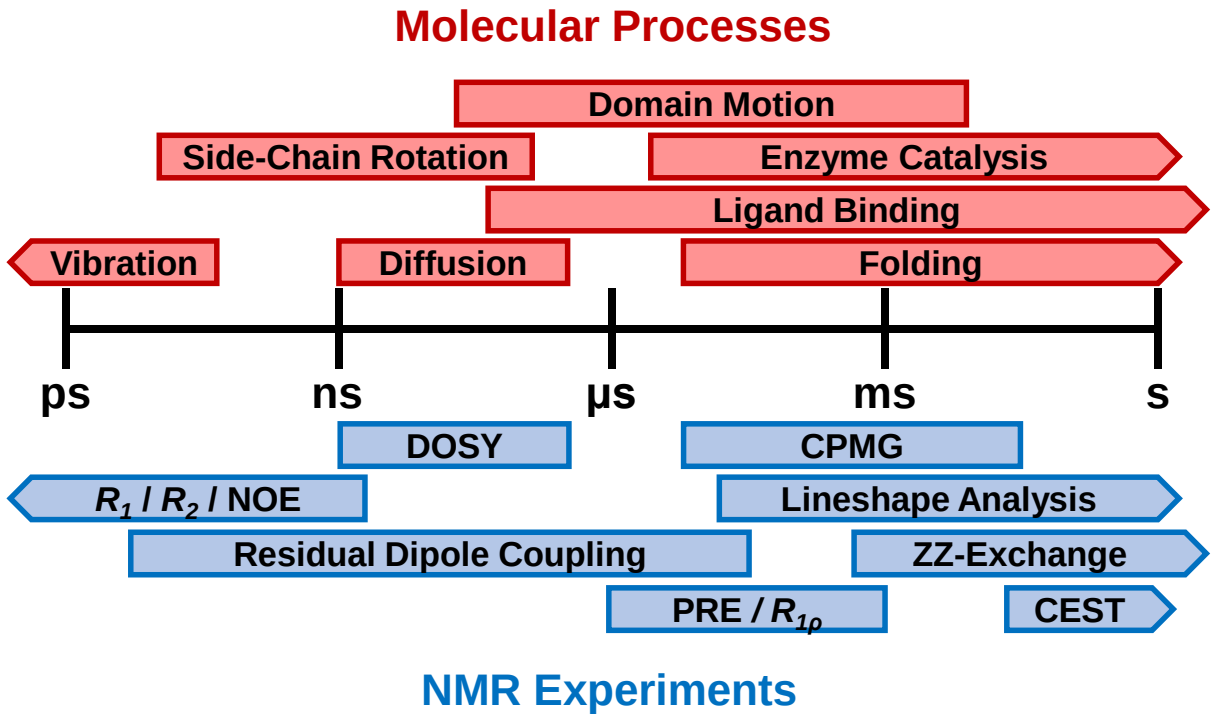


Figure 1.9. Timescale comparison between NMR experiments measuring dynamics and molecular processes. Dynamic molecular processes occurring in proteins are shown in red boxes against a logarithmic timescale axis. NMR experiments probing dynamics are shown in blue. In both cases, the typical timescale range involved in these processes or considered by these experiments is shown, thus outliers are not accounted for.[20]

1.4.2.5 ZZ-Exchange

At the slow end of the scale of protein dynamics, on the microsecond to second timescale, are larger amplitude molecular motions such as protein folding and unfolding, domain motions, and conformational rearrangements. Depending on the amplitude of motion involved, these processes nonetheless range over several timescale orders of magnitude. On the millisecond to second timescale, motion may be slow enough to give rise to two distinct, resolvable populations in an NMR spectrum, potentially allowing the use of ZZ-exchange experiments to characterize these motions.[137] In a ZZ-exchange spectrum, chemical shifts are acquired for ^{15}N and ^{1}H much like in an HSQC. However, a mix time is introduced between the indirect detection of the ^{15}N chemical

shift and transfer of magnetization back to ^{1}H for detection. During this mix time, which typically ranges from 10 ms to 1 s or longer, the protein is allowed to freely undergo conformational exchange. Thus, during acquisition, the ^{15}N resonance from one conformation will be coupled to the ^{1}H resonances of both conformations, in a ratio that is proportional to both the length of the mixing period and the rate of conformational exchange. These couplings give rise to four peaks per residue; two peaks arising from ^{15}N chemical shifts coupled to their corresponding ^{1}H chemical shifts from the same protein conformation, termed diagonal peaks, and two peaks arising from ^{15}N chemical shifts coupled to the ^{1}H chemical shift from the alternate protein conformation, termed crosspeaks. As mixing time is increased, the crosspeak to diagonal peak ratio increases in kind, until mixing time is long enough for conformational equilibrium to be reached. Exchange curves can then be fitted to a set of ZZ-exchange-derived peak intensities versus mixing times to determine the forward and reverse rates of exchange. Simultaneously with exchange however, relaxation takes place, resulting in decreased signal over time. Thus, the modeled fits must account both for forward and reverse exchange rates, which are proportional to an increase in crosspeak intensity and to a decrease in diagonal peak intensity as a function of mix time, and relaxation rates for both conformational states, which are inversely proportional to the intensity of all four peaks as a function of mix time. The resulting set of four equations, each detailing the dependence of one peak's intensity on mixing time, thus depends on four variables, k_{ij} and k_{ji}, which are respectively the forward and reverse rates of exchange between conformational states i and j, as well as R_i and R_j, which are the relaxation rates for conformational states i and j respectively. Diagonal peak intensities are defined by Equations 1.16 and 1.17, while crosspeak intensities are defined by Equations 1.18 and 1.19. $I_{ii}(T)$ and $I_{jj}(T)$ represent the diagonal peak intensities for conformational states i and j with mixing time T, $I_{ij}(T)$ and $I_{ji}(T)$ represent the crosspeak intensities for

conformational exchange from state *i* to *j* and *j* to *i* respectively over a mixing time *T*, $I_i(0)$ and $I_j(0)$ represent the diagonal peak intensities for conformational states *i* and *j* with a mixing time of 0 s, a_{21} and a_{12} correspond to $-k_{ij}$ and $-k_{ji}$ respectively, and a_{11}, a_{22}, λ_1, and λ_2 are defined by Equations 1.20 – 1.23.

$$(Eq.\,1.16)\quad I_{ii}(T) = I_i(0) \times \frac{\left(-(\lambda_2 - a_{11})e^{-\lambda_1 T} + (\lambda_1 - a_{11})e^{-\lambda_2 T}\right)}{(\lambda_1 - \lambda_2)}$$

$$(Eq.\,1.17)\quad I_{jj}(T) = I_j(0) \times \frac{\left(-(\lambda_2 - a_{22})e^{-\lambda_1 T} + (\lambda_1 - a_{22})e^{-\lambda_2 T}\right)}{(\lambda_1 - \lambda_2)}$$

$$(Eq.\,1.18)\quad I_{ij}(T) = \frac{I_i(0)\left(a_{21}e^{-\lambda_1 T} - a_{21}e^{-\lambda_2 T}\right)}{\lambda_1 - \lambda_2}$$

$$(Eq.\,1.19)\quad I_{ji}(T) = \frac{I_j(0)\left(a_{12}e^{-\lambda_1 T} - a_{12}e^{-\lambda_2 T}\right)}{\lambda_1 - \lambda_2}$$

$$(Eq.\,1.20)\quad a_{11} = R_n + k_{ij}$$

$$(Eq.\,1.21)\quad a_{22} = R_u + k_{ji}$$

$$(Eq.\,1.22)\quad \lambda_1 = \frac{1}{2}\left[(a_{11} + a_{22}) + \left((a_{11} - a_{22})^2 + 4k_{ij}k_{ji}\right)^{0.5}\right]$$

$$(Eq.\,1.23)\quad \lambda_2 = \frac{1}{2}\left[(a_{11} + a_{22}) - \left((a_{11} - a_{22})^2 + 4k_{ij}k_{ji}\right)^{0.5}\right]$$

As ZZ-exchange directly reports exchange rates for a conformational exchange between two states, this experiment can also be repeated at different temperatures and fit to Eyring's equation to find free energy, enthalpy, and entropy at the transition state, and thus the height of the energy barrier to exchange.

1.4.2.6 CPMG Relaxation Dispersion

Moving towards faster timescales and reaching faster catalytic, folding, and binding processes, as well as smaller amplitude domain motions, ZZ-exchange becomes less useful as crosspeak intensity saturates too rapidly to be measurably detected, and eventually becomes unusable once exchange rates are fast enough to cause diagonal peaks to converge. In this timescale range, around 100 μs to 10 ms, CPMG relaxation dispersion becomes the technique of choice. CPMG can also be used to study slower motions involving weakly populated states that are not sufficiently occupied to give rise to a quantifiable set of peaks in ZZ-exchange.[138] CPMG experiments detect the transverse relaxation rate, R_2, as a function of a CPMG field strength, and are specifically designed to detect the chemical exchange component of R_2, thus providing quantitative information about micro- to millisecond timescale motions such as conformational equilibrium.[139] Much like ZZ-exchange experiments, CPMG experiments are based on a modified ^{1}H-^{15}N HSQC with a mix time introduced into the pulse sequence. Unlike in ZZ-exchange, the CPMG mix time is of a constant duration. Over the course of this period, periodic 180° pulses are applied to the sample, refocusing magnetization that has become dephased due to conformational exchange over the mixing time. As the frequency of these 180° pulses (CPMG frequency or CPMG field strength, ν_{CPMG}) increases, the effect of chemical exchange on R_2 is averaged out, resulting in a smaller R_2 and thus a reduction in chemical-exchange mediated peak broadening. Mathematically, the relationship between ν_{CPMG} and R_2 is known, thus by calculating R_2 from peak intensities at various ν_{CPMG}, we can plot R_2 vs ν_{CPMG} and fit the relationship to the resulting data. Though the general relationship is complex, it is typically simplified to fit one of two regimes, depending on whether the exchange rate is greater or less than the $\Delta\omega$ between conformational

states, as defined by Equation 1.24 for fast exchange ($k_{ex} >> \Delta\omega$) and by Equation 1.25 for slow exchange ($k_{ex} << \Delta\omega$).

$$(Eq.\,1.24)\quad R_2 = R_2^0 + \frac{p_A p_B \Delta\omega^2}{k_{ex}}\left(1 - \frac{4\nu_{CPMG}}{k_{ex}}\tanh\left(\frac{k_{ex}}{4\nu_{CPMG}}\right)\right)$$

$$(Eq.\,1.25)\quad R_2 = R_2^0 + k_{AB}\left(1 - \frac{\sin(\Delta\omega\,\tau_{CPMG})}{\Delta\omega\,\tau_{CPMG}}\right)$$

In Equations 1.24 and 1.25, R_2^0 is the exchange-independent component of the transverse relaxation rate, p_A and p_B are the fractional populations of states A and B in solution ($p_B = 1 - p_A$), k_{ex} is the sum of exchange rate from A to B (k_{AB}) and from B to A (k_{BA}), τ_{CPMG} is the CPMG pulse train delay (the delay between 180° pulses is $2\tau_{CPMG}$), and R_2 is the measured CPMG field-modified relaxation rate, obtained by transforming peak intensities as shown by Equation 1.26, where T_{CPMG} represents the length of the CPMG block, $I(0)$ is a reference signal intensity at time zero, and $I(T_{CPMG})$ is the signal intensity measured following the CPMG block.

$$(Eq.\,1.26)\quad R_2 = -\left(\frac{1}{T_{CPMG}}\right)\ln\left(\frac{I(T_{CPMG})}{I(0)}\right)$$

Though it is possible to fit these equations to a single set of CPMG data, it can be very difficult to determine which equation is most appropriate, and several variables are being fit at once using a single equation. It is therefore common to run CPMG experiments at two different external magnetic field strengths. Exchange rates and populations are independent of field strength, while $\Delta\omega$ is field-dependent. Thus, a second set of data allows us to better fit the parameters by reducing the number of degrees of freedom to be satisfied.

1.4.2.7 Lineshape Analysis

An alternative to ZZ-exchange and CPMG relaxation dispersion at microsecond to second timescales is lineshape analysis. Unlike the other techniques presented so far, lineshape analysis uses a basic, unmodified ^{1}H-^{15}N HSQC. In any NMR spectrum, if the interscan delay is long enough to allow complete relaxation of bulk magnetization to equilibrium, the resulting peak shapes are characteristic of population distributions and exchange rates in solution (Fig. 1.10).[124,125] Under the assumption that relaxation rates are affected only by global protein tumbling, the relationship between peak width, fractional populations for the dynamic states being considered (p_A, p_B), and the chemical shifts of each state (δ_A, δ_B) is known. Though the general case equations are complex and will not be discussed here, several fitting algorithms exist for lineshape data that deconvolute partially overlapped NMR data, fit Lorentzian or Gaussian models to peaks, and determine kinetic parameters from these.[140] However, not all parameters can be accurately fit with no prior knowledge of the system. For slow exchange regimes, populations are separately resolved, fractional populations and chemical shifts can be directly extracted from the spectrum and used to determine exchange rates. For intermediate exchange regimes where the peaks have coalesced, this is however not possible. Therefore, δ_A and δ_B are typically determined through variable temperature measurements to slow exchange and resolve individual peaks for the two species to improve the accuracy of fits. As temperature dependency of chemical shifts is generally linear,[141] with a few measurements of δ_A and δ_B at low temperature, values can be extrapolated back for chemical shifts at the desired working temperature. Finally, for fast regimes, where peak broadening from transverse relaxation is counteracted by local motion that is faster than the global protein tumbling rate, lineshape analysis ceases to be capable of extracting a

quantitative k_{ex}. However, the peak shapes observed still inform us about local dynamics, where increasingly rapid motions lead to further increased peak intensity.

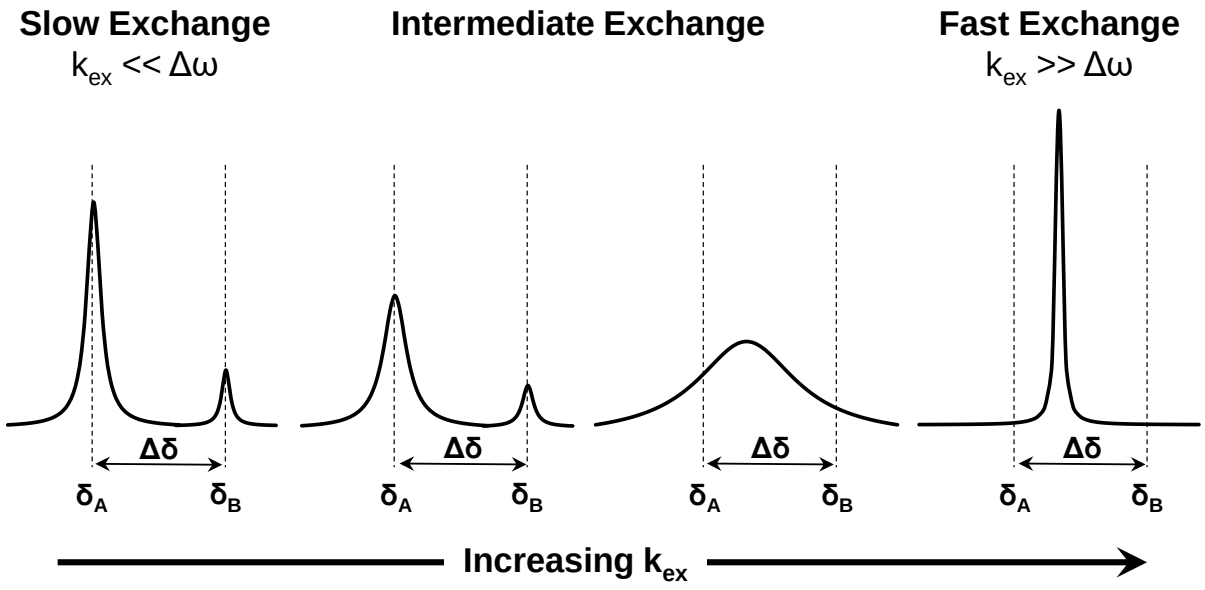


Figure 1.10. Effects of conformational exchange rate on NMR peak lineshapes. Example peak lineshapes for a residue experience two conformations, A and B as a function of exchange rate k_{ex}. As exchange rate increases, peaks for populations A and B broaden, then coalesce. In fast regimes, rapid motions reduce relaxation rates and increase signal. Lineshape analysis cannot quantitatively determine k_{ex} in a fast exchange regime.

1.4.2.8 Measurement of T_1 and T_2 Relaxation Rates

The final NMR-based dynamics characterization technique that will be discussed herein is used to characterize dynamics at the nanosecond to picosecond timescale. Unlike other techniques discussed, motions on these timescales are generally not described as interconversion between concrete states, but rather as quasi-harmonic fluctuations within a single energy well. As such, experiments probing motions on this timescale can provide an indirect metric of slower protein motions. Since local protein motions with rates comparable to or slower than that of molecular tumbling has a quantifiable effect on R_1 and R_2 relaxation rates.[142] In rigid residues on this timescale, magnetic field anisotropy generated by the local chemical environment is partially

averaged out by global molecular tumbling. In fast-moving residues, dynamic motions further contribute to this averaging of anisotropy, leading to increased T_2 relaxation times. T_1 relaxation on the other hand is the result of a loss of spin energy back to the surrounding lattice until equilibrium is restored. As increased motion increases the probability of spin interactions with lattice elements that stimulate energetic transitions back to the ground state, increased motion therefore generally correlates with decreased T_1 relaxation time for the range of correlation times that are common to proteins.[118] Together, T_1 and T_2 relaxation times can be used to approximate apparent correlation times ($\tau_c^{app} \propto T1 / T2$) throughout the protein structure, which are measurements of apparent tumbling and thus linked to fast dynamics. Experimentally, modified ^{1}H-^{15}N HSQCs are used to measure T_1 and T_2 throughout the protein backbone as a decay in signal intensity I over a delay period t.[143] In proteins, as T_2 dominates global relaxation of transverse magnetization, T_1 determination experiments are carried out via inversion recovery to minimize the effect of T_2 on relaxation rates, and T_2 determination is carried out using a series of refocusing pulses to exclude the contribution of chemical exchange to T_2. The resulting intensities from both experiments are then plotted as a function of mix time, and fit to a monoexponential decay characteristic of relaxation for each of T_1 and T_2, as defined by Equations 1.26 and 1.27 respectively and where I(t) is the measured peak intensity following a relaxation period of length t, and I_0 is the reference peak intensity with no relaxation delay. Once T_1 and T_2 are obtained, they can then be converted to a correlation time through Equation 1.28.

$$(Eq.1.26) \quad I(t) = I_0 e^{-\frac{t}{T_1}}$$

$$(Eq.1.27) \quad I(t) = I_0 e^{-\frac{t}{T_2}}$$

$$(Eq.\,1.28)\quad \tau_c^{app} \approx \frac{1}{4\pi\nu_N}\sqrt{6\frac{T_1}{T_2}-7}$$

In regions of the protein where dynamic motions are slower than global protein tumbling, τ_c^{app} is expected to reflect the global tumbling rate, although it is also subject to inflation by T_2 relaxation components stemming from slower chemical exchange processes that are not fully suppressed by the CPMG train during T_2 determination. In regions of the protein where dynamics are faster than global tumbling rate, a depression of τ_c^{app} proportional to the frequency of these motions can be observed, allowing for the identification of regions that are undergoing fast protein dynamics. Though T_1 and T_2 relaxation measurements are sufficient to characterize fast dynamics in proteins, heteronuclear NOE (hetNOE) measurements are also often reported as an additional measure of fast dynamics. As the efficiency of the NOE is a function of several factors including geometry and motion, studying the NOE between ^{1}H and ^{15}N in the protein backbone provides information about the motion of individual N-H bond vectors. When local motion is rapid in comparison to tumbling rate, NOE intensity is decreased relative to the average. To measure hetNOEs, a modified HSQC is used in which proton saturation is achieved during the relaxation delay prior to the starting 90 ° pulse, and the relative ratio of peak intensities for spectra with ($I_{saturated}$) and without ($I_{equilibrium}$) this proton saturation gives rise to the hetNOE measurement through Equation 1.29.

$$(Eq.\,1.29)\quad hetNOE = \left(\frac{I_{saturated}}{I_{equilibrium}} - 1\right)$$

Overall, NMR provides a thorough and high-resolution means to study protein dynamics in relatively small proteins. However, due to the contributions of molecular size to tumbling rate and therefore to transverse relaxation rates, it becomes less sensitive as protein size increases. In addition, due to its low sensitivity, high concentrations of isotopically-labeled protein are

necessary to acquire high quality NMR spectra. For this reason, NMR spectroscopy is usually reserved for detailed characterization of few proteins, and other less time- and reagent-intensive techniques such as circular dichroism are typically used for a more diagnostic characterization of protein structure before moving on to NMR.

1.4.3 Circular dichroism spectroscopy (CD)

Circular dichroism (CD) is a measurement of the differential absorption of left- and right-handed circularly polarized light by a chiral analyte. When plane-polarized light, comprising both left- and right-handed circularly polarized light components, encounters a chiral chromophore, this differential absorption gives rise to measurable ellipticity, typically represented as an angle with units of millidegrees (mdeg).[144] As this raw signal is a function of chromophore concentration and optical path length, it is standard to normalize it against these values. In proteins, both residue side-chains and peptide bonds can act as chromophores, depending on the region of the CD spectrum being studied, thus it is commonplace to use the concentration of peptide bonds in the sample to normalize ellipticity, giving rise to a mean residue ellipticity (MRE or θ) using units of $deg \cdot cm^2 \cdot dmol^{-1}$ as a standard. So treated, MRE can then be used to rapidly and quantitatively study the structure of target proteins. When CD spectroscopy is done on proteins, two primary regions are generally considered. Near UV measurements at wavelengths of 260 – 320 nm report primarily on protein tertiary structure and amino acid mobility through the absorbance of Phe, Tyr, and Trp aromatic side-chains.[144-146] On the other hand, far UV CD measurements at wavelengths of 180 – 250 nm, which we will focus on in this thesis, are the most commonly used for CD measurements on proteins, and report on protein secondary structure through $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ electronic transitions within protein backbone amides.[147,148]

As far UV CD spectroscopy signals are generated by the UV absorbance of amide bonds, they report primarily on the geometry of the adjacent φ and ψ backbone dihedral angles, and thus are closely related to protein secondary structure.[148] α-Helices, β-strands, turns, and random coil backbone configurations each give rise to characteristic CD spectra (Fig 1.11), while the spectrum produced by an entire protein is represented by an additive combination of the spectra produced by each of its chromophores.[149] Though deviations from ideality in the secondary structure elements of most proteins are reflected in their CD spectra, a linear combination of characteristic spectra for the various secondary structure elements can nonetheless be fitted to the experimental spectrum obtained, providing an estimate of the protein's secondary structure content. However, to reduce error stemming from non-ideality, deconvolution algorithms such as SELCON,[150] VARSLC,[151] and CONTIN,[152] which approximate secondary structure elements using reference datasets derived from protein structure databases rather than less realistic idealized peptides, are commonly used. Even so, while contributions from secondary structure dominate the far UV CD spectrum, other features such as tertiary structure and aromatic side-chains can also contribute to the measured signal,[152,153] resulting in a large degree of uncertainty in the prediction of the absolute secondary structure content of a protein.[154,155]

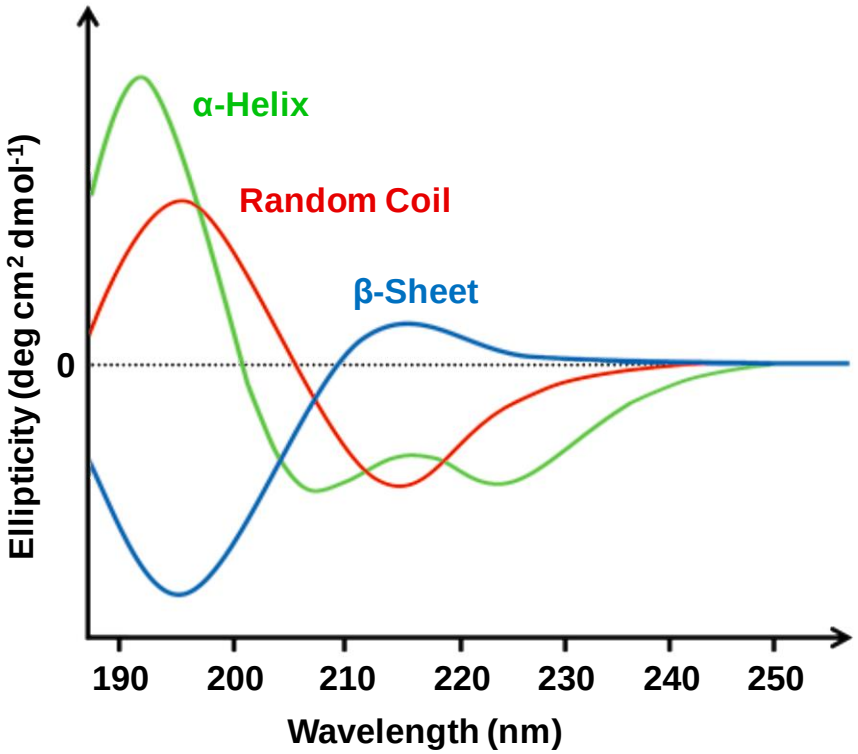


Figure 1.11. Characteristic CD spectra for idealized secondary structure elements. Far UV CD spectra for three primary secondary structure elements are shown, with an α-helix spectrum in green, a β-sheet spectrum in blue, and a random coil spectrum in red.

Where CD shines however is in the detection of conformational changes in otherwise closely related proteins. Though a CD signal is, as mentioned previously, a complex combination of several elements, it is extremely sensitive to changes in secondary structure.[156] CD spectroscopy can therefore be used to detect stimuli that alter a protein's conformation, such as the binding of a ligand,[157,158] or as in this thesis, to confirm that a mutant of a protein retains a similar fold.[159] Though its resolution is far inferior to that of atomistic techniques like X-ray crystallography or NMR, CD spectroscopy is also a simpler technique in terms of both experimental runtime and data analysis. This makes it an ideal platform for multi-point analyses such as protein thermal denaturation curves,[160] and secures CD a critical spot in the toolbox of biophysical tools at our disposal for the study of protein structure.

1.5 Thesis objectives

In Chapter 1, we discussed the link between protein structure, dynamics, and function, as well as limitations in treating protein motion in computational protein design. As many proteins depend on motion to accomplish their functions, keeping dynamics removed from CPD hinders our capacity to design complex functional proteins, and thus this thesis treats on the inclusion of dynamics in CPD, as well as the characterization of dynamics in two protein systems that are highly relevant to the development of CPD methods capable of handling and designing protein dynamics. In Chapter 2, we outline a new CPD methodology, *meta*-MSD, which utilizes large conformational ensembles in a *meta*-analysis to design dynamics into the Gβ1 model protein. Through *meta*-MSD, we developed Gβ1 mutants exhibiting novel conformational exchange termed DANCERs, which were then characterized via NMR spectroscopy to validate the accuracy of our predictions. In Chapter 3, we go on to characterize the effects of each mutation present in one DANCER on the designed exchange trajectory to understand how these mutations gave rise to these novel dynamics and how future designs for dynamics might be guided. In Chapter 4, we characterize structural dynamics in a family of RFPs to guide an upcoming novel computational protein design of functional protein dynamics where we will redesign dynamics in the RFP barrel to rigidify the RFP chromophore and enhance its brightness. In Chapter 5, we discuss the implications of *meta*-MSD to CPD as well as potential future steps to take to further the understanding of the protein energy landscape and the design of protein dynamics.

Chapter 2: Rational design of proteins that exchange on functional timescales

James A. Davey[‡], Adam M. Damry[‡], Natalie K. Goto[*] & Roberto A. Chica[*]

[‡]Co-first authors

[*]Co-corresponding authors

Initially published in: Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A. Rational design of proteins that exchange on functional timescales. Nat Chem Biol ***13****, 1280-1285 (2018)*

2.1 Statement of contribution

James A. Davey performed all computational experiments. Adam M. Damry and James A. Davey performed biophysical characterization experiments. Adam M. Damry performed all NMR experiments. Natalie K. Goto and Adam M. Damry designed NMR experiments and analyzed data. James A. Davey and Roberto A. Chica conceived the project, designed computational experiments and analyzed data. All authors wrote the manuscript.

2.2 Introduction

Proteins are the molecular machines of life, carrying out complex physical and chemical processes that often require concerted motions of local structural elements. Previous efforts to design new proteins for applications in research, industry, and medicine have focused on the creation of sequences that stably adopt a single target structure, ignoring the potential impact of protein dynamics in function. Although computational protein design (CPD) has enjoyed considerable success in creating new proteins using this approach,[91,105,161-163] most have failed to match the efficiencies that are found in nature.[97,164,165] This result suggests that fundamental aspects of protein structure that are not currently considered in design strategies still remain to be incorporated in order to approach the efficacy of naturally occurring systems. Given the demonstrated importance of dynamics in a range of protein functions including enzyme

catalysis,[32,166] allosteric regulation,[167] and molecular recognition,[168] rational design of a defined mode of dynamics into a protein fold has great potential to address this shortfall. Moreover, the ability to engineer a protein that can undergo conformational exchange between multiple states should expand the range of functionalities that can be designed, paving the way for applications that are currently inaccessible using natural proteins.

Rational design of protein dynamics requires the prediction of sequences that are able to adopt all conformations required for exchange between predefined states. The recent development of multistate design (MSD) approaches applicable to large structural ensembles[54,115,116] has provided a method for the evaluation of protein sequence energies in the context of a large number of possible conformational states. Thus, MSD can in principle be used to assess the energy landscape of a target protein and identify sequences that can exchange between distinct states. However, introduction of functionally-relevant conformational exchange into a stable protein fold is a difficult design problem as it requires *a priori* knowledge of the structural features of the relevant conformational states for dynamic exchange, including the intermediate states that the protein must adopt as it undergoes this conformational transition, which are often unknown. In addition, the multivariable optimization of sequences across many conformational states presents a significant computational challenge, since sequences must be designed that not only satisfy stability requirements for multiple target structures, but also yield an energy profile that would allow exchange between structures to occur on a functional timescale.

Herein, we have developed a general procedure that addresses these challenges and enables the rational design of protein dynamics, which we termed *meta*-multistate design (*meta*-MSD). *Meta*-MSD allowed the evaluation of protein conformational landscapes in order to predict sequences that can undergo spontaneous exchange between predefined states. We applied this

methodology to the design of sequences that adopt the global fold of Streptococcal protein G domain β1 (Gβ1) and spontaneously exchange between two conformations that have not been previously observed for this fold, providing the first demonstration of the rational design of dynamics into a specific protein target.

2.3 Results

2.3.1 *Meta*-MSD

A dynamic protein that spontaneously interconverts between two distinct conformational states adopts a continuum of unique configurations during exchange. However, the energy landscape is complex and the range of configurations that are sampled over the course of exchange cannot be completely defined. Nevertheless, we hypothesized that it should be possible to engineer a user-defined exchange trajectory by identifying sequences that stabilize configurations having structural characteristics postulated to facilitate this exchange. To simplify the exchange reaction coordinate, the conformational landscape can be conceptually divided into three states: a major, a minor, and a transition state. In the context of this work, we treated each of these states as a collection of unique configurations that we referred to as microstates. Microstates were generated by optimizing side-chain rotamers for predefined sequences on an ensemble of backbone templates using MSD, which also returned an energy value for each microstate that reflects its predicted stability (Fig. 2.1a–c). Following MSD, microstates were partitioned into their corresponding states according to their structural features (Fig. 2.1d), and the energy of each state calculated from the energy of its constituent microstates. This produced an energy profile for the exchange reaction for each sequence (Fig. 2.1e), with relative energy differences between states determining whether sequences would give rise to a dynamic or static protein (Fig. 2.1f). We called this framework *meta*-MSD because both state and dynamic behavior were assigned after rotamer optimization by

MSD. *Meta*-MSD can be used to identify sequences that stably populate the target major and minor states, with a transition state barrier that is small enough to create dynamic proteins that interconvert between these two states.

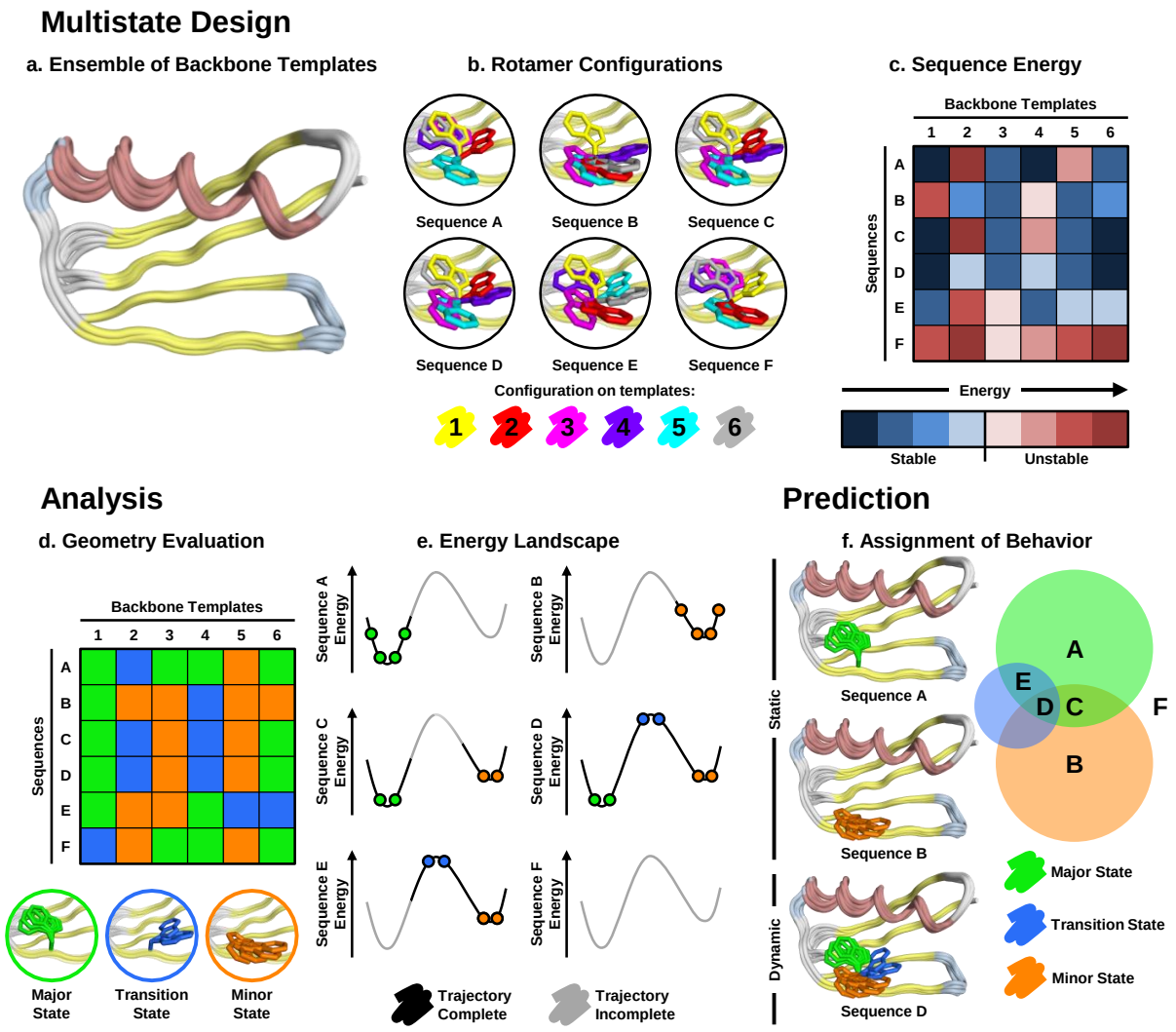


Figure 2.1. The *meta*-MSD framework for design of conformational exchange. Multistate design (MSD) with an ensemble of backbone templates approximating the conformational landscape for dynamic exchange between targeted states (a) was used to generate microstates by solving the lowest energy rotamer configuration for each sequence on each backbone template (b). MSD also returned an energy value for each microstate that reflects its predicted stability (c). Geometry-based analysis of the rotamer-optimized microstates (d) allowed assignment of each microstate to major, minor, or transition state regions of the energy landscape (e). Prediction of conformational dynamics was then done based on an evaluation of the relative energies of these states (f). For *meta*-MSD to predict a sequence as dynamic, all three states must be stable, with an energy profile that is compatible with exchange (e.g., sequence D). The schematic shows sequences A, B, C and E with predictions to produce static proteins because they either stabilize a single state or cannot stabilize the transition state. The prediction for Sequence F is for an unfolded protein because it is unstable on all states.

2.3.2 Design of conformational exchange into Gβ1

To validate our *meta*-MSD framework, we targeted the introduction of dynamics on a functionally relevant timescale (i.e., μs–ms)[21] into the structure of Gβ1. Native Gβ1 does not exhibit large amplitude motion on this timescale,[169,170] with a small size (56 amino acids) that facilitates characterization of its dynamic properties at atomic resolution. Additionally, Gβ1 possesses a single tryptophan residue (Trp43) that in high-resolution structures of Gβ1 and its natively folded variants[48,53,171-178] exclusively occupies a single side-chain conformation with χ_1 and χ_2 dihedrals of –74 ± 9° and +75 ± 11°, respectively. We named this conformation +*g*(–) due to its positive χ_2 dihedral angle and its *gauche*(–) χ_1 dihedral (Fig. S2.1). In Gβ1, the Trp43 side chain is mostly solvent-inaccessible, making intimate contacts with several residues that comprise the hydrophobic core. This makes it an attractive target for the design of conformational exchange, with one state being buried, and the other being excluded from the hydrophobic core in a solvent-exposed conformation. Exchange between such states is expected to involve the disruption of side-chain packing interactions and require some flexibility in backbone structure, which should be most effectively accommodated by collective motions that are characteristic of dynamics on functionally relevant timescales.[21,179] Moreover, with exchange of the tryptophan side chain set as our target for the design of dynamics, tryptophan side-chain dihedral angles provided a convenient metric for the assignment of microstates to one of the target states defined in our *meta*-MSD approach.

Using *meta*-MSD, we designed Gβ1 sequences that could adopt the native fold and also undergo conformational exchange between a state where the Trp43 indole is solvent-exposed [−*g*(+)] and a state where the indole is sequestered from the solvent in the hydrophobic core [–

g(−)] (Fig. S2.1e). Notably, we avoided selection of the native Trp43 conformation [+*g*(−)] for the core-buried state, since CPD has a tendency to overemphasize the stability of the native rotamer relative to non-native configurations.[180] A final and critical aspect of our conformational exchange design was the definition of an intermediate state with the Trp43 side chain in the *–t* conformation, since this state is necessary to provide a model of transiently populated microstates that are sampled along an exchange trajectory defined by rotation around the χ_1 dihedral of Trp43 (Fig. S2.1e). Use of the *–t* conformation as a proxy of the transition state thus allowed estimation of kinetic barriers between states along this defined exchange pathway, enabling the elimination of sequences predicted to stably adopt two end-states separated by large kinetic barriers that would not exchange on functional timescales.

To ensure adequate sampling of the range of structures that may be required to accommodate the designed conformational exchange, we generated an ensemble of 12,648 templates using a combination of several template generation procedures (Fig. S2.2, Table S2.1, Methods). Using this ensemble, we performed MSD to optimize rotamers on each template for a library of 1,296 Gβ1 sequences comprising combinations of core-residue mutations (Fig. S2.3) from experimentally verified folded Gβ1 variants.[54] MSD thus yielded >16 million microstates and corresponding energies (12,648 templates × 1,296 sequences), giving rise to an approximation of the accessible conformational landscape of Trp43 in the native Gβ1 fold.

Prior to *meta*-analysis of the MSD output, we eliminated sequences having a Boltzmann-weighted average of microstate energies greater than that of the wild-type sequence, since these are less likely to adopt a stable Gβ1 fold.[116] For the remaining 195 sequences, we then measured χ_1 and χ_2 dihedrals of the Trp43 side chain for each corresponding microstate allowing the assignment of each microstate to one of the six states defined in Figure S2.1c. After assigning all

microstates to states, for each sequence in the library we used the energy of the most stable microstate assigned to each state to represent the energy for that sequence and state. The resulting energy profile for each sequence constructed from these state energies (Fig. 2.1e), enabled identification of 35 sequences predicted to allow conformational exchange between the target core-buried [–*g*(–)] and solvent-exposed [–*g*(+)] conformations (Methods). Notably, *meta*-MSD performed with reduced-size ensembles comprising subsets of these 12,648 templates (Table S2.2) resulted in the prediction of fewer stable sequences (≤ 131) and of at most 3 dynamic sequences, an unsurprising result given that the use of larger ensembles in MSD decreases the number of false-negative predictions by addressing the fixed-backbone approximation more effectively than the use of a small ensemble.[181] Thus, we selected four of the 35 designed dynamic Gβ1 variants for experimental characterization and named them DANCERs, for *Dynamic And Native Conformational ExchangeRs* (Table 2.1).

Table 2.1. Predicted and Experimental Properties of Gβ1 variants

		Meta-MSD Predictions			Stability	Exchange [e]			
Protein	**Mutations**	**Behavior [a]**	**ΔE_{eq} [b] (kcal/mol)**	**$\Delta E^{\ddagger}$ [c] (kcal/mol)**	**ΔG_U [d] (kcal/mol)**	**k_{-1} [f] (s^{-1})**	**k_1 [g] (s^{-1})**	**$\Delta G^{\ddagger}$ [h] (kcal/mol)**	**ΔG_{eq} [i] (kcal/mol)**
Wild type		+*g*(−)	8.6	15.2	4.1 ± 0.2				
DANCER-0	Y3F/L5A/L7I/A34F/V39I	−*g*(+) ↔ −*g*(−)	0.9	7.8	1.5 ± 0.2				
DANCER-1	Y3F/L5A/L7I/A34F/V39L/V54I	−*g*(−) ↔ −*g*(+)	3.7	8.4	2.2 ± 0.1	30 ± 10	110 ± 50	18.9 ± 0.3	0.3 ± 0.1
DANCER-2	Y3F/L5A/L7I/A34F/V39L	−*g*(+) ↔ −*g*(−)	1.3	9.4	1.7 ± 0.1	[j]	[j]	[j]	1.4 ± 0.7
DANCER-3	Y3F/L7I/A34F/V39L/V54I	−*g*(−) ↔ −*g*(+)	2.9	13.7	2.0 ± 0.3	3.9 ± 0.2	23 ± 5	20.65 ± 0.08	1.3 ± 0.3
NERD-S	Y3F/L7I/A34F/V39I/V54I	−*g*(+)	4.3	14.7	2.7 ± 0.1				
NERD-C	Y3F/L7I/F30L/V39I	+*g*(−)	12.2	15.3	4.0 ± 0.3				

[a] Static variants (NERD-S and NERD-C) are predicted to occupy a single state while DANCER proteins are predicted to exchange between major and minor states (major state ↔ minor state)
[b] Energy difference between the two lowest energy states
[c] Energy barrier to conformational exchange (*i.e.*, energy difference between the −*t* and most stable states)
[d] Free energy of unfolding determined by chemical denaturation with guanidinium chloride at 25 °C (n = 3, mean ± s.d.)
[e] Kinetic parameters (k_1, k_{-1}, $\Delta G^{\ddagger}$) reported at 15 °C, ΔG_{eq} at 25 °C
[f] Rate constant for exchange from major to minor state (n = 6, mean ± s.d.)
[g] Rate constant for exchange from minor to major state (n = 6, mean ± s.d.)
[h] Energy barrier for exchange from major to minor state (n = 6, mean ± s.d.)
[i] Free energy difference between major and minor states (n = 6, mean ± s.d.)
[j] Exchange peaks were observed but could not be quantified

2.3.3 Folded DANCERs undergo conformational exchange

Although the four DANCERs each contained between five and six mutations, representing approximately 10% of the Gβ1 total sequence length, they expressed as soluble monomers (Fig. S2.4a), adopted the native Gβ1 fold (Fig. S2.4b), and were stably folded (Table 2.1, Fig. S2.5). Chemical denaturation experiments could be fit to a two-state model with m-values similar to that of wild-type Gβ1 (Fig. S2.5c), indicating a similar extent of change in solvent-accessible surface area upon unfolding.[182] We used solution NMR to assess the dynamic properties of DANCER proteins, with ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra showing immediate evidence that DANCER proteins exist in two distinct conformational states (Fig. S2.6). Specifically, spectra for DANCER-1, DANCER-2, and DANCER-3 all showed a subset of low intensity peaks not seen in spectra of wild-type Gβ1 (Fig. S2.7). The only exception was DANCER-0, which instead showed significant peak broadening, suggesting that it is dynamic on a faster timescale.[124] ^{1}H-^{15}N HSQC ZZ-exchange experiments for DANCER-1, DANCER-2, and DANCER-3 (Fig. 2.2a, Fig. S2.8a) all showed cross peaks between low intensity peaks and assigned Gβ1 peaks, indicating that the low intensity subset of peaks arose from an alternate state of Gβ1 undergoing exchange with the major species. Free energy differences between major and minor states determined from these spectra (Table 2.1, ΔG_{eq}) were all smaller than free energies of unfolding which, combined with the large proton chemical shift dispersion of minor state spectra, confirmed that the exchange did not involve the unfolded state. Mixing-time dependent changes in peak intensities acquired over a range of temperatures could be fit to kinetic and thermodynamic parameters of exchange (Table 2.1, Fig. S2.8b), indicating that conformational exchange was occurring on the millisecond timescale. DANCER-1 exhibited approximately 10-fold faster exchange than DANCER-3, with an activation barrier that was 1.75 kcal/mol smaller

in magnitude. DANCER-2 also showed evidence of conformational exchange, although the small population of the minor state (< 10%) prevented quantitative measurement of kinetic parameters for this mutant.

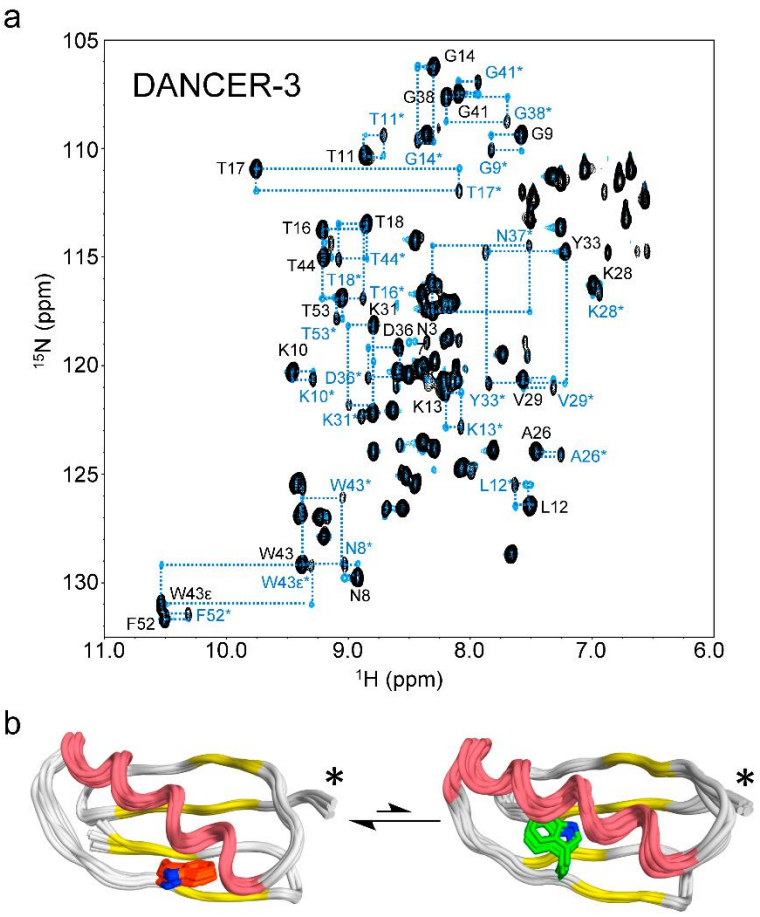


Figure 2.2. DANCER variants undergo conformational exchange. (a) The ZZ-exchange spectrum of DANCER-3 (blue) is shown overlaid with its ^{1}H-^{15}N HSQC spectrum (black) to highlight the presence of exchange peaks. Residues for which peaks could be unambiguously assigned between major and minor species are highlighted with dotted lines that connect exchange correlations with peaks from major and minor species. Peak assignments for these residues are shown in black (major state) and blue with an asterisk (minor state). (b) NMR ensembles of 10 conformers for DANCER-2 (with the Trp43 side chain in sticks, colored orange and green for major and minor states, respectively). The minor species of DANCER-2 is a model generated using NOESY data that excluded a small subset of peaks involving the Trp43 indole protons that were assigned to the major species during the automatic NOE assignment in CYANA (Methods). The N-terminus is indicated by an asterisk, while α-helices, β-strands, and loops, as assigned by chemical shift index results, are colored in red, yellow, and white, respectively.

2.3.4 Structural characterization of states sampled by DANCERs

To obtain structural evidence that the two conformations sampled by our dynamic Gβ1 variants matched structural states predicted by *meta*-MSD, we solved the solution NMR structure of the major state of DANCER-2 (Fig. 2.2b, Table S2.3). As predicted, this structure shows a native Gβ1 fold with χ_1 and χ_2 dihedrals for Trp43 that correspond to the solvent-exposed $-g(+)$ conformation (Table 2.2). However, there was also a secondary network of low intensity NOEs involving the Trp43 side chain that were not compatible with this structure (example shown in Fig. S2.9), but could be used to determine a structural model for the alternate, minor state (Methods). According to this model (Fig. 2.2b, Table S2.3), the configuration of Trp43 in the minor state is in the core-buried $-g(-)$ state (Table 2.2), as predicted by *meta*-MSD. Taken together, these data demonstrate that we have successfully designed a sequence that adopts the Gβ1 fold while undergoing exchange on a millisecond timescale between two previously unobserved conformational states that were the targets of our design protocol.

Table 2.2. Comparison of predicted and experimental structures

Protein	TM-score to 1PGA [a]	Predicted Trp43 Conformation	Experimental χ_1 (°) [b]	Experimental χ_2 (°) [b]
DANCER-2				
Major species	0.67	$-g(+)$	+75 ± 2	–74 ± 1
Minor Species	0.66	$-g(-)$	–95 ± 1	–110 ± 2
Static Gβ1 variants				
NERD-S	0.66	$-g(+)$	+54 ± 4	–89 ± 2
NERD-C	0.85	$+g(-)$	–84 ± 4	+80 ± 4

[a] TM-score has a value between 0 and 1, where 1 indicates a perfect match between two structures. Two proteins with a TM-score greater than 0.5 are considered to adopt the same fold.[183]
[b] Average over the NMR ensemble of 10 lowest energy conformers (n = 10, mean ± s.d.)

To evaluate the reliability of our *meta*-MSD predictions, we also characterized the structure and dynamics of DANCER-1 and DANCER-3. While the exchange parameters for these mutants made it impractical to determine structures (Methods), ^{1}H-^{15}N HSQC spectra of the major species showed similarities with those of other structurally characterized variants, suggesting a high degree of structural similarity with these states. Specifically, the DANCER-1 spectrum showed only small chemical shift differences from that of DANCER-2 (Fig. 2.3a), suggesting that the major species of DANCER-1 also contains Trp43 in the solvent-exposed $-g(+)$ state. Likewise, the ^{1}H-^{15}N HSQC spectrum for DANCER-3 was highly similar to that of a variant that we determined to thermodynamically and kinetically favor the $-g(+)$ state as predicted by *meta*-MSD (Fig. 2.3b), called NERD-S, for *Non-Exchanging Rigid Design* with a *Solvent-exposed* Trp43 side chain (Methods, Tables 2.1–2.2, Table S2.3, Fig. S2.10a). Therefore, in all three mutants predicted by *meta*-MSD to be dynamic for which structural information could be obtained, the major conformation was the Gβ1 structure with Trp43 being in the solvent-exposed $-g(+)$ state. This result is in disagreement with *meta*-MSD prediction of the core-buried $-g(-)$ state as the favored state for DANCER-1 and DANCER-3, a discrepancy that likely results from existing challenges in CPD algorithms that prevent perfect correlation of predicted and experimentally determined stabilities.[54,116,184]

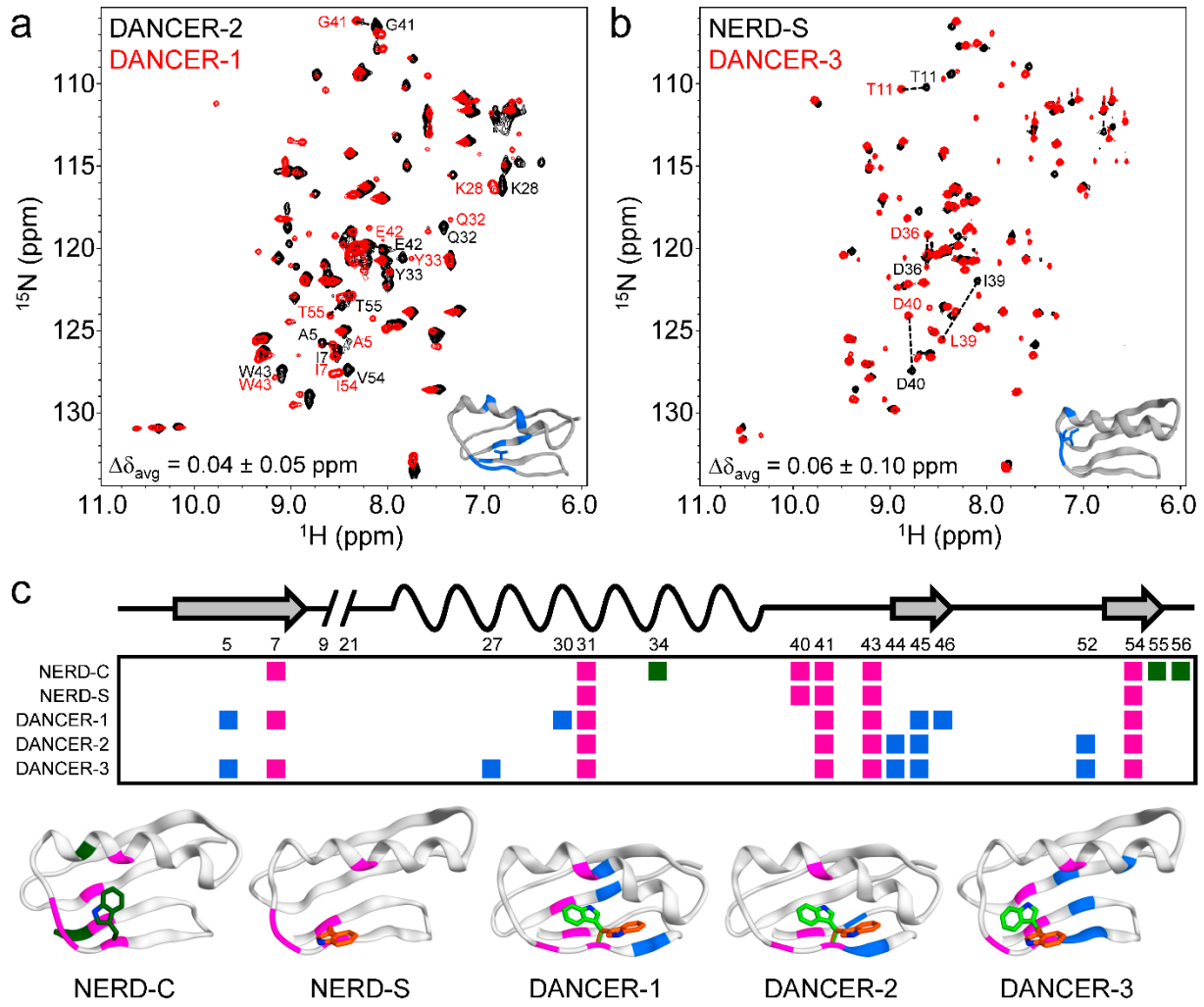


Figure 3. Structural analysis of DANCER-1 and DANCER-3. (a) Superimposed ^{1}H-^{15}N-HSQC spectra of DANCER-2 and DANCER-1 reveal high structural similarity between major states. Residues showing significant average amide shift differences between the major state of DANCER-2 and that of DANCER-1 ($\Delta\delta > \Delta\delta_{avg}$ + 1 s.d.) are labeled and highlighted in blue on the inset DANCER-2 structure. These residues are all proximal to the single amino acid that differs between the two DANCER proteins (shown as sticks). (b) ^{1}H-^{15}N-HSQC spectra demonstrating that the major state of DANCER-3 has the same structure as NERD-S. (c) Summary of NOE correlations involving the Trp43 indole N-H shown on a position map (secondary structure elements on top) and on each structure. Correlations are colored green, blue, or magenta, if they are observed in static, dynamic, or both variants, respectively. Each structure shows Trp43 side-chain in the conformation(s) consistent with observed NOEs. Included in this analysis is the solution NMR structure of NERD-C (*Non-Exchanging Rigid Design* with a *Core-buried* Trp43 conformation, Methods, Table 2.1), which was highly similar to the wild-type Gβ1 crystal structure (PDB ID: 1PGA[53]), with an all-atom RMSD of 1.10 ± 0.05 Å for residues comprising the hydrophobic core, and with Trp43 adopting the native +*g*(–) configuration (Fig. S2.10b, Table 2.2). NERD-C showed several unique indole N-H NOE correlations that were not observed in any of the DANCER variants, confirming that the DANCER proteins did not sample the native +*g*(–) configuration.

We used ^{15}N-NOESY spectra focusing on the Trp43 indole NH proton to gain insight into the structure of the minor states being sampled in the conformational exchange exhibited by DANCER-1 and DANCER-3. These spectra show NOE correlations (Fig. S2.11) to similar regions of the protein as was observed in DANCER-2 (Fig. 2.3c), consistent with exchange between core-buried –*g*(–) and solvent-exposed –*g*(+) states. Furthermore, comparison of NOEs involving the Trp43 indole NH proton confirmed that these correlations do not correspond to the core-buried state found in the wild-type structure [+*g*(–)] (Fig. 2.3c, Fig. S2.10b). Taken together, our NMR results confirm that Trp43 in both DANCER-1 and DANCER-3 exchanged between the solvent-exposed –*g*(+) and core-buried –*g*(–) conformations that were the targets of our design. These results also suggest that exchange occurred via a coordinated change in side-chain configurations for a triad of aromatic residues (Phe34, Trp43, Tyr45) in a process we called an aromatic relay (Fig. S2.12).

2.4 Discussion

The *meta*-MSD framework described here enabled the rational design of proteins that spontaneously exchange between two predefined states on the millisecond timescale without the need for an external stimulus to induce exchange. To our knowledge, this work represents the first successful application of CPD to design a specific mode of conformational exchange into a stable protein fold. Although previous CPD-based designs have accounted for intrinsic native-state dynamics via the use of conformational ensembles to improve protein design predictions, these examples have focussed on the design of ligand binding affinity[185,186] and protein stability,[54,115] but not conformational exchange. CPD has also been used to help identify point mutations that differentially alter the energy of ground and excited states of a protein undergoing conformational exchange, resulting in a shifted dynamic equilibrium that favors the species that had been sparsely

populated.[187] However, in this case the system already existed in a conformationally dynamic state, and hence there was no need to consider the energy barrier to exchange. Perhaps the most similar CPD-based example to our study is the design of a protein sequence that can reversibly switch between zinc finger and trimeric coiled-coil folds, depending on pH or the presence of transition metals.[164] In that system, exchange relied on the presence of a metal that was critical for the formation of the zinc finger structure. In contrast, our design of conformational exchange between two end-states in the absence of a ligand or other external stimulus required modeling of three states (e.g., major, minor, and transition states) as well as consideration of both the relative energies between the two target end-states (ΔE_{eq}) and the barrier to conformational exchange ($\Delta E^{\ddagger}$). Without estimation of both of these energy differences, it would not have been possible to distinguish between dynamic and static sequences (*e.g.*, both ΔE_{eq} and $\Delta E^{\ddagger}$ values for DANCERs were lower than for NERDs and wild-type Gβ1, Table 1).

Another unique feature of our design strategy that was essential for the successful design of all DANCER variants was the utilization of a *meta*-analysis-based approach, which enabled the use of a significantly larger structural ensemble than has previously been utilized in MSD. This was of critical importance, since the full complement of template generation procedures was required to approximate the energy landscape of the designed exchange trajectory with enough accuracy to predict the DANCER variants reported here (Table S2.2). In addition, the large ensemble size made it possible to design exchange in the absence of experimentally-derived structures corresponding to each state, in contrast with the metal-triggered conformational exchange that was previously designed using available crystal structures for the two end-states[164] and the designed equilibrium shift towards the excited state that was supported by CPD calculations using high-resolution structural data of both end-states.[187]

Our results showed that the introduction of dynamics on a functional timescale cannot be achieved via a single mutation and that instead dynamics are conferred through subtle interactions across a network of residues. For example, the A34F mutation, which was previously shown to induce dimerization of Gβ1 without altering the Trp43 conformation,[56,177] is common to all DANCER proteins and an integral component of the aromatic relay that underlies exchange (Fig. S2.12). However, this mutation alone was not sufficient to introduce conformational exchange into the Gβ1 fold, since the variant NERD-S also possessed this mutation but did not undergo exchange on the millisecond timescale (Table 2.1). Introduction of the conservative and isosteric I39L mutation into the NERD-S sequence appears to be sufficient to introduce the targeted conformational exchange, as it gave rise to the dynamic variant DANCER-3. These results highlight the challenges of attempting to infer dynamics from simple sequence characteristics, and demonstrate the power of *meta*-MSD to design conformational exchange into proteins even without prior knowledge of the mechanism of exchange.

The *meta*-MSD framework presented here is in principle applicable to the design of specific conformational exchange into any globular protein. In the future, *meta*-MSD could be used to design proteins with functions that rely on the ability to spontaneously access more than one conformational state (*e.g.*, molecular rotors, multi-substrate enzymes, biosensors, *etc.*). Alternatively, *meta*-MSD could be used to enrich functionally relevant but low occupancy states from an ensemble of dynamic configurations to improve function.[18] Moreover, while we have demonstrated the introduction of dynamics into a rigid protein, dampening of dynamics should in principle also be possible, as demonstrated by our design of NERD-C and NERD-S. This potential for *meta*-MSD to be used for the rigidification of highly dynamic regions in proteins without

adversely affecting the overall structure, in effect imitating conformational selection *in silico*, opens the door to the design of proteins with a wider range of functions than previously possible.

2.5 Methods

2.5.1 Structure preparation and ensemble generation

Coordinates for Streptococcal protein G domain β1 (Gβ1) were retrieved from the Protein Data Bank (PDB ID: 1PGA[53]). A set of eight seed structures was prepared by threading eight tryptophan rotamers (Table S2.1) at residue position 43 on the 1PGA template using the Molecular Operating Environment (MOE) software (Chemical Computing Group) and adding hydrogen atoms to each individual structure using the Protonate 3D utility.[188] To enable reshaping of the Gβ1 scaffold local structure space so as to accommodate all tryptophan conformations used here and ensure favorable scoring of all states by multistate design (MSD), the resulting 8 seed structures were used as input templates in a stepwise ensemble generation strategy (Fig. S2.2) that involved the introduction of backrub motions[189] using the RosettaBackrub server with default settings,[190] and the application of the *coordinate perturbation followed by energy minimization* (PertMin) algorithm.[116,181] RosettaBackrub was chosen for its ability to rapidly generate an ensemble of templates having large structural diversity while maintaining structural similarity to the original Gβ1 fold, a feature that is crucial to ensure that native ensembles are on-target.[116] The PertMin algorithm was chosen to finalize the ensemble generation procedure as it can produce low energy structures populating local minima about the structure space of the input structure.[181]

Each seed was used to generate 50 “backrubbed” node structures, and all seed plus node structures were explicitly solvated in a water cube with a minimum depth of 6 Å from the protein surface using MOE. PertMin was initiated by randomly perturbing the coordinates of all heavy

atoms (including water molecules) by ± 0.001 Å along each Cartesian coordinate axis to generate 30 perturbed structures from each seed and node. The perturbed structures were then energy minimized using 50 iterations of truncated Newton energy minimization[191] with the AMBER99 force field[192] in a reaction field implicit solvent model implemented in the MOE program. Unperturbed seed and node structures were also energy minimized following a similar protocol. This procedure yielded a set of 12,648 total backbone templates. It should be noted that the six Trp43 states (Fig. S2.1) were not produced in equal proportion in the final ensemble since the RosettaBackrub protocol allows side-chain rotamer moves.

2.5.2 MSD

All calculations were performed using the PHOENIX protein design software[54,97,193] with the fast and accurate side-chain topology and energy refinement (FASTER) algorithm[194,195] for sequence optimization. The backbone dependent Dunbrack rotamer library with expansions of ± 1 standard deviation around χ_1 and χ_2[196] was used to provide side-chain conformations to be threaded onto each fixed backbone template. Side-chain rotamers of Gβ1 core residues (Fig. S2.3a) were optimized on each template structure using amino acids observed at these positions in folded Gβ1 variants[54] as well as the wild-type amino acid at each position (numbered according to the 1PGA crystal structure): Residue 3 (Y and F), 5 (A, I, and L), 7 (F, I, L, and V), 30 (F, I, and L), 34 (A and F), 39 (I, L, and V), and 54 (A, I, and V). Side-chain rotamers of core residues W43 and F52 were also optimized. The designed sequence space thus consisted of 1,296 Gβ1 sequences (Fig. S2.3b). Sequences were scored using a four-term potential energy function consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field[98] with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal/mol and an equilibrium donor-acceptor distance of 2.8 Å,[161] an electrostatic energy term modelled using

Coulomb's law with a distance-dependent dielectric of 40, and a surface area-based solvation penalty term.[197] 408 individual MSD calculations were carried out on individual 31-member ensembles comprising 30 PertMin templates and one minimized seed or node structure (Fig. S2.2). Sequence stabilities during MSD optimization were calculated as a Boltzmann-weighted average at 300 K of the individual energies for each backbone from the 31-member ensemble.

2.5.3 Identification of DANCERs and NERDs by *meta*-MSD

Following MSD, the energy of every microstate for each sequence was recombined into a Boltzmann-weighted average energy (300 K) reflecting its overall stability across the total ensemble of 12,648 templates. These Boltzmann-weighted average energies were then used to identify 195 sequences with a predicted stability on the native Gβ1 fold greater than that of the wild type (–74.75 kcal/mol at 300 K). The energy of each state for these 195 sequences was determined by taking the energy of the most stable microstate assigned to that state, and the resulting state energies were used to construct an energy profile for each sequence. Of the 195 sequences, 35 had energy profiles that were predicted to be dynamic based on the following two criteria: 1) the energy difference between target core-buried [–*g*(–)] and solvent-exposed [−*g*(+)] states (ΔE_{eq}) was less than 4.2 kcal/mol, a value selected because it corresponds to a theoretical population ratio of 1:1000 between these desired minor and major states; 2) the energy barrier to exchange ($\Delta E^{\ddagger}$, *i.e.* energy difference between the −*t* and most stable states) was higher than 4.2 kcal/mol but lower than 16.8 kcal/mol, a value chosen as it is slightly higher than the value calculated for wild-type Gβ1. Four of these sequences were selected for experimental characterization (DANCER-0, DANCER-1, DANCER-2, and DANCER-3) because their $\Delta E^{\ddagger}$ values spanned a range that allowed evaluation of the accuracy of this metric for predicting exchange barriers. Note that the native +*g*(–) state was correctly predicted for wild-type Gβ1 with

a large ΔE_{eq} with respect to the end-state that is closest to it in energy (Table 2.1). In addition, the Gβ1 variant "H" that we previously showed to be static on the microsecond to millisecond timescale[115] was predicted by *meta*-MSD to adopt only the native Trp43 conformation [+*g*(–)], since it has higher ΔE_{eq} and $\Delta E^{\ddagger}$ values compared to the wild-type sequence. Here we renamed this variant NERD-C, for *Non-Exchanging Rigid Design* with a *Core-buried* Trp43 side chain. Another mutant was also selected to represent the solvent-exposed −*g*(+) configuration, called NERD-S, based on its calculated ΔE_{eq} and $\Delta E^{\ddagger}$ values, which were both larger than those of DANCER variants (Table 2.1).

2.5.4 *Meta*-MSD predictions using reduced-size ensembles

To verify whether the full ensemble of 12,648 templates was necessary to predict the dynamic variants reported here (DANCER proteins), we also performed *meta*-MSD with ensembles comprising a subset of these templates (Table S2.2). Specifically, we used 8-, 248-, and 408-member ensembles containing minimized seed structures, minimized seeds and PertMin templates generated from each seed structure, or minimized seed and node structures, respectively. *Meta*-MSD with these reduced-size ensembles was performed as described in the previous paragraph with the exception that *meta*-analysis considered only the energies and structures of rotamer-optimized microstates derived from each subset of templates.

2.5.5 Protein expression and purification

Codon-optimized and his-tagged Gβ1 genes cloned into the pJ414 vector were obtained from DNA2.0. Proteins for chemical denaturation assays and circular dichroism (CD) spectroscopy experiments were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using Luria-Bertani (LB) broth supplemented with 100 μg/mL ampicillin. Proteins for NMR spectroscopy were expressed

using M9 minimal expression medium supplemented with 1 g/L ^{15}N-ammonium chloride and/or 3 g/L ^{13}C-D-glucose for isotopic enrichment. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of approximately 0.6 after which protein expression was initiated with 1 mM isopropyl β-D-1-thiogalactopyranoside. Following overnight incubation at either 15 °C or 37 °C with shaking (250 rpm) for cultures grown in LB or M9 medium, respectively, 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 (Qiagen), followed by gel filtration in 10 mM sodium phosphate buffer (pH 7.4) using an ENrich SEC 650 size-exclusion chromatography column (BioRad). Purified samples were concentrated using Amicon Ultracel-3K centrifugal filter units (EMD Millipore).

2.5.6 Thermal denaturation assays

CD spectroscopy was performed with a Jasco J-815 spectrometer using 650-µL aliquots of each Gβ1 sample at a concentration of 40 µM in 10 mM sodium phosphate buffer (pH 7.4). Samples placed in a 1-mm path-length quartz cuvette (Jasco) were heated at a rate of 1 °C per minute, and ellipticity at 208 nm was measured every 2 °C. Melting point (Tm) values were determined by fitting a 2-term sigmoid function with baseline correction[198] using nonlinear least-squares regression. Reversibility was confirmed by comparing CD spectra acquired before and after thermal denaturation experiments from 185–250 nm at 25 °C.

2.5.7 Chemical denaturation assays

Chemical denaturation assays were performed in triplicate using protein samples at a 1 mg/mL concentration. Protein aliquots of 25 µL in individual wells of UV-Star 96-well plates (Greiner Bio-One) were mixed with 175 µL of 0–5 M guanidinium chloride solutions (12 points, evenly

spaced) and incubated at room temperature for an hour. Fluorescence emission spectra were measured for each sample from 300 nm to 450 nm (excitation at 295 nm and step size of 2 nm) using an Infinite M1000 plate reader (Tecan). Fluorescence was integrated and converted into fraction of unfolded protein values. Error on these values is reported as standard deviation from three replicates at each denaturant concentration for each Gβ1 variant. Cm values (concentration of denaturant at midpoint of denaturation) were determined by fitting a 2-term sigmoid function using nonlinear least-squares regression.

2.5.8 NMR spectroscopy

^{15}N- and ^{13}C-labelled Gβ1 samples for NMR consisted of 0.1–2.0 mM protein in 10 mM sodium phosphate buffer (pH 7.4), 10 μM EDTA, 0.02% sodium azide, and 10% D_2O for experiments requiring detection of amide protons or 99% D_2O otherwise. DANCER-1 and DANCER-2 ^{13}C-labelled samples additionally contained 1× cOmplete EDTA-free Protease Inhibitor Cocktail (Roche). 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. Chemical shift assignment and NOESY experiments were performed at 25 °C and ZZ-exchange experiments were performed at temperatures varying from 5 °C to 25 °C. NMR data sets were processed with the NMRPipe software package[199] and spectra were analyzed with NMRViewJ (One Moon Scientific).[200] 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 for backbone assignments, and NH-TOCSY (with mixing time (τ_{mix}) = 75 ms), CCH-TOCSY (τ_{mix} = 32.5 ms) and HCCH-TOCSY (τ_{mix} = 27.5 ms) spectra for side-chain assignments. Average backbone amide chemical shift differences (Δδ) between major and minor states for DANCER

proteins were calculated using $\Delta\delta = ((\Delta\delta_{HN})^2 + (\Delta\delta_N/5)^2)^{0.5}$ where $\Delta\delta_{HN}$ and $\Delta\delta_N$ are chemical shift differences calculated for amide proton and nitrogen atoms, respectively. To obtain distance restraints, ^{15}N-edited and ^{13}C-edited HSQC-NOESY (τ_{mix} = 120 ms) spectra were acquired. Secondary structure predictions from backbone chemical shifts[134] were performed in NMRViewJ. ZZ-exchange spectra were analyzed by fitting a Gaussian model to peaks in distinct clusters and integrating using NMRDraw.[199] Rates of exchange were determined by fitting 4-term relaxation and exchange curves[137] using nonlinear least-squares regression, and thermodynamic parameters were determined by fitting exchange rates and temperatures to the Eyring equation.

2.5.9 Structure determination

TALOS+[132] was used to determine secondary structure propensities and backbone dihedral restraints for Gβ1 variants on the basis of measured chemical shifts for $^{1}H_{\alpha}$, ^{15}N, $^{13}C'$, $^{13}C_{\alpha}$, and $^{13}C_{\beta}$ chemical shifts. Simultaneous NOE assignment and structure calculation was performed for NERD-C, NERD-S, and the major state of DANCER-2 using CYANA 2.1.[135] It was not possible to solve structures for DANCER-1 since the significantly populated minor state and millisecond timescale exchange between major and minor states gave rise to NOEs between atoms that contain contributions from both states. Similar difficulties also hindered structure determination for DANCER-3, even though its minor state was populated to a smaller extent. In the case of DANCER-2, peaks from the minor species were weakly populated and broad relative to those of the major state, minimizing the contribution of these resonances to NOESY spectra. In addition, during NOESY peak-picking, efforts were made to avoid including the relatively small number of broad correlations that were observed from the minor species in this spectrum. Chemical shifts from cross peaks in 3D ^{15}N-edited and ^{13}C-edited NOESY-HSQC spectra were used as input, in addition to TALOS+ derived dihedral angle restraints. A total of 753, 507, and 417 unique and

non-redundant distance restraints were used to calculate 100 conformers for NERD-C, NERD-S, and the major state DANCER-2, respectively, and NMR ensembles were represented by the 10 lowest energy conformers. No distance violations >0.5 Å or torsion angle violations >5 ° were observed in any of the structures calculated. Structural statistics are reported in Table S2.3.

2.5.10 Model generation of the DANCER-2 minor species

The millisecond timescale of exchange in DANCER-2 made it possible for strong NOE correlations from the minor species to appear in NOESY spectra used to determine the structure of the major state (Fig. 2.3c, Fig. S2.11). This included a secondary network of weaker NOEs that could not be assigned in a way that was compatible with the [−*g*(+)] Trp43 conformation determined in the major state structure, a state that was largely dictated by an extensive network of NOEs involving the Trp43 side chain, including strong correlations to Glu27 and Ala31. The alternate network of NOEs included unambiguous correlations between Trp43 indole protons and the side chains of Ile15, Val62 and Phe60. To generate a model of the alternate state being detected in the NOESY experiments, another set of NOE assignments coupled with structure calculation were performed in CYANA[58] using the same datasets, but omitting the strong correlations made by Trp43 with Glu27 and Ala31 from the original peak lists. In this process 427 unique and non-redundant distance restraints were assigned, although it should be acknowledged that this model of the minor state structure (Fig. 2.2b, Table S2.3) might suffer from inaccuracies arising from the assumption that most NOE correlations beyond the Trp43 interaction site are common between major and minor species.

2.5.11 Code availability

All PHOENIX scripts used are available upon request.

2.5.12 Data availability

Structure coordinates have been deposited in the Protein Data Bank with accession codes 5UB0 (NERD-C), 5UBS (NERD-S), 5UCE (major state of DANCER-2), and 5UCF (minor state of DANCER-2). NMR data has been deposited in the Biological Magnetic Resonance Data Bank with accession codes 30220 (NERD-C), 30221 (NERD-S), 30222 (DANCER-2), 27030 (DANCER-0), 27031 (DANCER-1), and 27032 (DANCER-3).

2.6 Supplementary Information

Table S2.1. Tryptophan rotamers used to create seed structures

Seed	Conformation	χ_1 (°)	χ_2 (°)
1[a]	+*g*(−)	−68	+74
2	−*g*(+)	+63	−93
3	+*g*(+)	+63	+94
4	+*t*	−175	+76
5	−*g*(−)	−61	−91
6	+*g*(−)	−61	+111
7	−*t*	−173	−113
8	+*g*(−)	−86	+56

[a] Wild-type conformation found in the 1PGA crystal structure

Table S2.2. Meta-MSD predictions using reduced-size ensembles [a]

	***Meta*-MSD Predictions [f]**		
Ensemble [b]	**Unstable**	**Static**	**Dynamic**
Full (n = 12,648)	0	160 (160, 0)	35 (0, 35)
Seed (n = 8) [c]	113 (93, 20)	82 (67, 15)	0
PertMin (n = 248) [d]	64 (60, 4)	130 (100, 30)	1 (0, 1) [g]
Backrub (n = 408) [e]	69 (61, 8)	123 (99, 24)	3 (0, 3) [g]

[a] Only the 195 sequences that had been predicted to be stable by *meta*-MSD using the complete 12,648-member ensemble were evaluated with the reduced-size ensembles in this test.
[b] See Supplementary Fig. 2 for definition of seed and node structures.
[c] Contains 8 minimized seed structures.
[d] Contains 8 minimized seed structures and 240 PertMin templates generated from each seed (30 per seed).
[e] Contains 8 minimized seed and 400 minimized node structures.
[f] Numbers of sequences reported as: total (static, dynamic). Total is the total number of sequences that were predicted using the reduced-size ensemble to be unstable, static or dynamic, as indicated in the column heading. These are broken down into number of sequences within this category that were predicted with the full ensemble to be static or dynamic.
[g] None of the sequences predicted to be dynamic when using these reduced-size ensembles are DANCER variants characterized in this study.

Table S2.3. Summary of NOE restraints and structural statistics [a]

	DANCER-2 Major	DANCER-2 Minor	NERD-S	NERD-C
PDB ID	5UCE	5UCF	5UBS	5UB0
Distance Restraints Statistics				
Number of NOEs	417	427	507	753
Short Range ($\|i-j\| \leq 1$)	266	267	276	367
Medium Range ($1 < \|i-j\| < 5$)	59	63	66	133
Long Range ($\|i-j\| \geq 5$)	92	97	165	253
MolProbity Ramachandran Plot Statistics (%)				
Residues in most favored regions	95.2	97.8	98.0	98.1
Residues in allowed regions	4.8	2.2	2.0	1.9
Residues in disallowed regions	0.0	0.0	0.0	0.0
Average RMSD to mean (Å)				
Backbone (mean ± s.d.)	0.48 ± 0.08	0.64 ± 0.14	0.48 ± 0.06	0.31 ± 0.08
Heavy Atom (mean ± s.d.)	0.99 ± 0.12	1.16 ± 0.13	0.95 ± 0.08	0.72 ± 0.07
RPF Scores				
Recall	0.825	0.882	0.891	0.933
Precision	0.699	0.725	0.854	0.857
F-measure	0.757	0.770	0.872	0.893
DP-score	0.590	0.633	0.698	0.818
Structure Quality Factors (Raw / Z-score [b])				
MolProbity clash score	10.15 / –0.22	3.73 / 0.89	15.65 / –1.16	15.91 / –1.20
Procheck G-factor (phi & psi)	–0.61 / –2.08	–0.30 / –0.87	–0.46 / –1.49	–0.43 / –1.38
Procheck G-factor (all)	–0.8 / –4.73	–0.44 / –2.60	–0.57 / –3.37	–0.47 / –2.78
Verify3D	0.29 / –2.73	0.28 / –2.89	0.37 / –1.44	0.40 / –0.96
ProsaII (negative)	0.43 / –0.91	0.37 / –1.16	0.66 / 0.04	0.64 / –0.04

[a] Analyzed for the 10 lowest energy structures for each designed protein using CYANA,[135] MolProbity,[201] and PSVS[202,203]

[b] With respect to mean and standard deviation for a set of 252 X-ray structures with sequence lengths ≤ 500, resolution ≤ 1.80 Å, and R-free ≤ 0.28

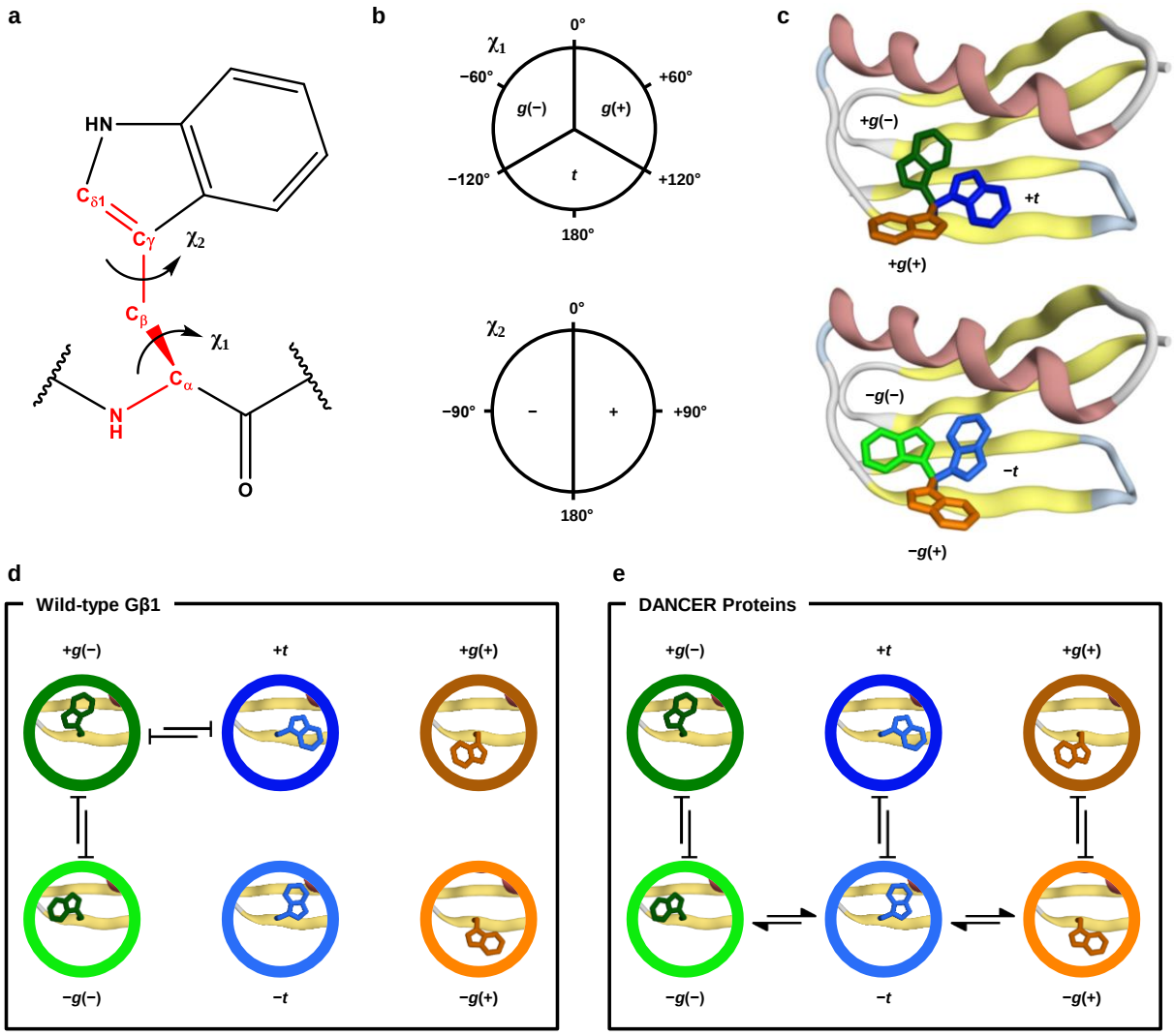


Figure S2.1. Trp43 conformations and exchange trajectories. a, The tryptophan side chain contains two rotatable bonds, χ_1 (N–C_α–C_β–C_γ) and χ_2 (C_α–C_β–C_γ–$C_{\delta1}$). **b,** χ_1 dihedrals are clustered into one of three bins centered at −60°, 180°, and +60°, named *g*(−), *t*, and *g*(+) for *gauche*(−), *trans*, and *gauche*(+), respectively. χ_2 dihedrals are clustered into two bins centered at −90° and +90°, which are named "–" and "+", respectively. Although the tryptophan side-chain χ_2 dihedral can also be centered at 0°, this configuration was not included in our analysis as it is sparsely populated.[196] **c,** The combination of χ_1 and χ_2 dihedral angles yields six conformations shown on the Gβ1 backbone (PDB ID: 1PGA).[53] The Trp43 side chain is colored according to its state as defined by its χ_1 and χ_2 dihedral angles, with core-buried, intermediate, or solvent-exposed states shown in green, blue, or orange, respectively. The Trp43 conformation found in wild-type Gβ1 is in dark green [+*g*(–)]. **d,** Trp43 in wild-type Gβ1 occupies the core-buried +*g*(−) conformation that does not undergo exchange. **e,** Trp43 in DANCER proteins was designed to exchange between the core-buried −*g*(−) and solvent-exposed −*g*(+) conformations through a trajectory traversing the intermediate –*t* conformation. Other exchange trajectories were not allowed.

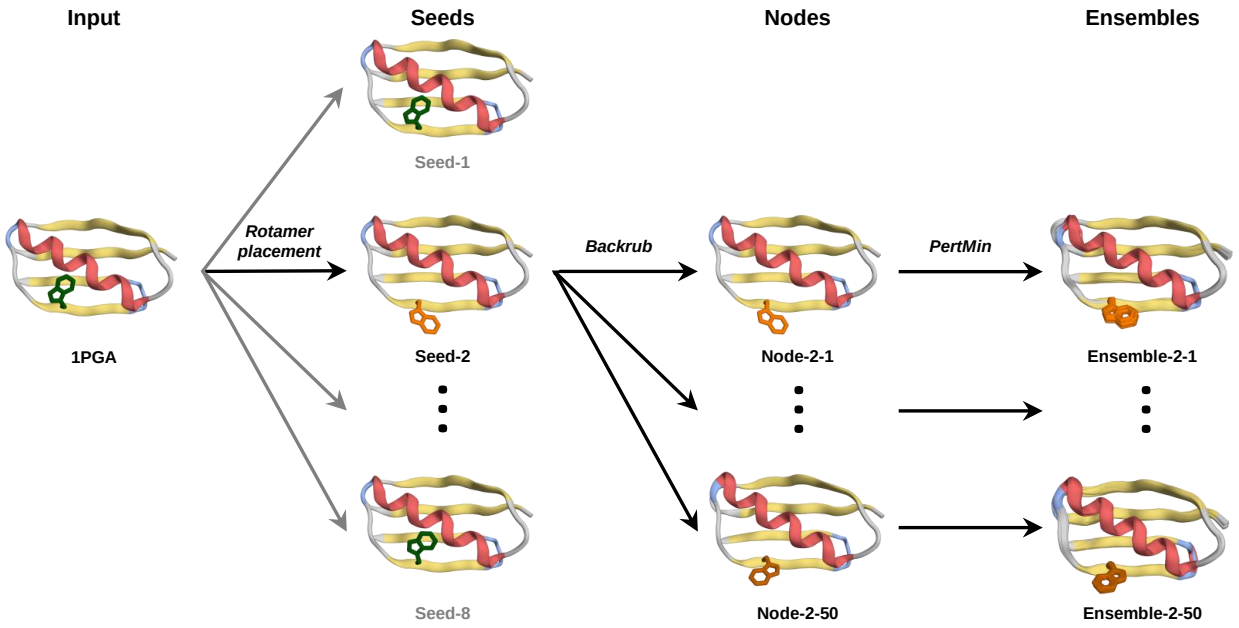


Figure S2.2. Template generation procedure. Native ensembles of Gβ1 templates were prepared in a step-wise procedure consisting of the generation of 8 seed structures using rotamer placement of one of eight tryptophan rotamers (Supplementary Table 1) on the wild-type crystal structure at position 43, the creation of 50 node structures from each seed by remodeling the backbone using Backrub motions,[189] and application of the *coordinate perturbation followed by energy minimization* (PertMin) procedure[116,181] to each node structure to generate 30-member PertMin ensembles. For example, a tryptophan rotamer corresponding to the solvent-exposed –*g*(+) conformation is threaded onto the 1PGA crystal structure to generate Seed-2. Seed-2 is then subjected to RosettaBackrub[190] 50 times, yielding the Node-2-1 to Node-2-50 structures. Because RosettaBackrub allows side-chain moves, the Trp43 rotamer can be altered from that of the seed structure following this procedure, as seen on Node-2-50, which now contains the solvent-exposed +*g*(+) conformation. Application of the PertMin procedure to each seed and node results in a set of 30-member ensembles of PertMin structures (overlaid). Each seed and node structure was also energy-minimized, and the resulting structure was added to the corresponding PertMin ensemble. Thus, a total of 12,648 unique templates were generated from the 8 seed and 400 node structures (408 structures × 31 ensemble members).

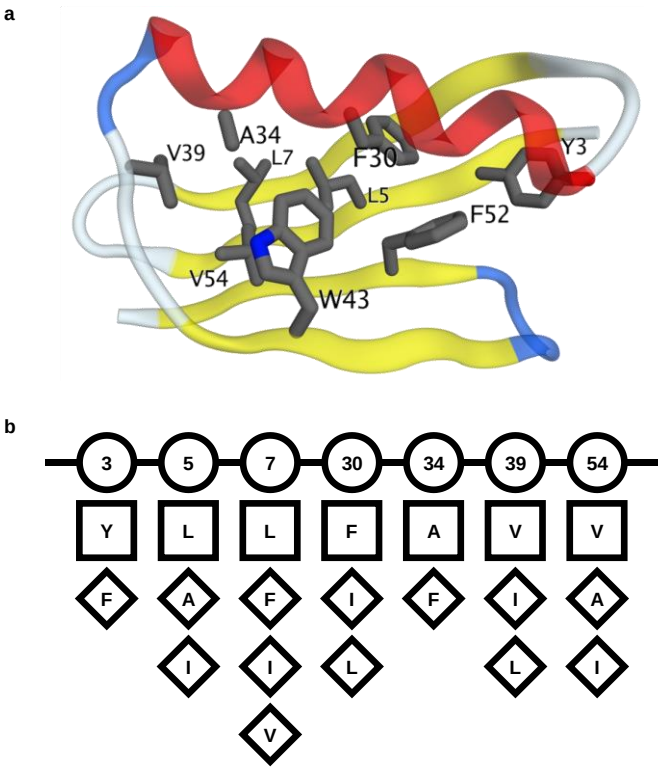


Figure S2.3. Gβ1 design space. a, Crystal structure of wild-type Gβ1 (PDB ID: 1PGA)[53] with core-buried residues highlighted as sticks. **b,** Designed sequence space comprising 1,296 sequences. Circles indicate the residue position. Wild-type residues and mutations are indicated in squares or diamonds, respectively. W43 and F52 were not mutated but their rotameric configuration was allowed to vary during multistate design.

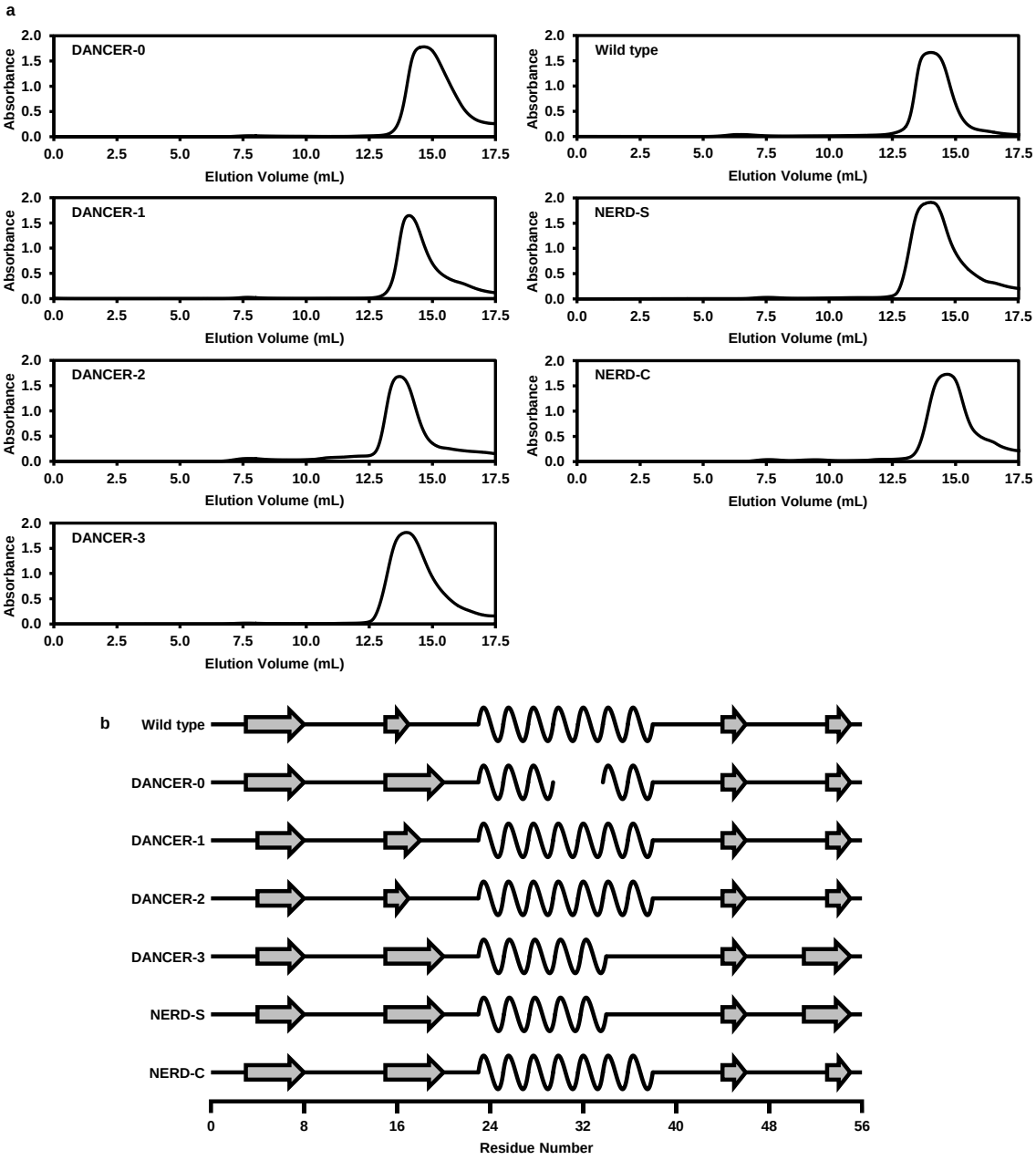
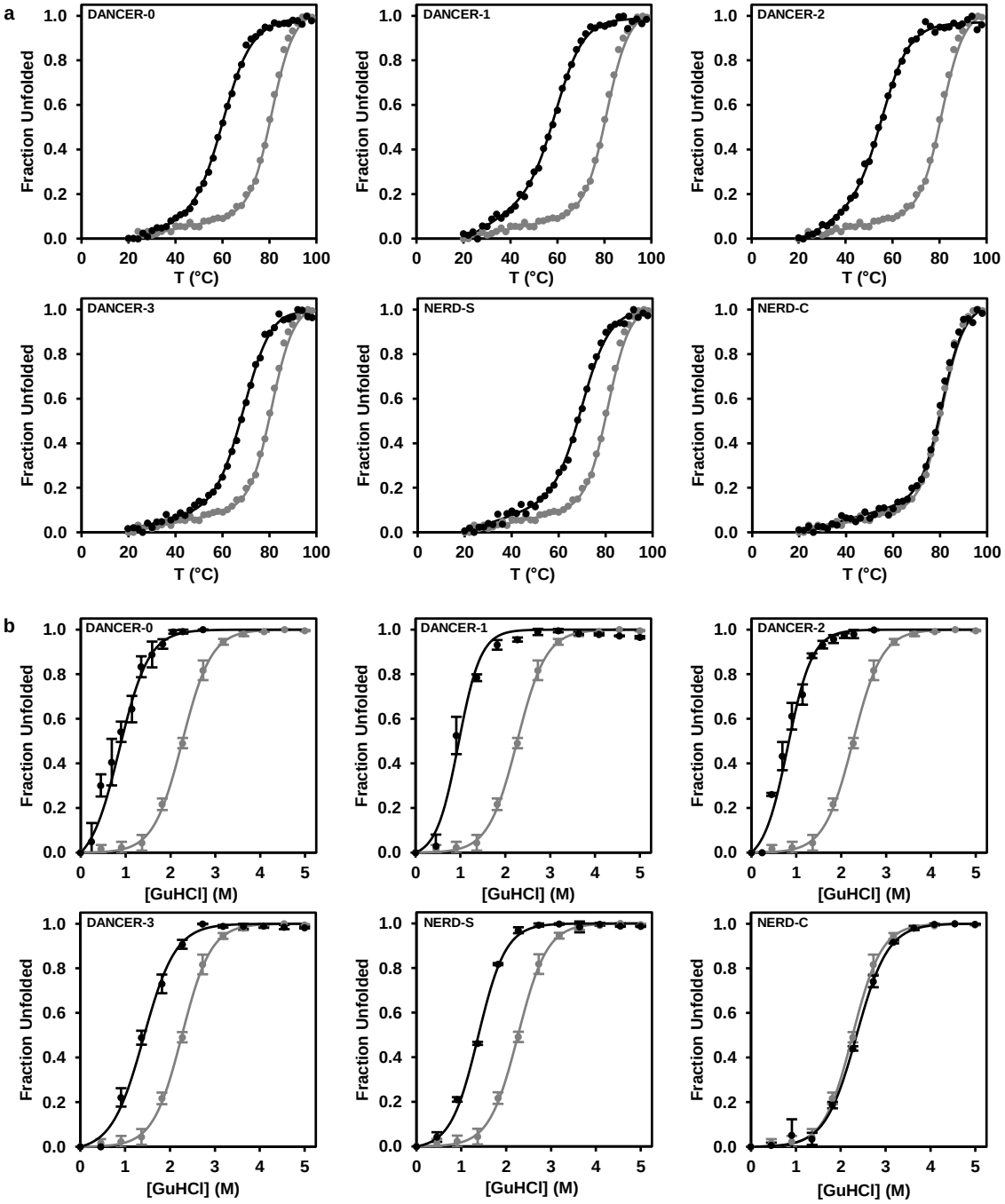


Figure S2.4. Gβ1 variants adopt a similar fold. a, Size-exclusion chromatography elution profiles monitored by absorbance at 280 nm during the final purification step show a high degree of similarity between wild-type, dynamic, and static variants of Gβ1. Even at the high concentrations used for these profiles (1–10 mM) there is no evidence of higher order oligomer formation. Small variations in elution volumes between samples were within instrument variability. **b,** Secondary structure content for each variant was evaluated using NMR chemical shift indices.[134] Secondary structure elements (α-helices and β-strands are indicated by wave and arrow symbols, respectively) are conserved between wild-type Gβ1 and its variants, confirming that they adopt the same folded structure. For DANCER-0, chemical shifts for residues 30 through 35 could not be assigned. This may be due to conformational exchange on an intermediate timescale that has broadened the peaks beyond detection in this part of the protein.



c

Protein	Tm (°C)	Cm [a] (M)	m-value [b] (kcal/mol·M)
Wild type	81.3	2.26 ± 0.04	1.8 ± 0.1
DANCER-0	61.0	0.85 ± 0.09	1.8 ± 0.2
DANCER-1	60.7	0.95 ± 0.05	2.3 ± 0.2
DANCER-2	56.7	0.81 ± 0.05	2.1 ± 0.1
DANCER-3	70.5	1.39 ± 0.05	1.5 ± 0.2
NERD-S	70.7	1.38 ± 0.01	1.9 ± 0.1
NERD-C	80.8	2.35 ± 0.02	1.7 ± 0.1

[a] Concentration of guanidinium chloride at midpoint of denaturation.
[b] Determined from chemical denaturation experiments at 25 °C.

Figure S2.5. Stability of Gβ1 variants. a, Thermal denaturation monitored by circular dichroism spectroscopy at 208 nm indicates that all Gβ1 variants are stable at room temperature (n = 1). Data was fit to a two-state unfolding model (lines).[198] Data from the thermal denaturation of wild-type Gβ1 is shown in grey for comparison. **b,** Chemical denaturation using guanidinium chloride (GuHCl) demonstrates that all proteins unfold according to a two-state model (n = 3). Fraction unfolded values were calculated by monitoring integrated tryptophan fluorescence ($\lambda_{excitation}$ = 280 nm, $\lambda_{emission}$ = 310–450 nm) relative to fluorescence at 0 M GuHCl. All experiments were performed in triplicate, and the average and standard deviation reported for each data point. The wild-type denaturation curve is shown in grey for comparison. **c,** Table showing melting temperatures (Tm), Cm, and m-values for Gβ1 variants.

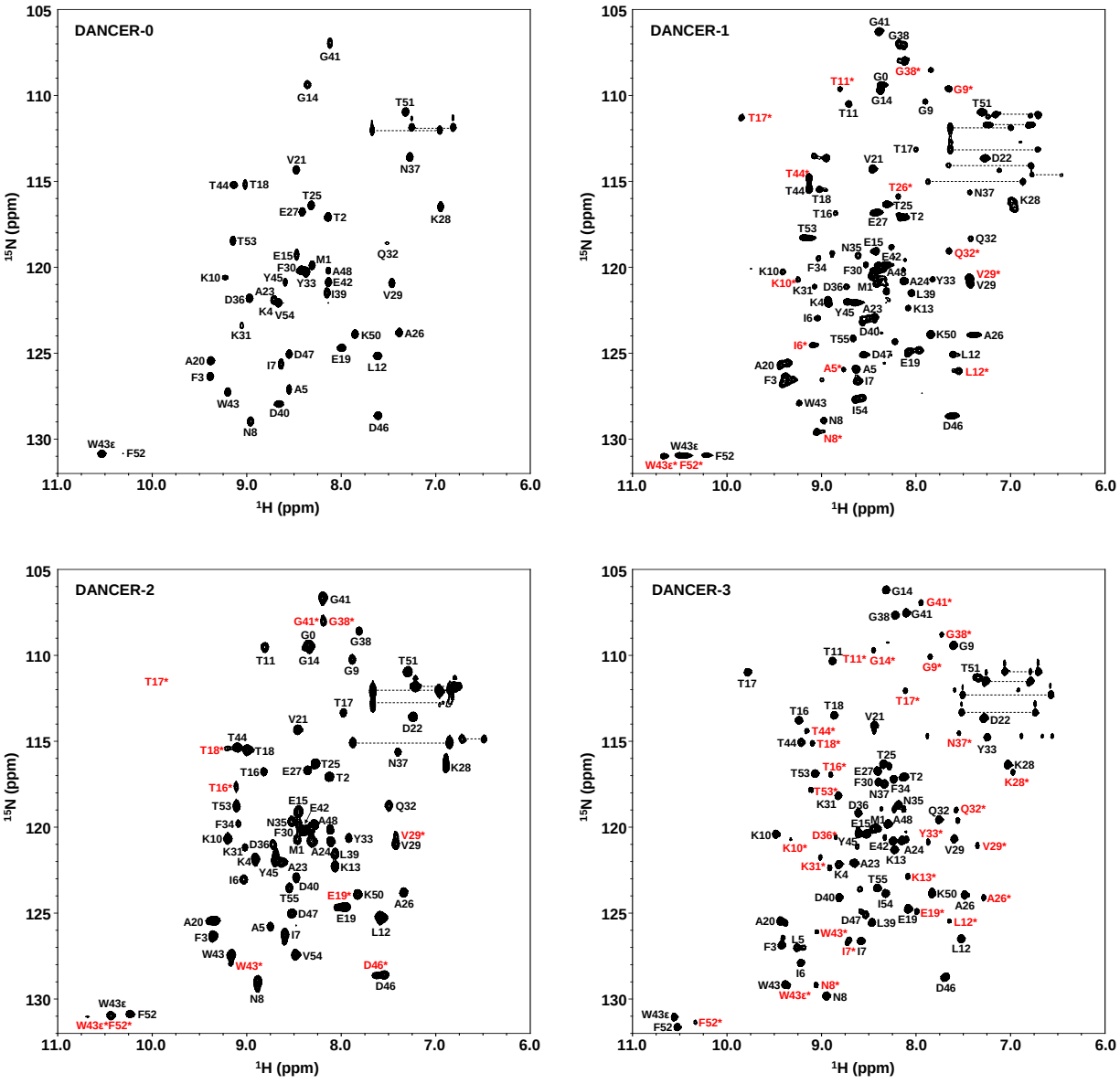


Figure S2.6. Assigned ^{1}H-^{15}N-HSQC spectra of DANCER Gβ1 variants. Peaks corresponding to the minor state are indicated with red assignments. Side-chain amide resonances from asparagine and glutamine residues are connected by horizontal lines.

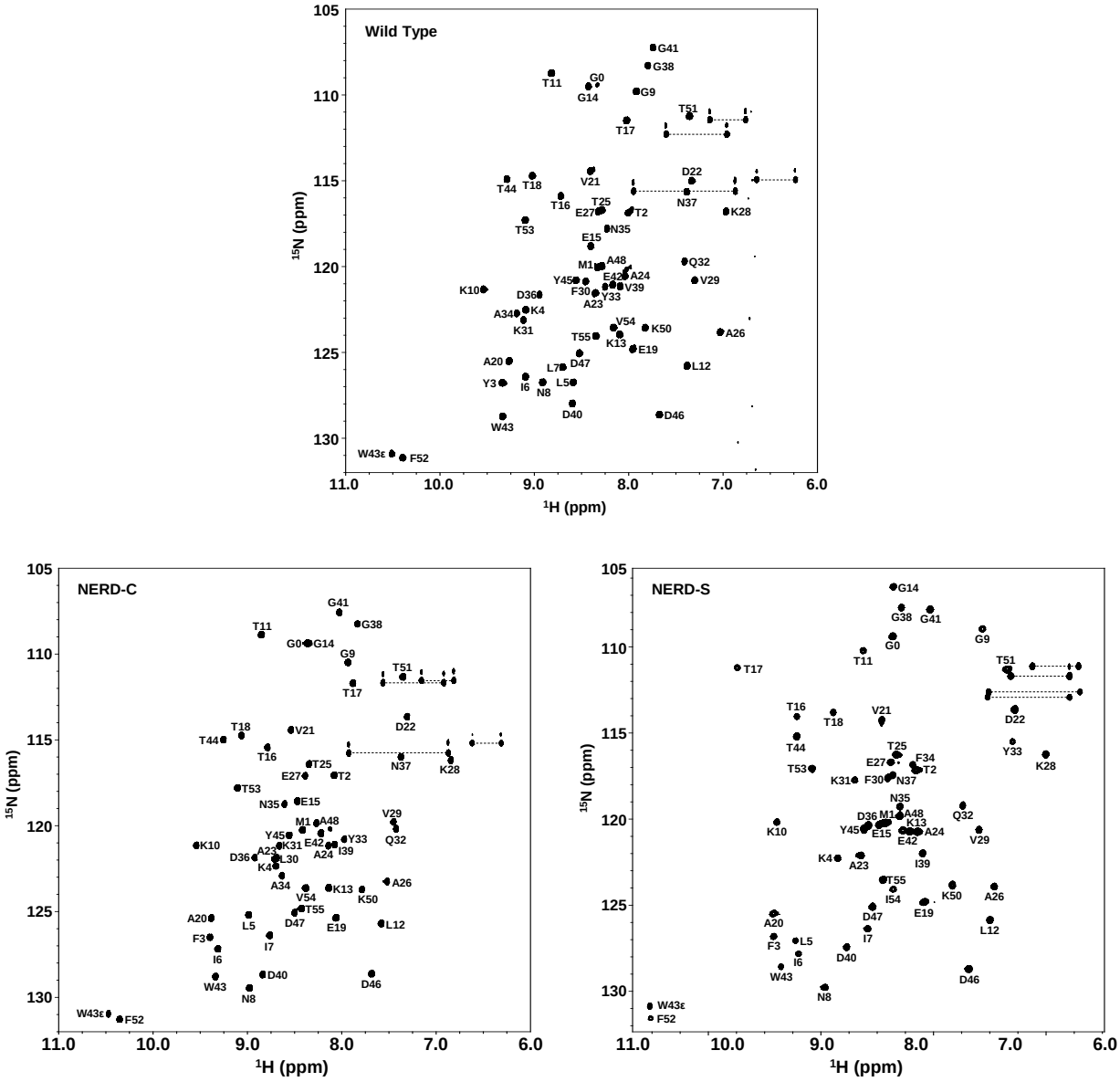


Figure S2.7. Assigned ^{1}H-^{15}N-HSQC spectra of wild-type and static Gβ1 variants. Side-chain amide resonances from asparagine and glutamine residues are connected by horizontal lines.

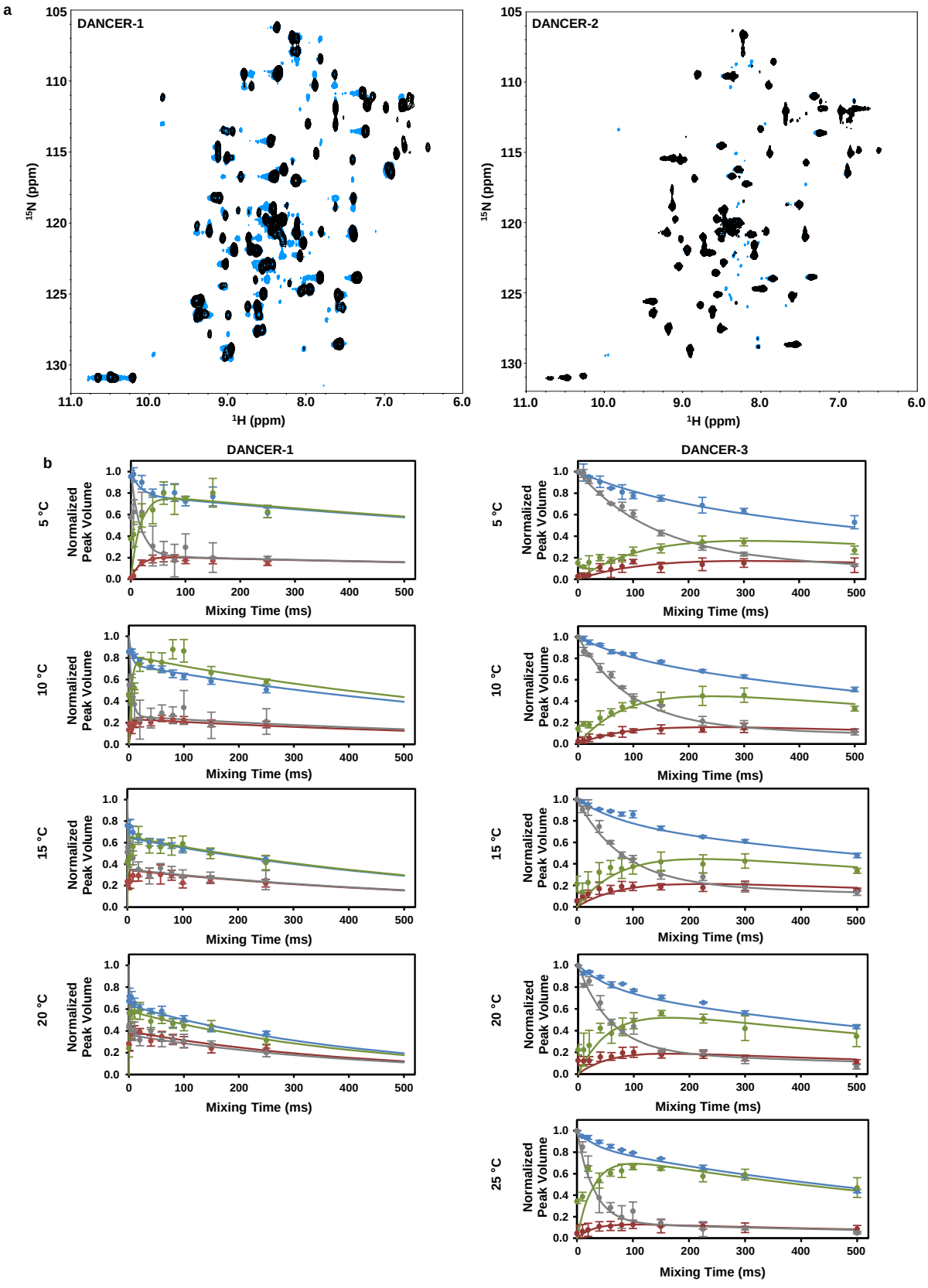

a
DANCER-1
DANCER-2
^{15}N (ppm)
^{1}H (ppm)
b
DANCER-1
DANCER-3
5 °C
10 °C
15 °C
20 °C
25 °C
Normalized Peak Volume
Mixing Time (ms)

Figure S2.8. ^{1}H-^{15}N ZZ-Exchange analysis of DANCER variants. a, Representative ZZ-exchange spectra are shown in blue and overlaid with ^{1}H-^{15}N HSQCs in black to highlight the presence of exchange peaks for DANCER-1 and DANCER-2. In the case of DANCER-2, exchange peaks were very low in intensity due to the small population of the minor state. **b,** Exchange profiles for normalized peak intensities from residues Thr11, Thr17 and Gly38, shown as a function of mixing time. Each point is the average of 2 replicates over these three residues (n = 6), and error bars show the standard deviation. Lines represent fitted models for exchange and relaxation,[137] for the $-g(+)$ peak (blue), $-g(-)$ peak (grey), the $-g(+)$ to $-g(-)$ exchange peak (red), and the $-g(-)$ to $-g(+)$ exchange peak (green).

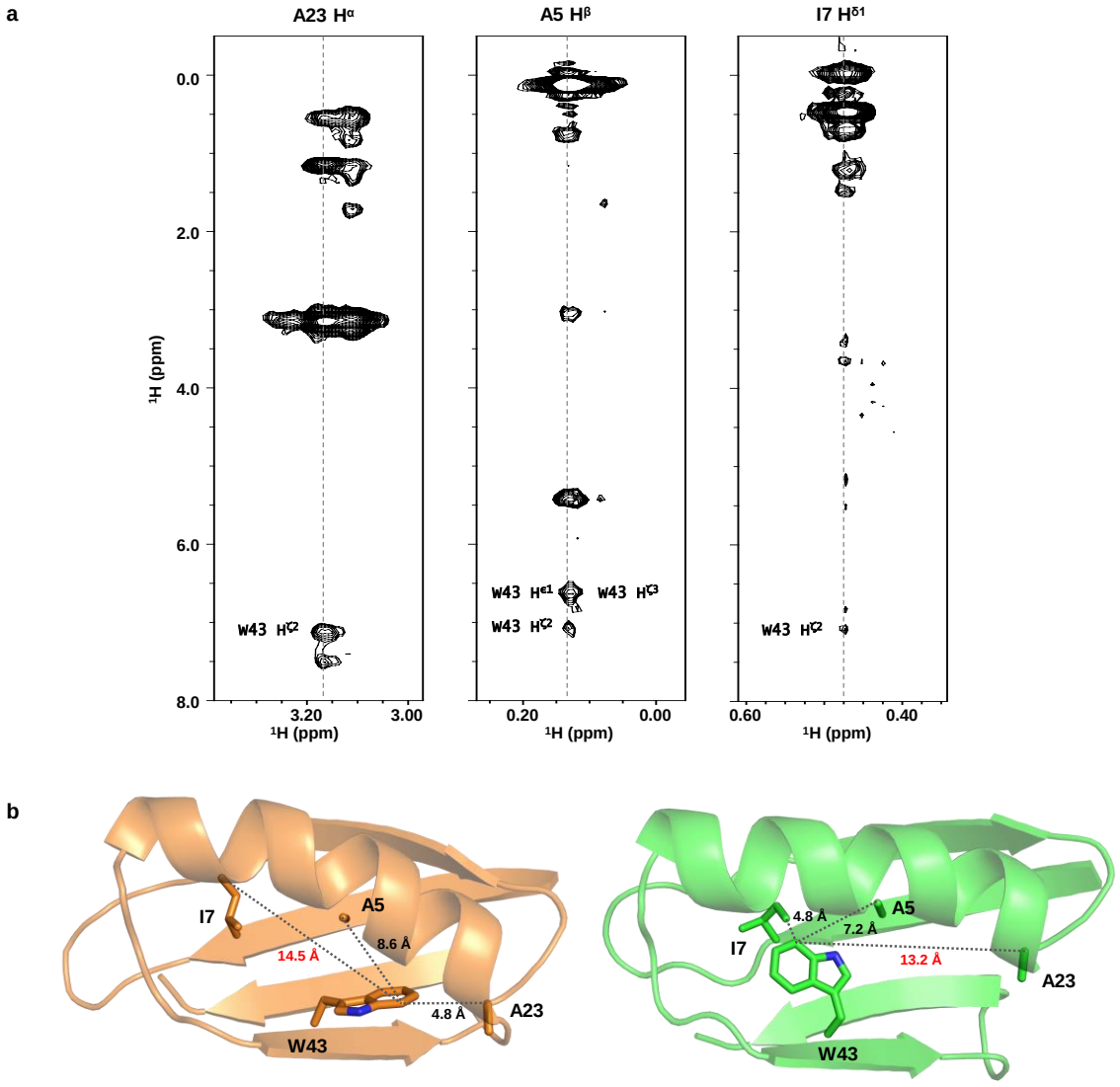


Figure S2.9. Unique NOE correlations to Trp43 observed in DANCER-2. a, Strip plots for atoms with NOE correlations to the characteristic Trp43 Hζ2 proton from 3D ^{13}C-NOESY spectra acquired with a mix time of 120 ms. **b,** DANCER-2 major and minor state average structures are shown in orange and green, respectively, with internuclear distances corresponding to each NOE correlation shown for each structure. Atoms correlated by NOEs indicated in the strip plot are connected by dotted lines in each structure, labeled with C–C distances to the Trp43 Cζ2 (red for distances incompatible with observed NOEs, and black for distances that could give rise to NOE correlations).

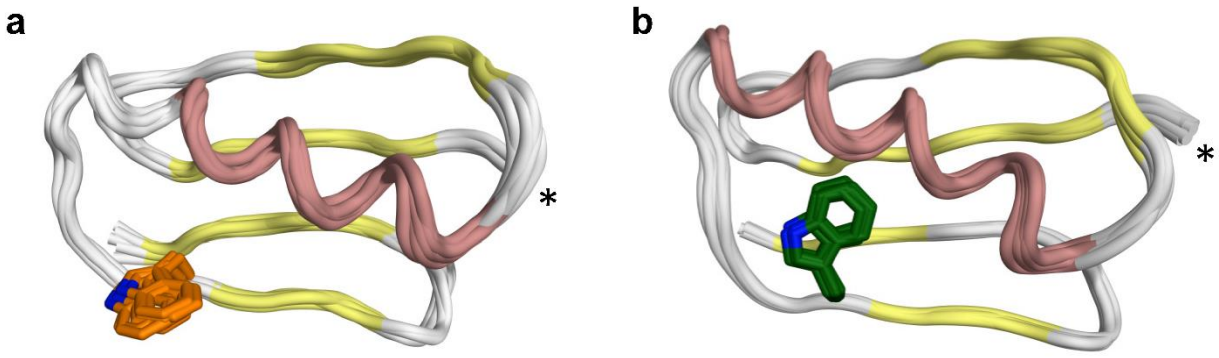


Figure S2.10. Solution NMR structures of NERD variants. NMR ensembles of 10 conformers are shown for **a,** NERD-S with the solvent-exposed *–g*(+) and **b,** NERD-C with the native core-buried +*g*(–) configuration of Trp43 as predicted by *meta*-MSD (Table 1), both of which exhibited structural and thermodynamic properties similar to those of all other mutants studied here (Table 1–2, Supplementary Fig. 4–5), and ^{1}H-^{15}N HSQC spectra showing a single conformation (Supplementary Fig. 7). The Trp43 side chain is shown as sticks, and the N-terminus is indicated by an asterisk. α-helices, β-strands, and loops, as assigned by chemical shift index results, are colored in red, yellow, and white, respectively.

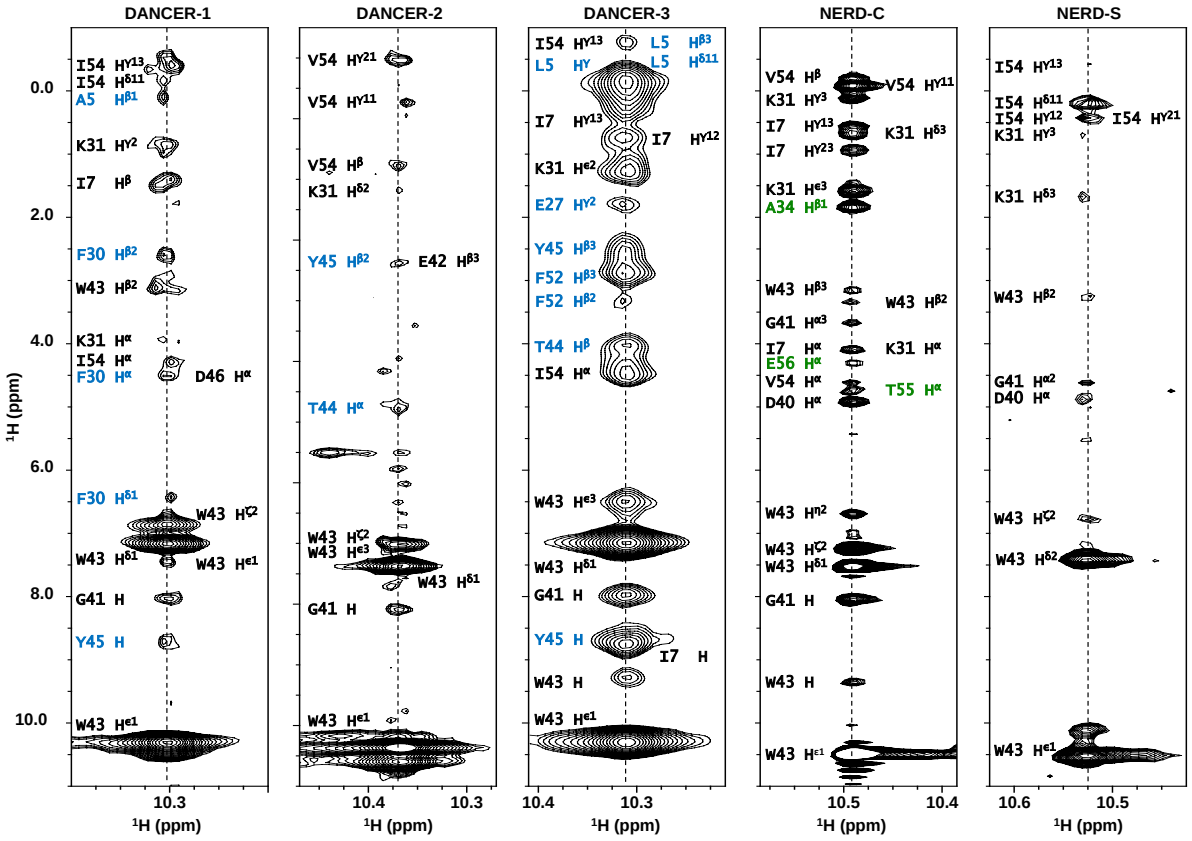


Figure S2.11. Trp43 ^{1}H-^{15}N-NOE correlations observed in Gβ1 variants. Strip plots for the Trp43 indole NH proton from 3D ^{15}N-NOESY spectra acquired with a mix time of 120 ms for dynamic (DANCER-1, DANCER-2, and DANCER-3) and static (NERD-C and NERC-S) variants. NOE assignments are labeled in green, blue, or black, if they are observed in static, dynamic, or both variants. Aromatic side-chain assignments for DANCER-1 and DANCER-3 were inferred from DANCER-2 and NERD-S assignments, respectively.

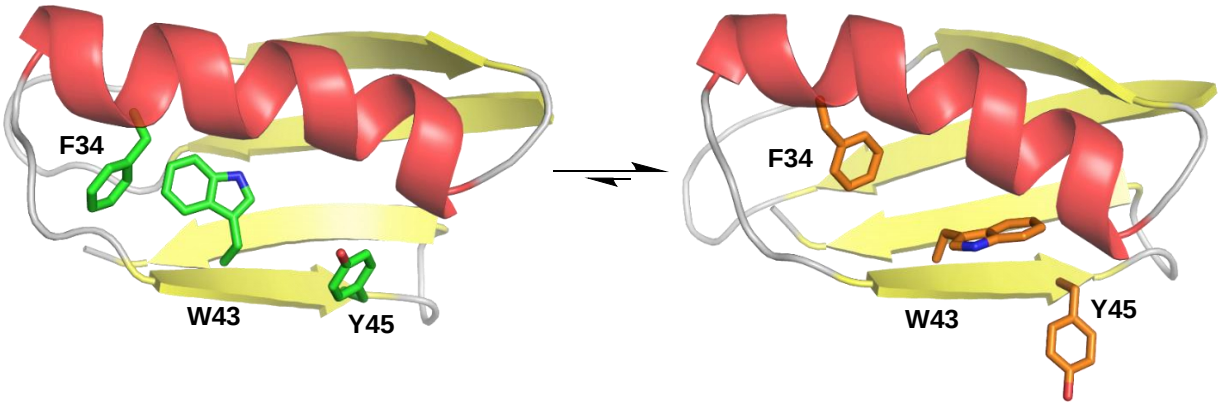


Figure S2.12. Aromatic relay model of exchange in DANCER variants. Analysis of chemical shift changes between major and minor species for DANCER-2, along with the corresponding average solution structures, suggests a model of conformational exchange that involves a relay of aromatic residues nucleated by the change in rotameric state of Trp43. When Trp43 is in its favored solvent-exposed state, Tyr45 also adopts a solvent-exposed configuration, while on the other side Phe34 occupies the hydrophobic core (side chains highlighted in orange). In the minor state, Trp43 interacts with the core and the Tyr45 side chain interacts with the cavity once occupied by the Trp43 indole while Phe34 is excluded from the hydrophobic core to face the solvent (side chains in green). This change is accompanied by a straightening and small displacement of the helix. The amide shift differences observed between major and minor species for DANCER-1 and DANCER-3 also follow a similar pattern, suggesting that all three variants undergo similar types of exchange between major and minor states.

Chapter 3: Origin of dynamics in a small globular protein

Adam M. Damry, Aron Broom, Marc M. Mayer, Natalie K. Goto & Roberto A. Chica

3.1 Statement of contribution

Marc M. Mayer and Adam M. Damry prepared Gβ1 mutants and performed biophysical characterization. Marc M. Mayer assigned chemical shifts for D3-F34A. All other NMR experiments were analyzed by Adam M. Damry. Molecular Dynamics simulations were run and analyzed by Aron Broom. Experimental design was devised by Adam M. Damry, Dr. Natalie K. Goto, and Dr. Roberto A. Chica. The chapter was written by Adam M. Damry and edited by Dr. Roberto A. Chica.

3.2 Introduction

Proteins are molecular machines that carry out complex physical and chemical processes through physical interactions with target molecules. Protein motion is often required throughout the complex pathways underpinning these interactions,[204-206] with dynamics known to occupy a vital role in as varied functions as enzyme catalysis,[32,166] allosteric regulation,[167] and molecular recognition.[168] Homology studies have also demonstrated that function-associated dynamic traits remain highly conserved in enzymes from a shared subfamily.[207] However, despite the demonstrated importance of protein dynamics to protein function, the link between protein sequence and protein dynamics remains poorly understood,[208] and epistasis resulting from the high complexity of the protein energy landscape complicates attempts to study how sequence elements contribute to dynamics and protein function.[209] The rational design of complex protein functions requiring the consideration and engineering of extensive dynamic properties thus remains an unsolved problem. Nonetheless, evolution-based studies have taken steps towards understanding

the molecular basis of how protein dynamics and function coevolve.[18,210,211] Key findings have shown that the development of novel protein functions can arise from new dynamic regimes that reorganize functional sites,[16,35] and do so with relatively few mutations. These studies however often focus on function-expanding dynamics developing in metastable regions of proteins where existing motions responsible for a promiscuous activity can be amplified and modulated with small changes to the global protein sequence and structure.[18,212] The root of how novel dynamics arise in proteins thus remains elusive, given the difficulty of finding an appropriate model system.

In Chapter 2, we rationally designed a series of Streptococcal protein G domain β1 (Gβ1) variants termed DANCERs that exhibit a mode of conformational exchange that is absent in the parent wild-type Gβ1, specifically a spontaneous exchange of Trp43 between two non-native states.[213] The successful generation of three DANCERs demonstrated that novel protein dynamics could be rationally designed. The specific sequence elements that led to these proteins' altered dynamics have yet to be determined however, and thus it is still not known why it is that DANCERs are dynamic while wild-type Gβ1 is rigid. Nonetheless, DANCERs are unique in that the designed Trp43 conformational exchange did not derive from preexisting dynamics in the parent, unlike many other systems that have been used to study the evolution of dynamics.[18,212] Gβ1 and DANCERs therefore comprise a system in which the development of novel dynamics can be traced back to their origin amongst a defined and bounded set of designed mutations by testing the impact of each on dynamics. A deeper understanding of the link between Gβ1 dynamics and sequence is thus critical to explaining how a rigid protein can be made dynamic, a key element to how future designs of dynamics should be carried out, as well as to our understanding of the protein energy landscape.

3.3 Results

3.3.1 Role of mutations in DANCER-3 conformational exchange

Existing Gβ1 DANCERs provide a system in which a novel mode of dynamic conformational exchange was successfully designed through the introduction of five or more core mutations into the sequence of wild-type Gβ1 (Fig. 3.1a).[213] To elucidate the role of these mutations in a pathway leading from the rigid wild-type parent to dynamic DANCERs, we individually reverted each mutation in DANCER-3, chosen due to its compatibility with standard NMR experiments used to study conformational exchange, to the wild-type amino acid at that position (Fig. 3.1b). The resulting five DANCER-3 mutants, D3-F3Y, D3-I7L, D3-F34A, D3-L39V, and D3-I54V, each named for the mutation they possess compared to DANCER-3, all adopted the native Gβ1 fold (Fig. S3.1) and were stably folded (Table 3.1, Fig. S3.2, S3.3). We used solution NMR to assess the dynamic properties of these DANCER-3 mutants, with ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra of four of the five mutants showing evidence for the existence of several distinct conformational states (Fig. S3.4, S3.5). Specifically, spectra for D3-F3Y, D3-I7L, D3-L39V, and D3-I54V showed a population of additional peaks not observed in the wild-type Gβ1 spectrum including an alternate W43ε peak, much like DANCER-3 itself.

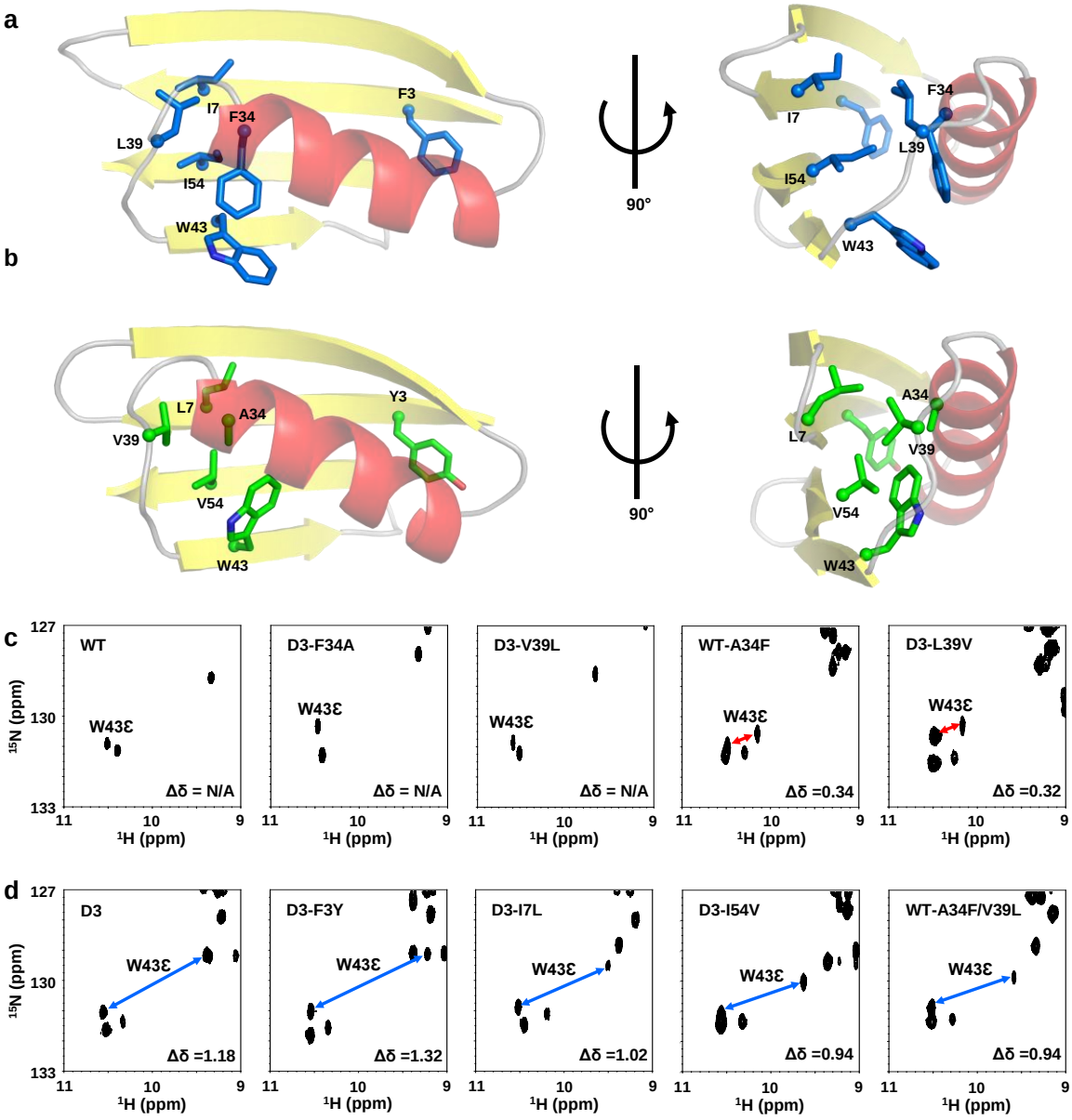


Figure 3.1. Chemical shift displacement analysis of selected Gβ1 variants. a, A model of the DANCER-3 major state structure is shown, with Trp43 and all mutated positions relative to wild-type Gβ1 shown as sticks. α-helices, β-strands, and loops are colored in red, yellow, and grey, respectively. **b,** The wild-type Gβ1 crystal structure (PDB ID: 1PGA)[53] is shown, with Trp43 and the wild-type residues at positions mutated in DANCER-3 shown as sticks. **c, d,** Cutouts from ^{1}H-^{15}N HSQC spectra highlighting the distinctive Trp43ε peaks of selected Gβ1 variants. When present, minor state peaks were identified using ZZ-exchange experiments, shown linked to the major state peak with a double-sided arrow (red for monomer-dimer exchange and blue for Trp43 conformational exchange), and Δδ measurements between major to minor state peaks are inset. Variants were sorted according to observed Δδ, with those possessing no or small (< 0.5 ppm) Δδ shown in **c**, and those possessing large (> 0.5 ppm) Δδ shown in **d**.

Table 3.1. Stability of Gβ1 variants measured through thermal and chemical denaturation.

Protein	Tm [a] (℃)	Cm [b] (M)	ΔG_U [c] (kcal/mol)
WT	81.3	2.66 ± 0.04	4.3 ± 0.3
WT-A34F	57.0	1.02 ± 0.04	2.7 ± 0.2
WT-V39L	73.6	2.32 ± 0.02	4.3 ± 0.4
WT-A34F/V39L	60.5	1.34 ± 0.01	3.0 ± 0.1
D3	70.5	1.72 ± 0.05	3.4 ± 0.4
D3-F3Y	65.5	1.66 ± 0.06	3.3 ± 0.3
D3-I7L	59.5	1.30 ± 0.03	2.8 ± 0.4
D3-F34A	84.2	2.74 ± 0.07	4.3 ± 0.1
D3-L39V	67.5	1.75 ± 0.05	2.9 ± 0.1
D3-I54V	64.6	1.82 ± 0.06	3.2 ± 0.4

[a] Thermal denaturation midpoint temperature determined through loss of circular dichroism signal at 208 nm
b Chemical denaturation midpoint denaturant concentration
c Free energy of unfolding determined by chemical denaturation with guanidium chloride at 25 ℃ (n=3, mean ± s.d.)

Before testing for the presence of Trp43 conformational exchange in these mutants, we noted that D3-I7L presented a poorly populated Trp43 minor state, with a minor state peak volume of <5 % compared to the major state peak. To improve spectral quality, we raised the concentration of D3-I7L in the sample from 0.2 mM to 0.5 mM, however against expectations, the intensity of the Trp43 minor state peak did not commensurately increase. This observation led us to investigate the possibility of dimer formation, as it is known that the A34F mutation induces dimer formation in wild-type Gβ1.[56,177] Size-exclusion chromatography using wild-type Gβ1 and the A34F point mutant (WT-A34F) as controls for monomeric and dimeric Gβ1 respectively reveals a ~1 mL shift in elution volume indicative of dimer formation for all proteins tested containing the A34F mutation (Fig. S3.6). K_d determination experiments (Methods, Table S3.1) further corroborate that all mutants tested save D3-F34A form partial dimers over the standard range of concentrations used for NMR.

NOE data from previous work has shown that the contacts made by Trp43 cannot be explained by a single Trp43 conformation in the context of the native Gβ1 fold,[213] which is maintained in the A34F-mediated dimer.[56] The presence of this dimeric species therefore adds complexity to our study of conformational exchange in DANCERs only insofar as NMR experiments probing dynamic exchange will report on both Trp43 conformational exchange and monomer-dimer exchange. Moreover, as NMR reports on local dynamics on a residue by residue basis, and the Gβ1 dimer interface lies on the opposite face of the protein from Trp43 (Fig. S3.7), NMR remains an ideal technique for characterizing Trp43 dynamics. Using WT-A34F as a negative control and DANCER-3 as a positive control for Trp43 conformational exchange, we observed a small chemical shift displacement (Δδ) of 0.34 ppm for the Trp43 side-chain NH in WT-A34F (Fig. 3.1c), in contrast to the Δδ of 1.18 ppm seen in DANCER-3 (Fig. 3.1d). This corresponds to an expected small change in Trp43 chemical environment when the non-exchanging WT-A34F variant dimerizes in comparison to a much larger change when Trp43 side-chain conformational exchange occurs in DANCER-3. Of the four DANCER-3 single mutants tested (as D3-F34A forms only one population of peaks thus does not exhibit any form of dynamic exchange), D3-F3Y, D3-I7L, and D3-I54V demonstrated a Δδ closer to that of DANCER-3 than WT-A34F, with Δδ values ranging from 0.94 to 1.32 ppm. D3-L39V on the other hand resembled WT-A34F, with a Δδ of 0.32 ppm. This suggested that D3-L39V may not exhibit the Trp43 conformational exchange characteristic of DANCERs.

Further validation was performed using a ^{1}H-^{15}N HSQC ZZ-exchange experiment contrasting rates of Trp43 dynamic exchange to those of Thr17, a Gβ1 dimerization reporter (Fig. S3.7-S3.9). In the non-exchanging WT-A34F control as well as in D3-L39V, increasing temperature caused a similar increase in exchange rates detected by both Thr17 and Trp43,

suggesting that both residues detect only monomer-dimer exchange (Table 3.2). Strikingly, a similar increase in temperature in DANCER-3 leads to the expected increase for all exchange rates detected by Thr17, while the minor state to major state exchange rate detected by Trp43 decreases instead (Fig. 3.2a). This marked difference confirms that Trp43 conformational exchange is occurring independently from monomer-dimer exchange, albeit with non-Arrhenius behavior suggesting a complex exchange trajectory (Fig. 3.2b). Specifically, as increasing temperature favors the dimeric form of DANCER-3, it is possible that Trp43 conformational exchange is inhibited in the dimer. Previously observed results with the L39I mutant of DANCER-3 (D3-L39I, previously reported as NERD-S) corroborate this hypothesis, as D3-L39I exhibits no detectable Trp43 conformational exchange at the concentration range used in our NMR experiments despite significant sequence and spectral similarity to DANCER-3,[213] but unlike DANCER-3 adopts an almost exclusively dimeric form at these concentrations (Table S3.1). Similar non-Arrhenius behavior to DANCER-3 Trp43ε was also observed for Trp43ε in D3-F3Y, D3-I7L, and D3-I54V (Table 3.2, Fig. S3.10), confirming that these three mutants exhibit DANCER-like Trp43 conformational exchange, while D3-F34A and D3-L39V do not. ^{1}H-^{15}N HSQC CPMG experiments[138] with concentration modulation were also used to validate that Trp43 conformational exchange is independent from monomer-dimer exchange in DANCER-1 and DANCER-2 (Fig. S3.11), which are poorly behaved in ZZ-exchange experiments. CPMG reveals that DANCER-1 and DANCER-2 both demonstrate altered relaxation-dispersion properties for Trp43 compared to Thr17, while WT-A34F and D3-V39L do not, confirming that independent Trp43 conformational exchange occurs in the DANCERs, and not in WT-A34F or D3-V39L.

Table 3.2. Monomer-dimer and tryptophan conformational exchange kinetics**.**

Sequence	Peaks	T (°C)	$k_{1\rightarrow2}$ [a] (s^{-1})	$k_{2\rightarrow1}$ [b] (s^{-1})	Arrhenius Behavior [c]
WT-A34F	W43ε	9	14 ± 1	15 ± 2	Yes
		12	16 ± 2	17 ± 1	
		15	19 ± 2	21 ± 2	
	T17	9	10 ± 1	16 ± 1	Yes
		12	12 ± 1	19 ± 2	
		15	14 ± 1	25 ± 2	
D3-L39V	W43ε	12	10 ± 1	25 ± 2	Yes
		15	14 ± 2	35 ± 2	
		18	19 ± 1	44 ± 4	
	T17	12	5.2 ± 0.4	9.0 ± 0.8	Yes
		15	7.9 ± 0.9	13 ± 1	
		18	10 ± 1	17 ± 2	
DANCER-3	W43ε	12	2.8 ± 0.4	39 ± 3	No
		15	3.2 ± 0.3	22 ± 3	
		18	3.8 ± 0.6	15 ± 1	
	T17	12	4.9 ± 0.7	25 ± 3	Yes
		15	5.6 ± 0.4	31 ± 3	
		18	6.3 ± 0.6	39 ± 4	
D3-F3Y	W43ε	12	2.7 ± 0.4	3.7 ± 0.6	No [d]
		25	3.3 ± 0.4	4.1 ± 0.7	
		30	3.6 ± 0.5	4.4 ± 0.7	
	T17	12	1.5 ± 0.3	2.1 ± 0.5	Yes
		25	4.2 ± 0.5	7 ± 1	
		30	6.0 ± 0.8	11 ± 2	
D3-I7L	W43ε [e]	N/A	N/A	N/A	N/A
	T18	12	3.0 ± 0.4	6.1 ± 0.5	Yes
		15	6.1 ± 0.6	11 ± 1	
		18	9.1 ± 0.7	15 ± 1	
D3-I54V	W43ε	12	17 ± 2	17 ± 2	No
		15	17 ± 3	17 ± 2	
		18	17 ± 2	17 ± 2	
	T17	12	21 ± 2	22 ± 2	Yes
		15	26 ± 2	28 ± 3	
		18	29 ± 4	35 ± 3	
WT-A43F/ V39L	W43ε	12	19 ± 2	19 ± 2	No
		15	20 ± 2	18 ± 2	
		18	21 ± 2	16 ± 1	
	T18	12	20 ± 2	25 ± 2	Yes
		15	24 ± 3	30 ± 3	
		18	28 ± 3	39 ± 4	

[a] Rate constant for exchange from the major state (the most populated state at the lowest temperature tested) to the minor state (the least populated state at the lowest temperature tested) (n = 2 analytical replicates, mean ± s.d.)

[b] Rate constant for exchange from the minor state to the major state (n = 2 analytical replicates, mean ± s.d.)

[c] Arrhenius behaviour is defined here as a measurable increase in both rates of exchange as temperature rises

[d] Although D3-F3Y W43ε rates of exchange rise with temperature, this increase is very weak compared to T17 (Fig. S3.7). As both rates are on the same order of magnitude, it is expected that for one process, the increases would be similar. This is not the case, thus we have classified D3-F3Y W43ε as exhibiting non-Arrhenius behavior.

[e] W43ε minor state peaks and crosspeaks could not be quantified for D3-I7L but are present

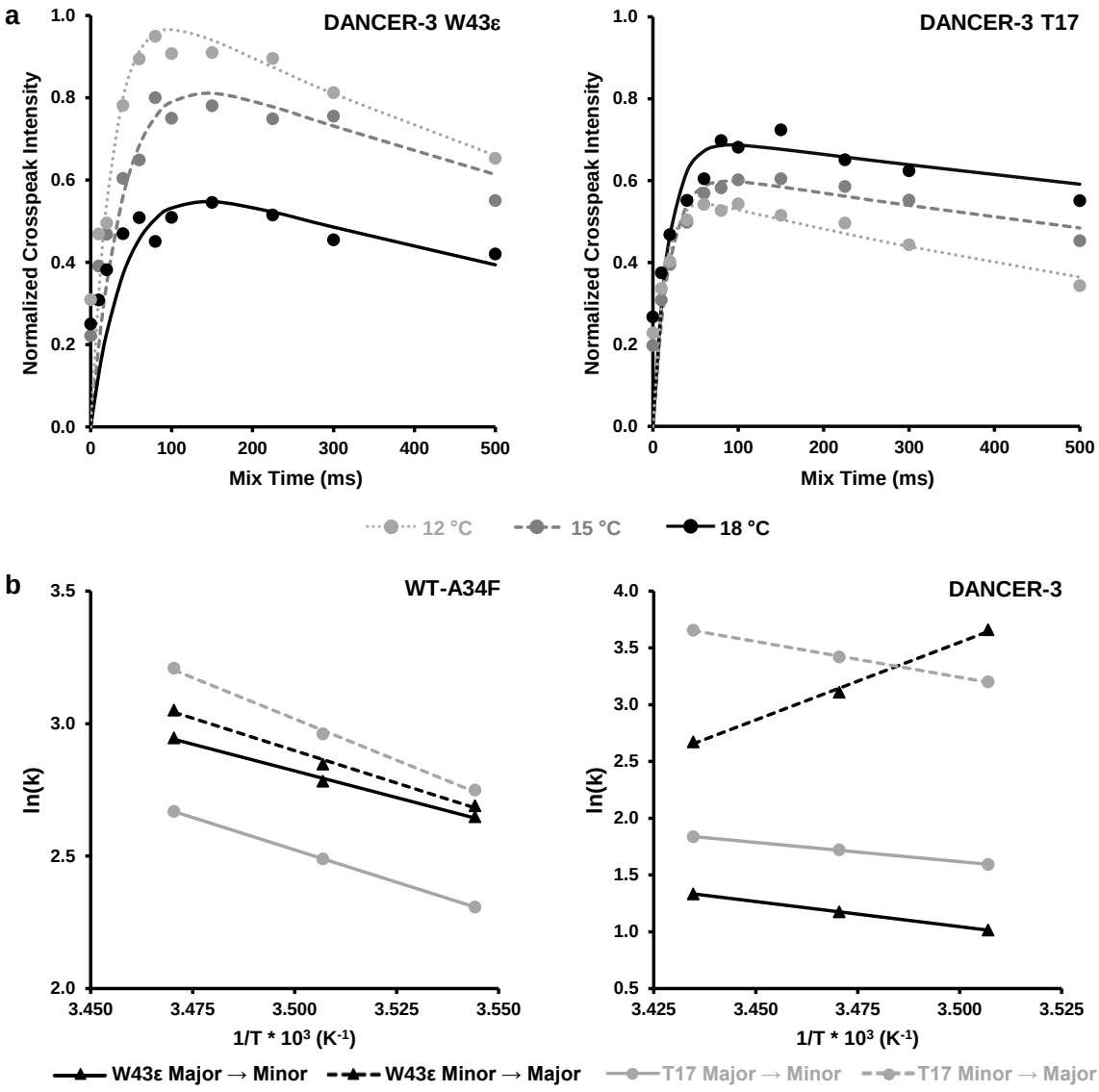


Figure 3.2. Trp43 conformational exchange manifests as non-Arrhenius behavior in DANCERs. a, ^{1}H-^{15}N HSQC ZZ-exchange crosspeak intensity curves for the Trp43ε minor to major state transition show that this transition slows in DANCER-3 as temperature increases. This unusual behavior is not recapitulated with Thr17, demonstrating a different mode of exchange for Trp43ε than monomer-dimer exchange. **b,** Arrhenius plots demonstrate Arrhenius behavior for both WT-A34F Trp43ε and Thr17 exchange profiles as shown by similar activation energies for all four transitions. DANCER-3 however demonstrates non-Arrhenius behavior where the Trp34ε minor to major state transition displays a non-physical negative activation energy indicative that it does not follow a simple two-state Arrhenius model, as opposed to the normal Arrhenius behavior demonstrated by Thr17.

3.3.2 The A34F mutation is necessary to escape the wild-type Gβ1 energy well

Having identified A34F and V39L as two key mutations required for Trp43 conformational exchange, we next wanted to understand how each is involved in mediating this exchange. Key to this was the effect of each of these mutations in the context of the wild-type Gβ1 sequence, thus free from convoluting effects attributable to other mutations. With the function of the A34F mutation in wild-type Gβ1 already known, we introduced the V39L point mutation into wild-type Gβ1 (WT-V39L). This mutant was folded like Gβ1, stable, and monomeric (Fig. S3.1 – S3.3, S3.6), and generated an HSQC with high spectral similarity to that of wild-type Gβ1 (Fig. S3.5, S3.12), suggesting that its Trp43 conformation is that of the wild-type and showing that the V39L mutation alone is insufficient to dislodge Trp43 from the highly stable wild-type conformation. With this in mind, we re-examined D3-F34A. This mutant's HSQC confirms that it adopts only a single conformation, yet it still possesses all four other mutations present in DANCER-3, which has been shown to adopt two non-native Trp43 conformations.[213] To confirm whether the single Trp43 conformation present in D3-F34A resembled either of the DANCER-3 non-native conformations, or rather the wild-type conformation, we solved the solution structure of D3-F34A by NMR (Fig. S3.13, Table S3.3), and found that this mutant adopts the wild-type Trp43 conformation. This confirmed that even with four other mutations present, the A34F mutation is critical for Gβ1 to adopt any of the non-native Trp43 conformations observed in DANCERs.

3.3.3 Subtle alterations in protein core packing give rise to new dynamic motions

The A34F mutation alone is however not sufficient to produce a DANCER. As shown above, the V39L mutation is also necessary for Trp43 conformational exchange to occur in the context of the DANCER-3 sequence. To test whether the interaction between Phe34 and Leu39 had given rise to Trp43 conformational exchange, or if complex interactions between these and

other DANCER-3 mutations were also required, we introduced both the A34F and V39L mutations into wild-type Gβ1 (WT-A34F/V39L). The resulting protein was folded like Gβ1, stable, and partially dimeric at the concentrations used (Fig. S3.1, S3.2, S3.4, Table S3.1). Its HSQC spectrum demonstrates the presence of two peak populations (Fig. S3.3) with a $\Delta\delta$ of 0.94 ppm for the Trp43 side-chain peak and ZZ-exchange spectra exhibiting non-Arrhenius behaviour (Fig. S3.8 – S3.10), indicating that Trp43 conformational exchange does occur in WT-A34F/V39L and that these two mutations are responsible for the designed dynamics in DANCER-3.

As we could not however easily explain the mechanism by which the conservative V39L mutation contributed to DANCER dynamics, we ran molecular dynamics (MD) simulations on model structures of DANCER-3, D3-F34A, D3-L39V, and D3-L39I to probe the interactions made by the F34 and L39 residues and the rest of the protein. By plotting the Trp43 χ_1 and χ_2 dihedral angles as a function of simulation time (Figure S3.14a), distinct conformational clusters arise for each variant (Figure S3.14b), and the number of transitions between these clusters is indicative of the dynamicity of the protein in question, with DANCER-3 undergoing the most transitions and D3-F34A the least (Fig. 3.3a). Notably, D3-L39V is shown to be significantly less dynamic than DANCER-3, paralleling results from our NMR experiments. D3-L39I however presents an intermediate dynamicity in the monomeric form used for simulations, unlike what was observed in NMR for its dimeric form, further validating that only the monomeric form of DANCERs is capable of undergoing Trp43 conformational exchange.

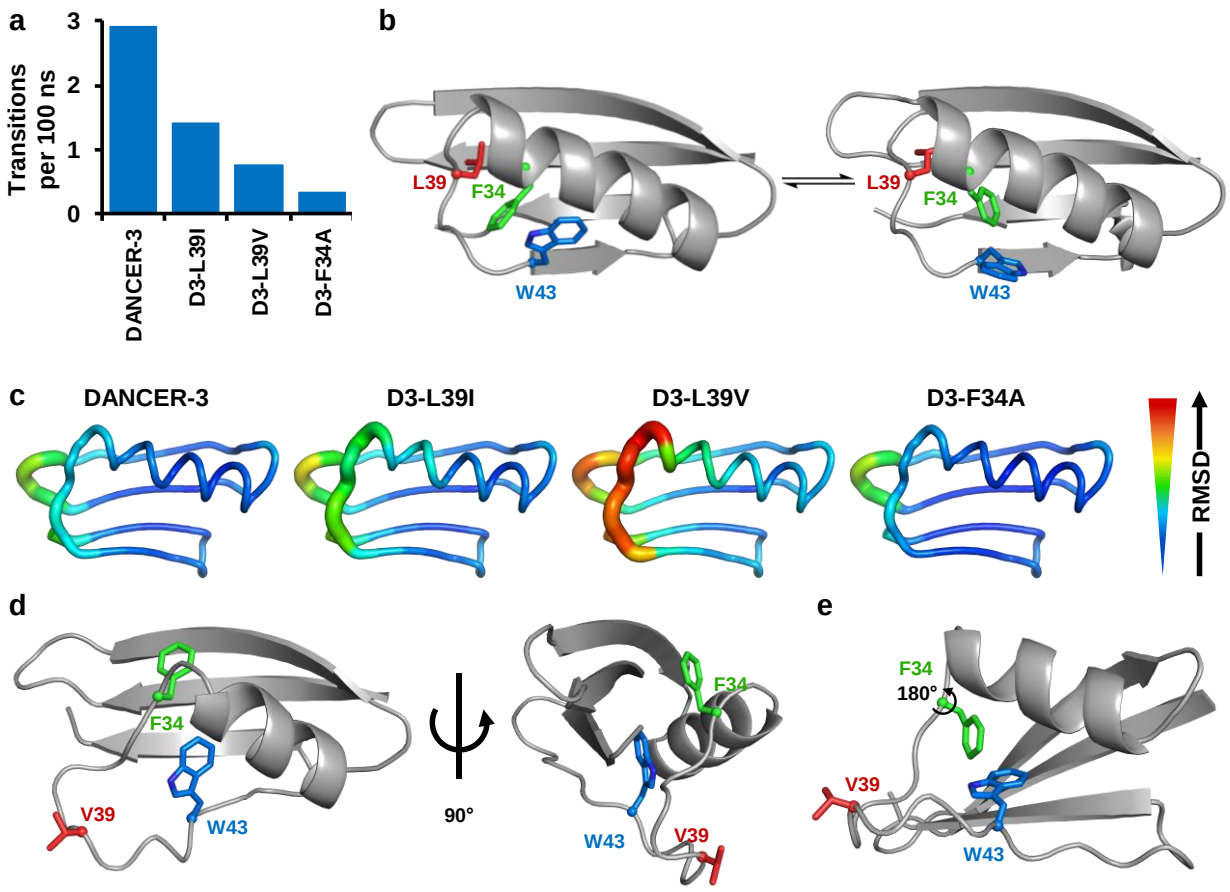


Figure 3.3. Molecular dynamics simulations suggest mechanisms by which the A34F and V39L mutations gate dynamics. a, The number of Trp43 conformational transitions occurring per unit of time for each simulation represents a measure of dynamicity. Conformational transitions represent a change in χ_1 or χ_2 over successive frames where χ_1 passes from one canonical conformation (gauche$^+$, gauche$^-$, or trans) to another, or χ_2 undergoes a sign alternation. **b,** Representative frames sampled from the DANCER-3 simulation for the gauche$^+$ and gauche$^-$ Trp43 conformations demonstrating the concerted motion Phe34 undergoes over the Trp43 conformational exchange pathway. **c,** Backbone RMSD over each simulation plotted on a model backbone for each respective protein. Increased RMSD seen in D3-L39V at the C-terminal extremity of the α-helix and in the loop containing V39 is indicative of increased mobility and ability to adopt alternate backbone conformations. **d,** D3-L39V populates a non-native backbone conformation comprising a destabilized α-helix C-terminus and untethered V39 loop. **e,** Phe34 is incapable of displacing Trp43 while satisfying the backbone conformation described in (d) as shown by a model demonstrating that a 180° χ_1 rotation of the Phe34 side-chain could be allowed without significant displacement of the Trp43 side-chain.

These MD simulations also allow us to see what other motions occur throughout the protein structure alongside Trp43 conformational exchange (defined here as a transition between gauche$^+$ and gauche$^-$ / trans conformations). One mode of Phe34 motion occurs in concert with Trp43

conformational exchange (Fig. 3.3b), where the Phe34 side-chain moves to occupy the space vacated when the Trp43 side-chain rotates out of the protein core. In D3-F34A, the smaller Ala34 side-chain cannot serve this purpose, and therefore Trp43 conformational exchange does not occur, remaining instead trapped in the wild-type gauche$^-$ conformation. The role of Leu39 is less clear, as it exhibits no obvious transitions concerted with Trp43 motion. What we do observe however is that D3-L39V exhibits enhanced motion in the loop containing position 39 compared to DANCER-3, D3-F34A, and D3-L39I (Fig. 3.3c). Moreover, D3-L39V commonly occupies a conformation that is only seldom observed in other examined variants, where the C-terminal portion of its helix is destabilized and partially unfolded (Fig. 3.3d), reminiscent of the loss of helicity observed in this region of the protein in the WT-A34F crystal structure (Fig. S3.7). In this conformation, the backbone is not conducive to a Phe34 conformation that would allow it to displace Trp43, allowing more free rotation of Phe34 independent of Trp43 (Fig. S3.14c) and suggesting why Trp43 conformational exchange is inhibited in D3-L39V. In DANCERs on the other hand, the L39 (or I39) residue tethers this loop in a backbone conformation that allows the concerted motion of Phe34 and Trp43.

3.4 Discussion

Together with NMR characterization, our MD simulations suggest how Trp43 conformational exchange takes place in DANCERs. Specifically, the Phe34 side-chain moves in concert with Trp43 to fill the space vacated by Trp43, while Leu39 (or Ile39) stabilizes the exchange pathway by preventing the C-terminal segment of the α-helix from the adopting non-productive conformations observed in MD simulations of D3-V39L. Without this stabilizing effect, brought about by the addition of a single specific methylene group to one specific amino acid, Trp43 conformational exchange does not occur, demonstrating how subtle structural changes

to key regions of a protein can give rise to seemingly disproportionate effects on the global structure and dynamics of the protein. This incredible plasticity arises from the ruggedness of the protein energy landscape, wherein a small sequence change can take a protein out from one energy well and into a different local minimum.[214] Even so, very few conservative mutations do go on to cause wide-reaching effects or even local disruptions to native protein structure, as was observed with the Y3F, L7I, and V54I mutations, none of which were primary contributors to the designed Trp43 conformational exchange pathway. The V39L mutation, which allows Trp43 conformational exchange to occur when introduced into WT-A34F yet is silent in the context of wild-type Gβ1, also highlights how the effect of a mutation on dynamics is complex and context-dependent. In fact, introducing all four of these mutations together in the wild-type (D3-F34A) was insufficient to shift Trp43 from its native conformation, demonstrating that for all its ruggedness, deeper energy wells also exist on the protein energy landscape that can trap proteins in a single conformation, as was the case for the rigid wild-type Gβ1. Here, the introduction of a bulky phenylalanine side-chain to disrupt the protein core was needed to escape the wild-type conformation's energy well, as opposed to the smaller effects of the other more conservative mutations. Though destabilizing, the A34F mutation pushed Gβ1 from its wild-type energy well, as evidenced by the alternate backbone conformation and dimer formation observed in WT-A34F. Only then was the subtler perturbation introduced by the V39L mutation capable of causing the emergence of Trp43 conformational exchange.

Our results show however that destabilization alone was not what allowed our DANCERs to exhibit Trp43 conformational exchange. Though all our DANCERs are less stable than wild-type Gβ1, our stability measurements (Fig. S3.2) show that they are also all more stable than the non-exchanging WT-A34F mutant. In fact, introducing the V39L mutation, though destabilizing

in wild-type Gβ1, both introduced Trp43 conformational exchange into WT-A34F and caused a modest but statistically significant increase in protein stability, suggesting that this mutation stabilizes the alternate protein conformations found in DANCERs. DANCER-3 is also nearly as stable as the non-exchanging and wild-type-like WT-V39L mutant, showing that the mutations introduced do not simply overpack the protein core. Our results therefore suggest that destabilization of the wild-type conformation and stabilization of the exchange trajectory we sought to design were both needed to obtain a DANCER.

In summary, the dynamic nature of DANCERs arose from the introduction of two mutations into wild-type Gβ1. The first, A34F, was needed to destabilize the wild-type conformation sufficiently for the protein to reach a metastable state, with the second, V39L, only then being capable of mediating the novel conformational exchange pathway developed. The specific effects of each of these mutations on the Gβ1 energy landscape (Fig. 3.4) however lends insight into the mechanisms by which novel dynamics might evolve in proteins both natural and engineered. Though WT-V39L was relatively uninteresting from the perspective of studying dynamics, as it strongly resembles wild-type Gβ1, evolution often favors silent mutations over ones that bring about unwanted and potentially deleterious secondary effects[215] such as the dimerization caused by the A34F mutation. Thus, if this function were to evolve naturally, we could expect to see the V39L mutation arising before the A34F mutation, acting as a silent precursor to the evolution of these novel dynamics. While the V39L mutation does not displace the mutant protein from its wild-type energy well, it does shape the surrounding energy landscape, allowing for novel dynamics and novel function to emerge only once the final perturbing mutation is introduced, and highlighting why tracing back the roots of novel function and dynamics through a protein's evolution can be difficult.

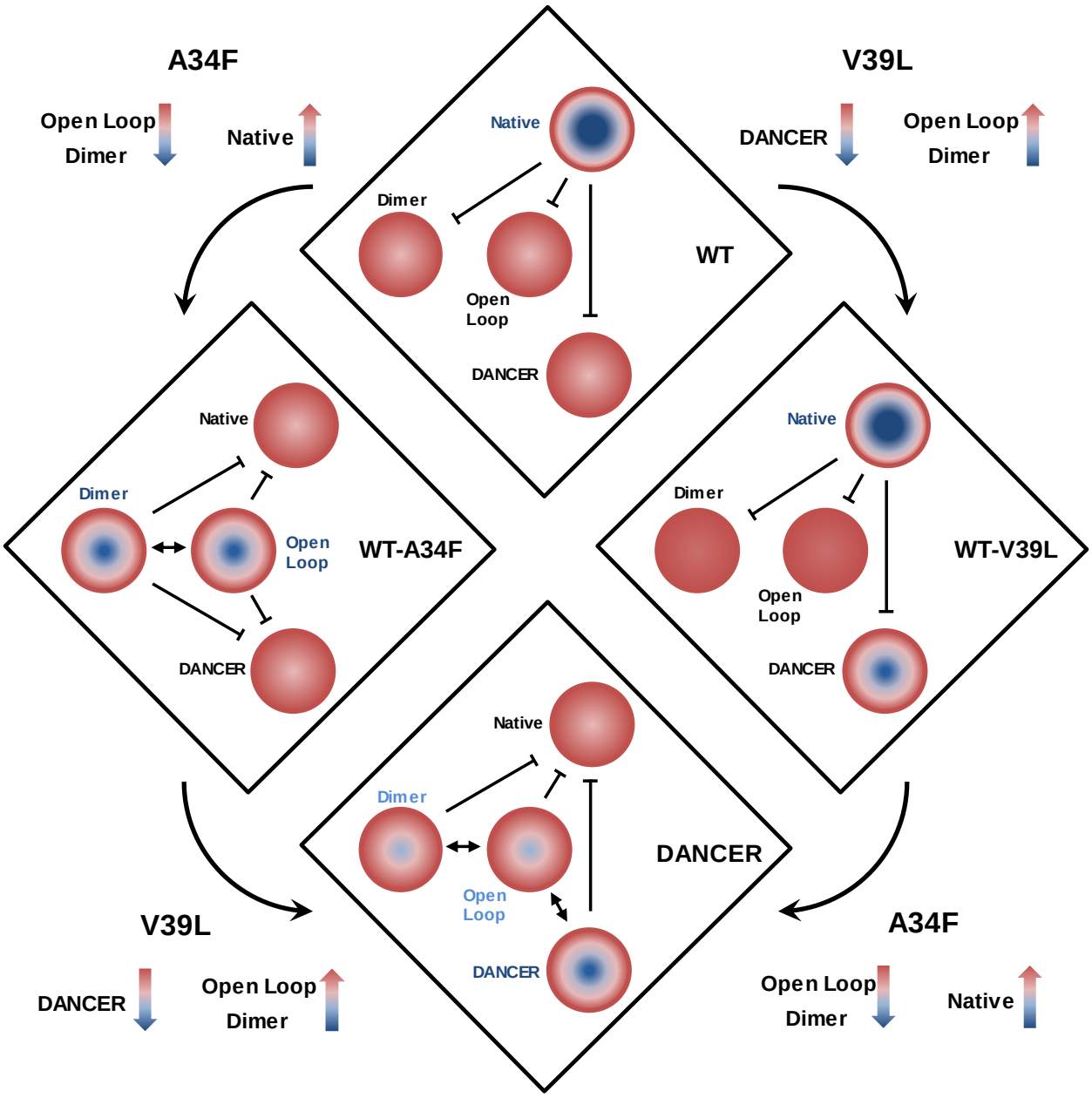


Figure 3.4. Summary of mutational effects on the Gβ1 energy landscape. Major conformational states sampled by Gβ1 variants are shown on a hypothetical energy landscape. Hypothetical energy well depths are based on NMR and MD data suggesting which states can be sampled by each respective variant and observable equilibria between states. Though DANCERs adopt two monomeric states, both are separate from either the native or altered loop states, and are therefore being shown as one state for simplicity.

Coming back to protein design however, we sought to answer the question of how novel protein dynamics might be engineered. The difficulty of this task is clear, with the critical V39L

mutation being nearly impossible to predict through visual inspection of the Gβ1 structure. Designing novel dynamics into a naturally rigid region of a protein is not a simple question of packing, overpacking, or destabilizing a protein, but rather requires the introduction of several mutations that destabilize unwanted conformational states and stabilize desired ones. Despite the complexity of the protein energy landscape however, DANCER sequences were successfully predicted by a computational protein design methodology, and our molecular dynamics simulations also correctly identified these positions as important to DANCER dynamics. These results thus highlight how modern computational tools are becoming necessary to the engineering of increasingly complex protein functions, and how by using them as tools to reshape the protein energy landscape, we might open the door to the design of proteins with a wider range of functions than previously possible.

3.5 Methods

3.5.1 Protein expression and purification

Codon-optimized and His-tagged genes for wild-type Gβ1 and DANCER-3 cloned into the pJ414 vector were obtained from DNA2.0. All other variants tested were generated through splicing by overlap extension mutagenesis using either the wild-type Gβ1 or DANCER-3 as starting templates and cloned into the pET-11a vector. Proteins for chemical denaturation assays and circular dichroism measurements were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using Luria-Bertani (LB) broth supplemented with 100 μg/mL ampicillin. Proteins for NMR spectroscopy were instead expressed using M9 minimal expression medium supplemented with 1 g/L ^{15}N-ammonium chloride and/or 3 g/L ^{13}C-D-glucose for isotopic enrichment. Cultures were grown at 37 °C with shaking to an OD600 of approximately 0.6 after which protein expression was initiated with 1 mM isopropyl β-D-1-thiogalactopyranoside. Following overnight incubation

at either 15 °C or 37 °C with shaking (250 rpm) for cultures grown in LB or M9 medium, respectively, 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 (Qiagen), followed by gel filtration in 10 mM sodium phosphate buffer (pH 7.4) using an ENrich SEC 650 size-exclusion chromatography column (BioRad). Purified samples were concentrated using Amicon Ultracel-3K centrifugal filter units (EMD Millipore).

3.5.2 Thermal denaturation assays

Circular dichroism measurements were performed with a Jasco J-815 spectrometer using 650-μL aliquots of each Gβ1 sample at a concentration of 40 μM in 10 mM sodium phosphate buffer (pH 7.4). Samples placed in a 1-mm path-length quartz cuvette (Jasco) were heated at a rate of 1 °C per minute, and ellipticity at 208 nm was measured every 2 °C. Tm values were determined by fitting a 2-term sigmoid function with baseline correction[198] using nonlinear least-squares regression. Reversibility was confirmed by comparing circular dichroism spectra acquired before and after thermal denaturation experiments from 185–250 nm at 25 °C.

3.5.3 Chemical denaturation assays

Chemical denaturation assays were performed in triplicate using protein samples at a 1 mg/mL concentration. Protein aliquots of 25 μL in individual wells of UV-Star 96-well plates (Greiner Bio-One) were mixed with 175 μL of 0–5 M guanidium chloride solutions (12 points, evenly spaced) and incubated at room temperature for an hour. Fluorescence emission spectra were measured for each sample from 300 nm to 450 nm (excitation at 295 nm and step size of 2 nm) using an Infinite M1000 plate reader (Tecan). Fluorescence was integrated and converted into

fraction of unfolded protein values. Error on these values is reported as standard deviation from three replicates at each denaturant concentration for each Gβ1 variant. Cm values (concentration of denaturant at midpoint of denaturation) were determined by fitting a 2-term sigmoid function using nonlinear least-squares regression.

3.5.4 NMR spectroscopy

^{15}N- and ^{13}C-labelled Gβ1 samples for NMR consisted of 0.1–2.0 mM uniformly labelled protein in 10 mM sodium phosphate buffer (pH 7.4), 10 μM EDTA, 0.02% sodium azide, 1× cOmplete EDTA-free Protease Inhibitor Cocktail (Roche), and 10% D_2O for experiments requiring detection of amide protons or 99% D_2O otherwise. All NMR experiments were performed on a Bruker AVANCEIII HD 600 MHz spectrometer equipped with a triple resonance cryoprobe. HSQC, CPMG, chemical shift assignment, and NOESY experiments were performed at 25 °C, and ZZ-exchange experiments were performed at temperatures varying from 5 °C to 25 °C. NMR data sets were processed with the NMRPipe software package[199] and spectra were analyzed with NMRViewJ (One Moon Scientific).[200] 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 for backbone assignments, and NH-TOCSY (with mixing time (τ_{mix}) = 75 ms), CCH-TOCSY (τ_{mix} = 32.5 ms) and HCCH-TOCSY (τ_{mix} = 27.5 ms) spectra for side-chain assignments of D3-F34A. Dimer K_d values were calculated using peak volumes for linked major and minor state peaks for Thr17 taken at four different concentrations per protein sample tested. Average backbone amide chemical shift differences (Δδ) between W43ε major and minor states for were calculated using $\Delta\delta = ((\Delta\delta_{HN})^2 + (\Delta\delta_N/5)^2)^{0.5}$ where $\Delta\delta_{HN}$ and $\Delta\delta_N$ are chemical shift differences calculated for protons and nitrogen atoms, respectively. To obtain D3-F34A distance restraints, ^{15}N-edited and ^{13}C-edited HSQC-NOESY

(τ_{mix} = 120 ms) spectra were acquired. ZZ-exchange spectra were acquired at temperatures between 9 °C and 30 °C for which Trp43ε and Thr17 peaks could be deconvolved from neighbouring peaks (or Trp43ε and Thr18 peaks for WT-A34F/V39L and D3-I7L as Thr17 crosspeaks could not be deconvolved at any temperature), and analyzed by fitting a Gaussian model to peaks in distinct clusters and integrating using NMRDraw.[199] Rates of exchange were determined by fitting 4-term relaxation and exchange curves[137] using nonlinear least-squares regression, and thermodynamic parameters were determined by fitting exchange rates and temperatures to the Eyring equation.

3.5.5 Structure determination

TALOS+[132] was used to determine secondary structure propensities and backbone dihedral restraints for D3-F34A on the basis of measured chemical shifts for $^{1}H_{\alpha}$, ^{15}N, $^{13}C'$, $^{13}C_{\alpha}$, and $^{13}C_{\beta}$ chemical shifts. Simultaneous NOE assignment and structure calculation was performed using CYANA 2.1.[135] Chemical shifts from cross peaks in 3D ^{15}N-edited and ^{13}C-edited NOESY-HSQC spectra were used as input, in addition to TALOS+ derived dihedral angle restraints. A total of 618 unique and non-redundant distance restraints were used to calculate 100 conformers, and NMR ensembles were represented by the 10 lowest energy conformers. No distance violations >0.5 Å or torsion angle violations >5 ° were observed in the resulting structures.

3.5.6 Molecular dynamics simulations

Structures of DANCER-3, D3-L39I, D3-L39V, and D3-F34A were modeled from the NMR structure of D3-L39I by introducing mutations using TRIAD (Protabit LLC, Pasadena CA). Prior to molecular dynamics, the structures were parameterized with the AMBER FF14SB forcefield using the LEaP program from the amber suite (http://ambermd.org/). The protein was

surrounded in a cubic box of TIP3P water with 6 angstroms as the shortest clearance on a single side, and the electrostatic charge neutralized by addition of sodium ions.

Molecular dynamics simulations were performed using in-house code written with the OpenMM API (version 7, http://openmm.org/). A Langevin thermostat was used with a temperature of 298.15 K and a timestep of 2 fs. A Monte Carlo barostat was employed with a pressure of 1 atm, periodic boundary conditions and particle mesh Ewald summation with a long-range cutoff of 12 angstroms. The lengths of all bonds to hydrogen atoms were constrained. Energy minimization was performed until a convergence of 10 kJ/mol, followed by 100 ns of equilibration. Following equilibration, 5 production replicas were run for 3 μs each, for a total of 15 μs of simulation time per protein. All analysis of the molecular dynamics trajectories was performed using VMD (www.ks.uiuc.edu/Research/vmd/) on the final 2 μs of each replica trajectory for each protein. Thus, a total of 10 μs simulation time per protein was used for analysis.

3.5.7 Data availability

Structure coordinates for D3-F34A have been deposited in the Protein Data Bank with the accession code 6NJF. NMR data for D3-F34A has been deposited in the Biological Magnetic Resonance Data Bank with the accession code 30532.

3.6 Supplementary Information

Table S3.1. Dimer dissociation constants of dimeric Gβ1 variants

Sequence	Dimer K_d (μM)
WT-A34F	27 ± 4 [56]
DANCER-1	560 ± 40
DANCER-2	1180 ± 60
DANCER-3	43 ± 3
D3-L39I	16 ± 2
D3-F3Y	66 ± 5
D3-I7L	62 ± 8
D3-L39V	31 ± 3
D3-I54V	240 ± 20
WT-A34F/V39L	116 ± 5

Table S3.2. CPMG analysis results.

Protein	Peak	Concentration	CPMG Ratio [a]	Concentration Effect [b]
DANCER-1	T17	High	3.21	1.84
		Low	1.74	
	W43ε	High	1.45	0.83
		Low	1.76	
DANCER-2	T17	High	3.11	2.01
		Low	1.54	
	W43ε	High	1.58	0.92
		Low	1.72	
WT-A34F	T17	High	0.97	0.94
		Low	1.02	
	W43ε	High	0.99	0.92
		Low	1.08	
D3-L39V	T17	High	1.11	0.95
		Low	1.17	
	W43ε	High	1.13	0.96
		Low	1.17	

[a] Ratio of peak intensities for a given peak in the 1000 Hz spectrum relative to the same peak in the 20 Hz spectrum

[b] Ratio of CPMG ratios at high and low protein concentrations for a given peak. As monomer-dimer exchange rates should be more heavily influenced by concentration than Trp43 conformational exchange rates, we expect to see different concentration effects for Trp43ε as for Thr17 in variants capable of undergoing Trp43 conformational exchange.

Table S3.3. Summary of NOE restraints and structural statistics [a]

PDB ID	6NJF
Distance Restraint Statistics	
Number of NOEs	618
Short Range ($\|i - j\| \leq 1$)	336
Medium Range ($1 < \|i - j\| < 5$)	88
Long Range ($\|i - j\| \geq 5$)	194
MolProbity Ramachandran Plot Statistics (%)	
Residues in most favored regions	96.3
Residues in allowed regions	3.7
Residues in disallowed regions	0.0
Average RMSD to mean (Å)	
Backbone (mean ± 1 S.D.)	0.29 ± 0.06
Heavy Atom (mean ± 1 S.D.)	0.79 ± 0.09
Structure Quality Factors (Raw / Z-score [b])	
MolProbity clash score	16.85 / -1.37
Procheck G-factor (phi & psi)	-0.26 / -0.71
Procheck G-factor (all)	-0.41 / -2.42
Verify 3D	0.42 / -0.64
ProsaII (negative)	0.56 / -0.37

[a] Analyzed for the 10 lowest energy structures for each designed protein using CYANA[135] and MolProbity[201]

[b] With respect to mean and standard deviation for a set of 252 X-ray structures with sequence lengths ≤ 500, resolution ≤ 1.80 Å, and R-free ≤ 0.28

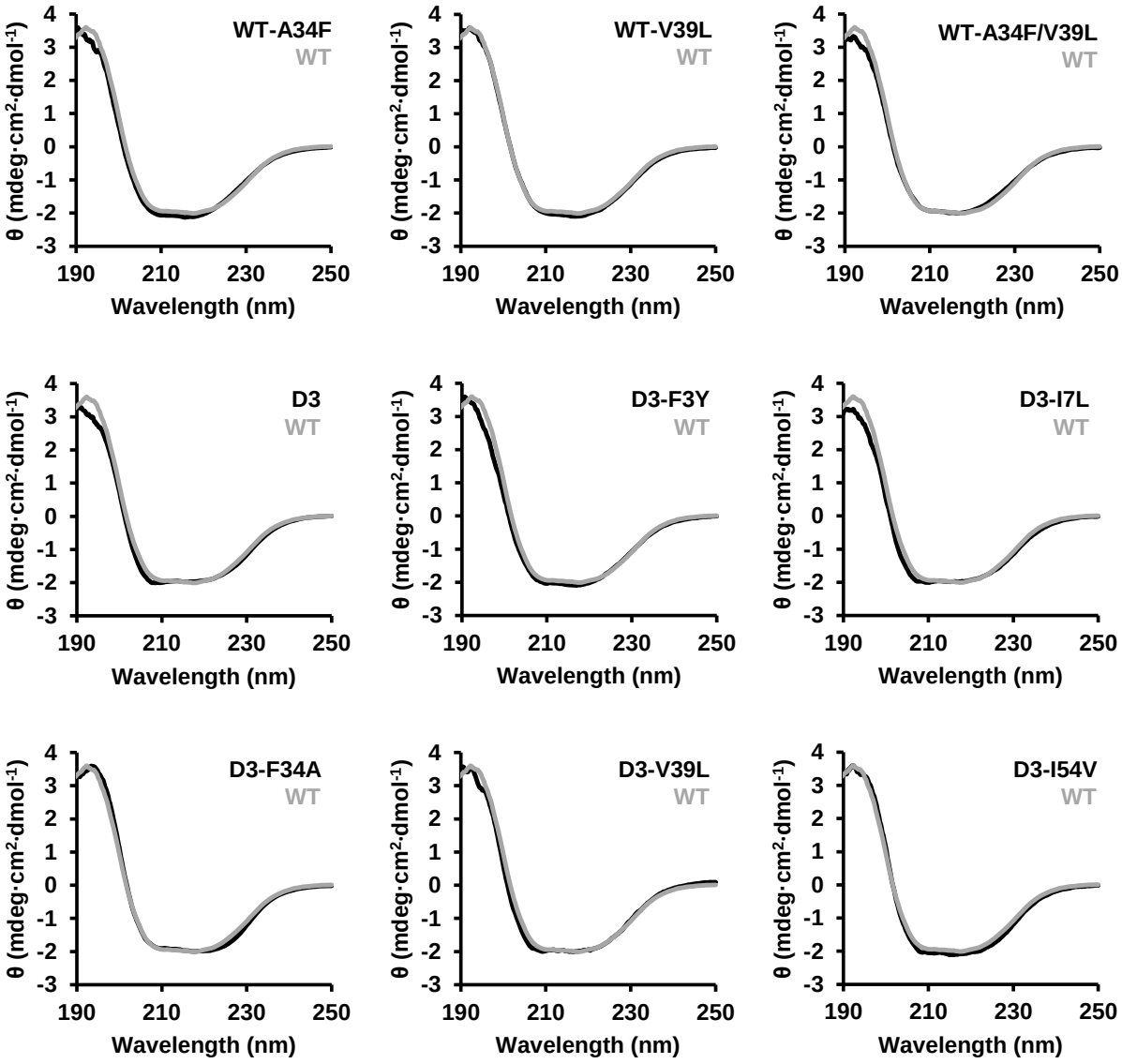


Figure S3.1. Gβ1 variants adopt a similar fold. Secondary structure content monitored by circular dichroism spectroscopy indicates that all Gβ1 variants possess similar secondary structure to the wild-type, suggesting a similar fold. Data shown as the average of 3 accumulations with a step size of 0.2 nm. Data from the wild-type Gβ1 is shown in grey for comparison.

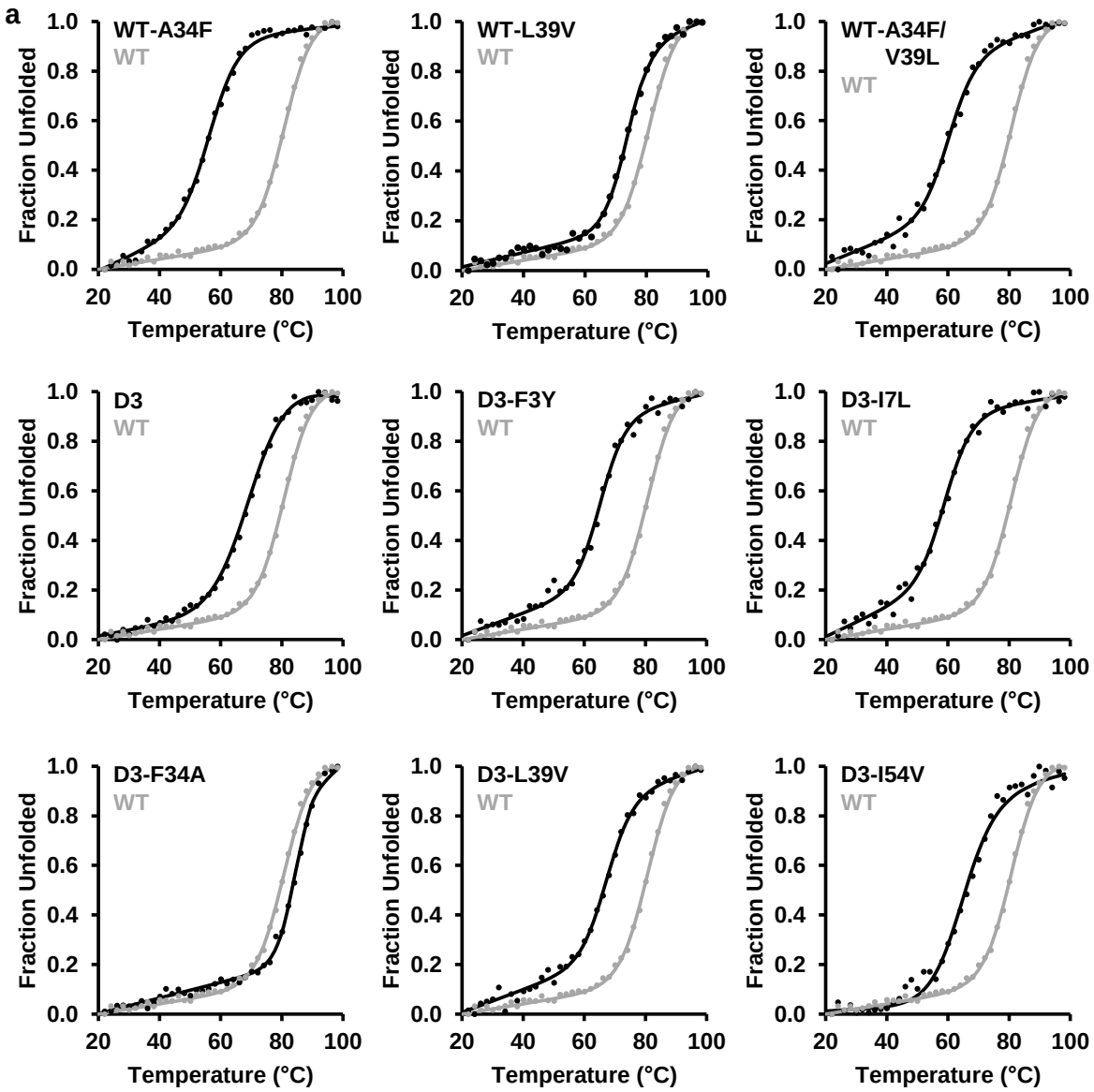


Figure S3.2. Thermal denaturation curves of Gβ1 variants. Thermal denaturation monitored by circular dichroism spectroscopy at 208 nm indicates that all Gβ1 variants are stable at room temperature (n = 1). Data was fit to a two-state unfolding model (lines).[198] Data from the thermal denaturation of wild-type Gβ1 is shown in grey for comparison.

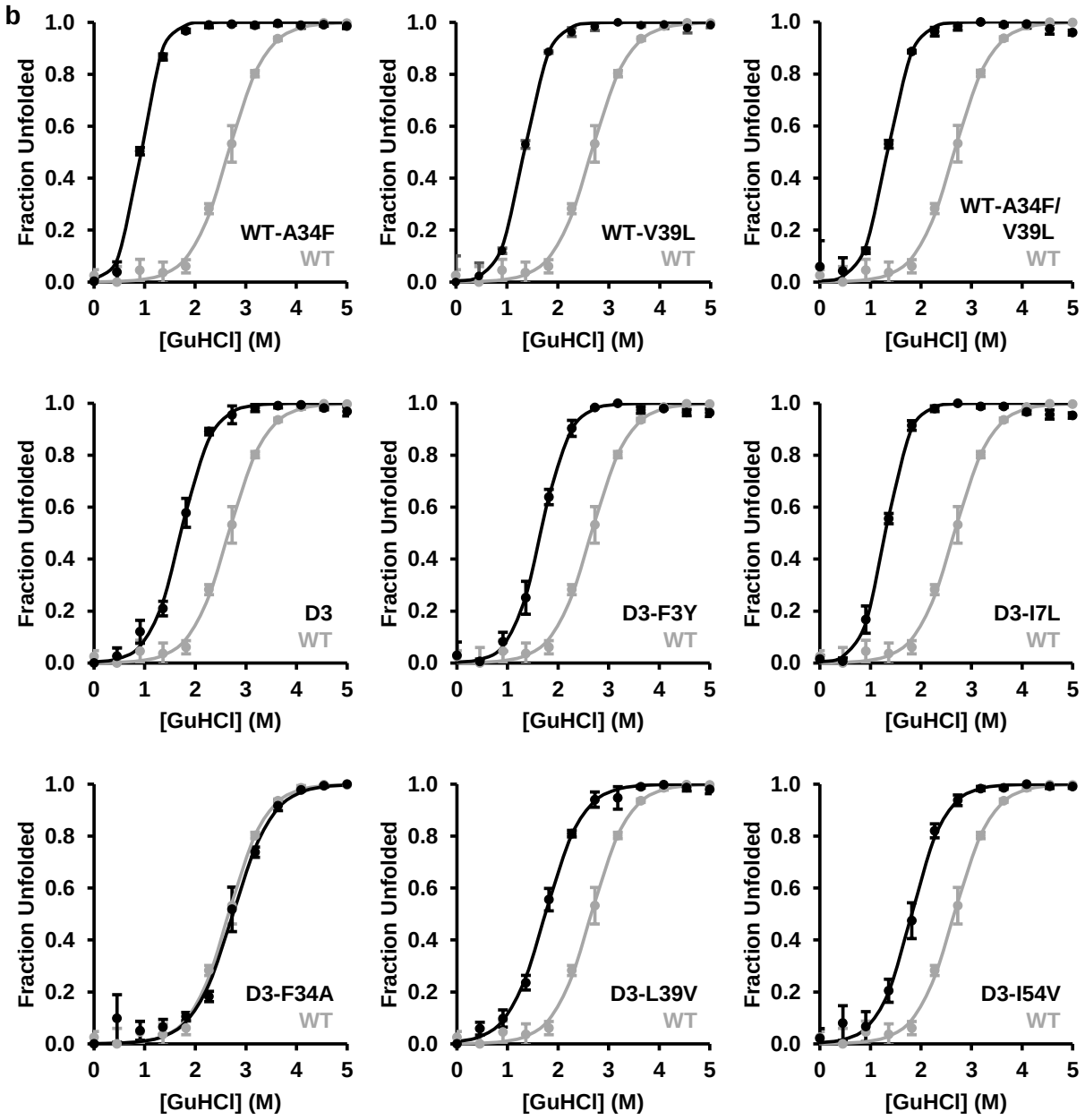


Figure S3.3. Chemical denaturation curves of Gβ1 variants. Chemical denaturation using guanidinium chloride (GuHCl) demonstrates that all proteins unfold according to a two-state model (n = 3). Fraction unfolded values were calculated by monitoring integrated tryptophan fluorescence ($\lambda_{excitation}$ = 280 nm, $\lambda_{emission}$ = 310–450 nm) relative to fluorescence at 0 M GuHCl. All experiments were performed in triplicate, and the average and standard deviation reported for each data point. The wild-type denaturation curve is shown in grey for comparison.

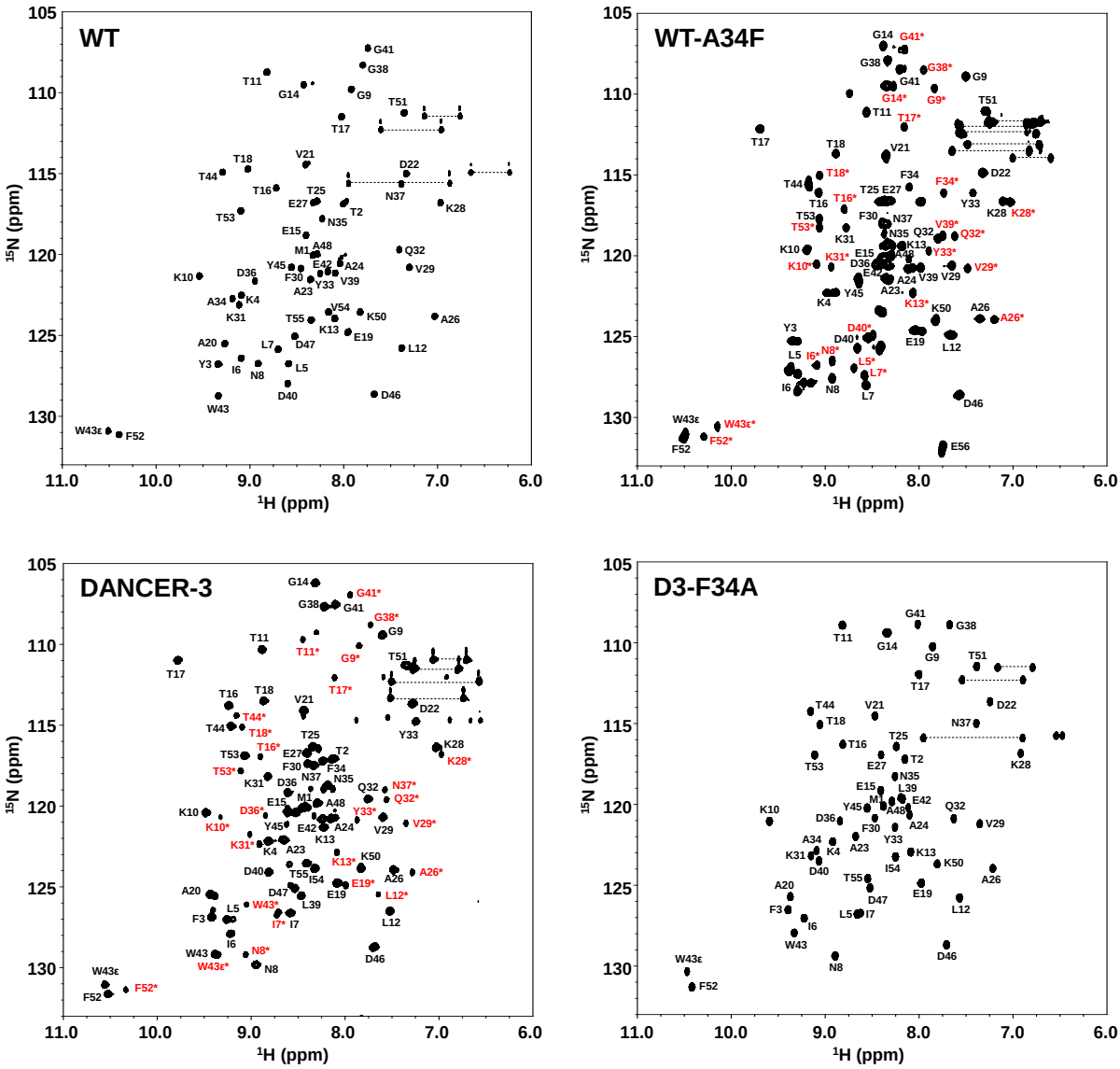


Figure S3.4. Assigned ^{1}H-^{15}N HSQC spectra of selected Gβ1 variants. Peaks corresponding to the minor state are indicated with red assignments. Side-chain amide resonances from asparagine and glutamine residues are connected by horizontal lines. WT-A34F assignments from Jee J., *et al*. (2008).

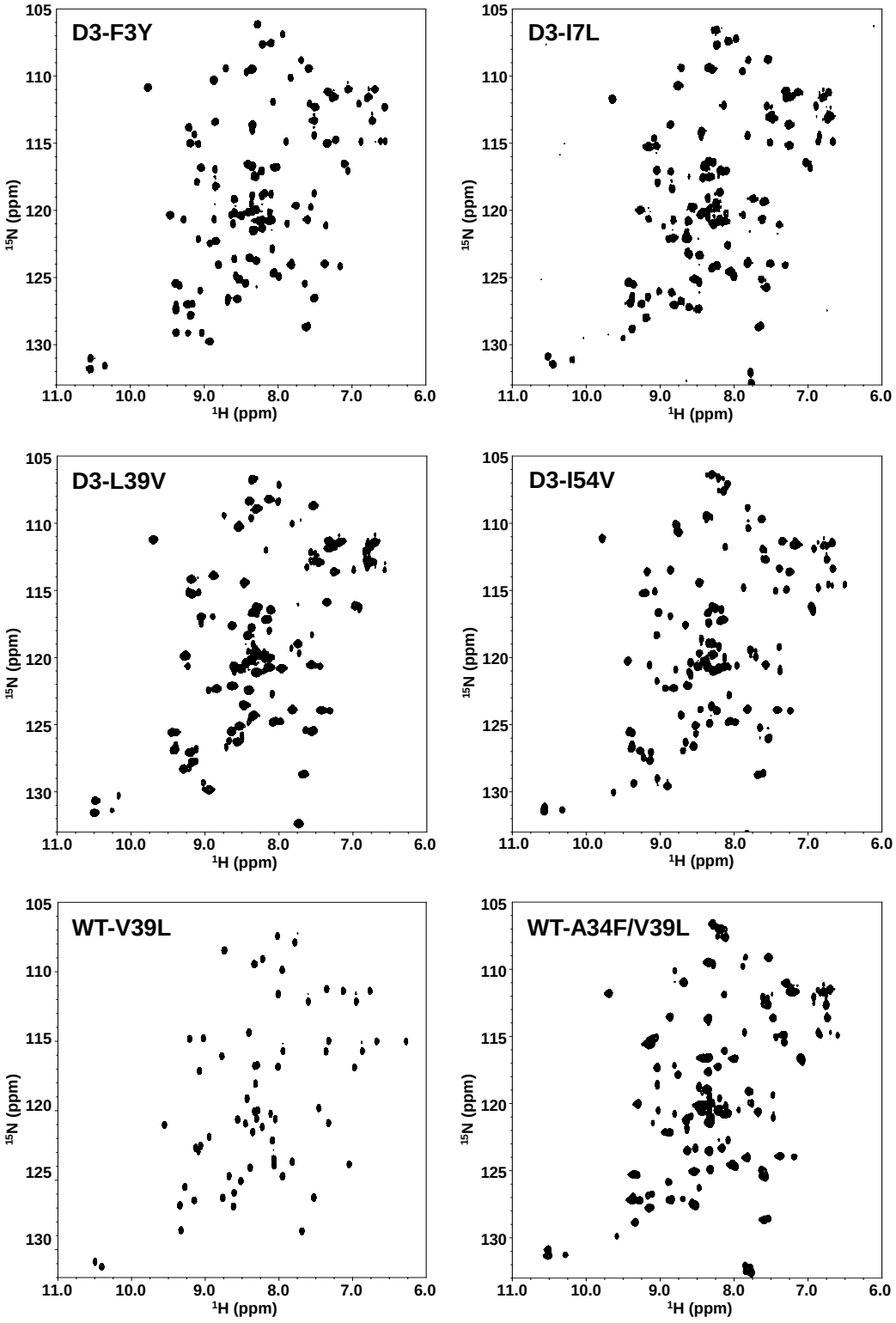


Figure S3.5. Unassigned ^{1}H-^{15}N HSQC spectra of selected Gβ1 variants.

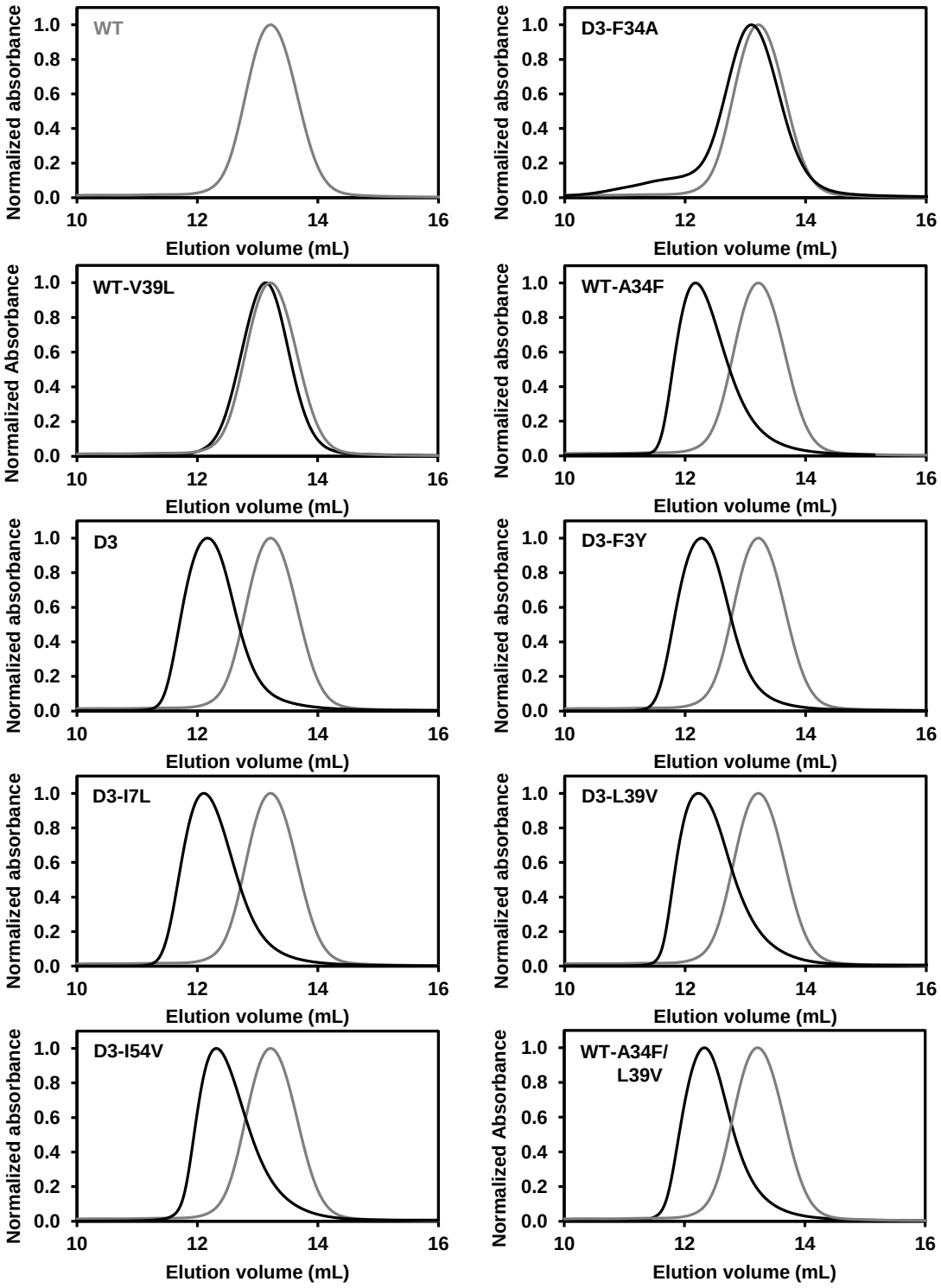


Figure S3.6. Size-exclusion chromatography profiles of Gβ1 variants. Size-exclusion chromatography elution profiles monitored by absorbance at 280 nm during the final purification step with an AKTA Superdex 75 10/300 column show a shift of ~1 mL in elution volume for dimeric Gβ1 variants relative to the monomeric wild-type, shown in grey for comparison, unlike what had previously been observed using a BioRad Enrich SEC650 column.[213]

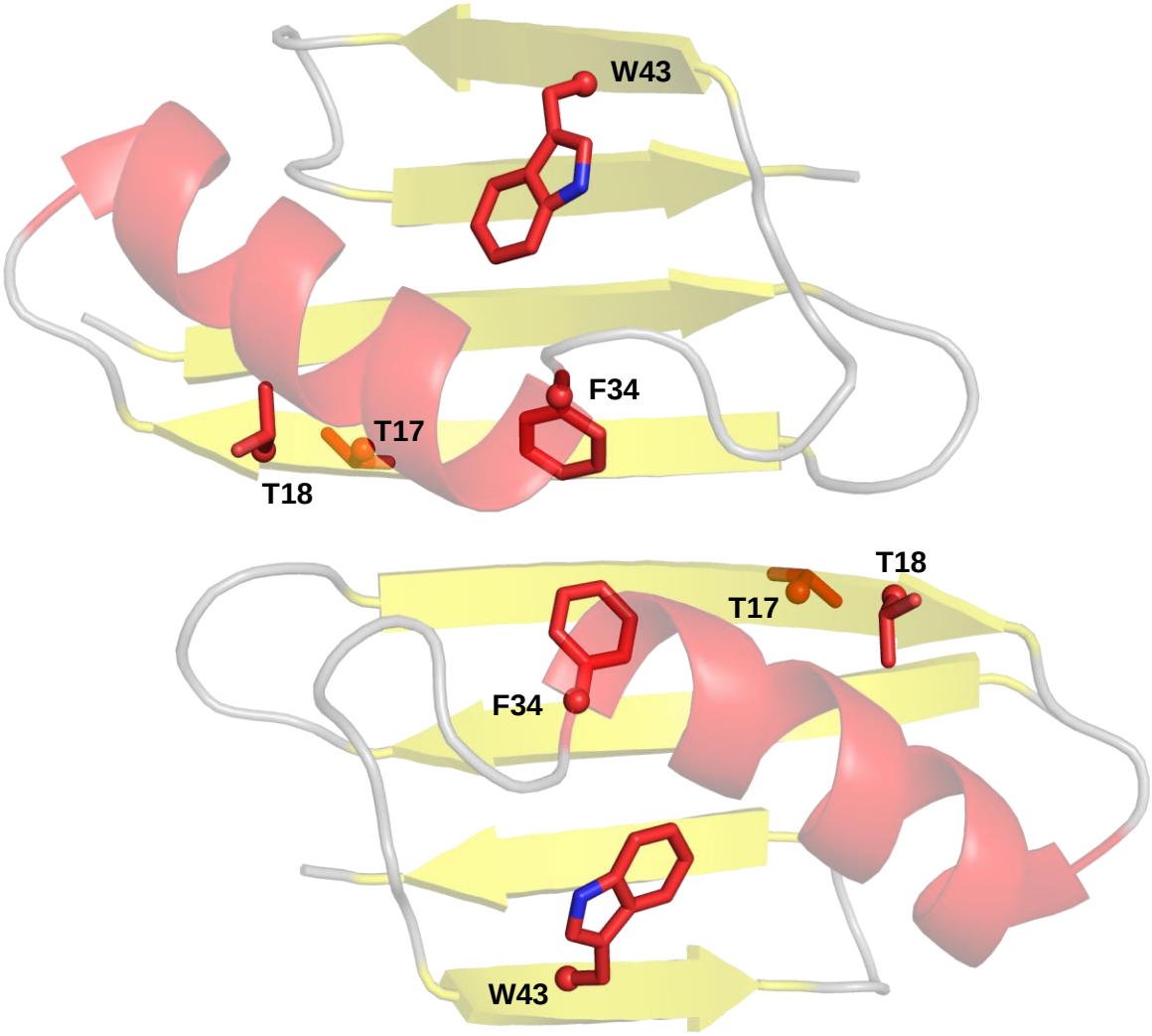


Figure S3.7. Structure of the WT-A34F dimer. Crystal structure of WT-A34F (PDB ID: 2RMM)[56] with selected residues highlighted as sticks. Phe34 is responsible for inducing dimerization, and Thr17 acts as a reporter of dimerization in NMR.

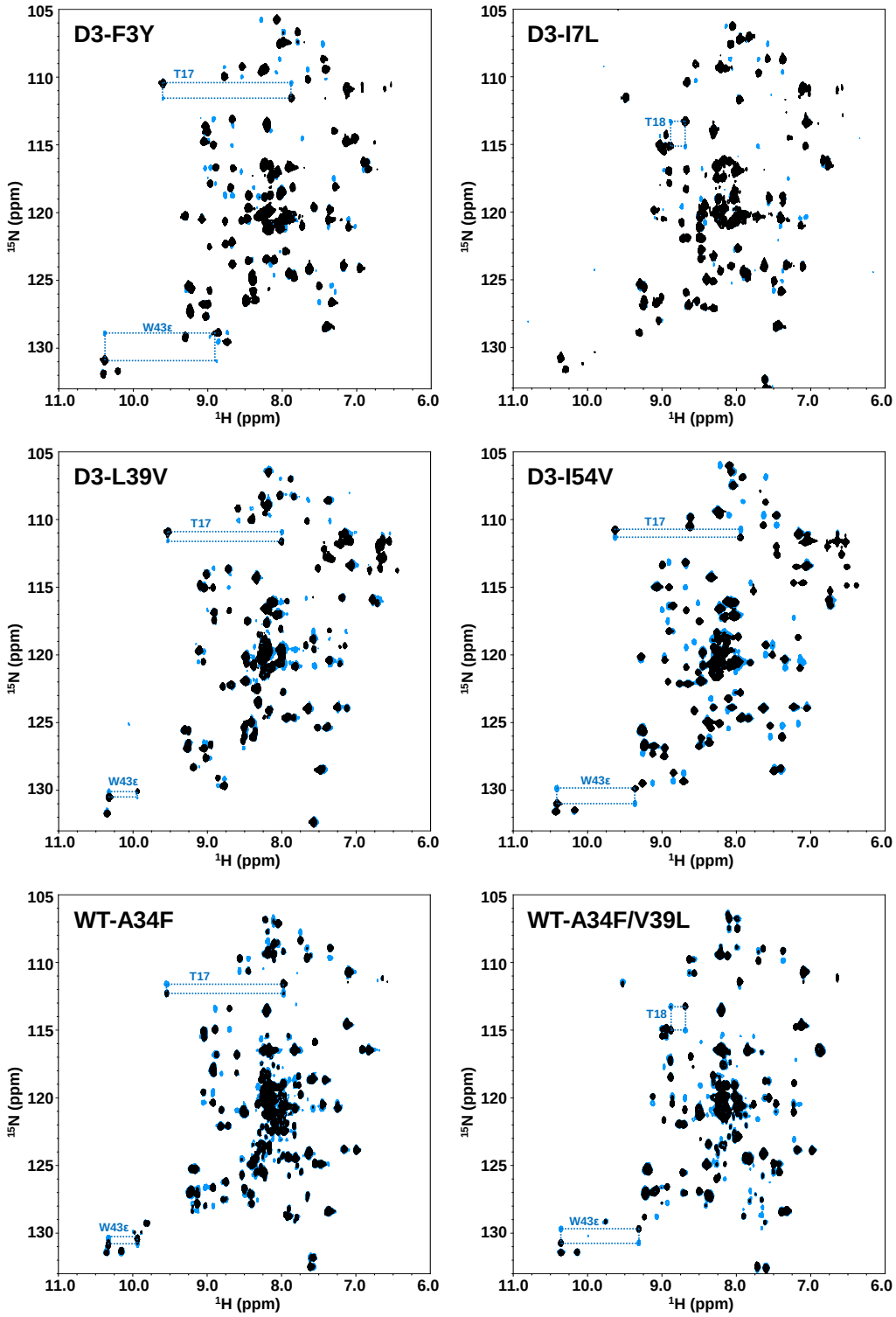
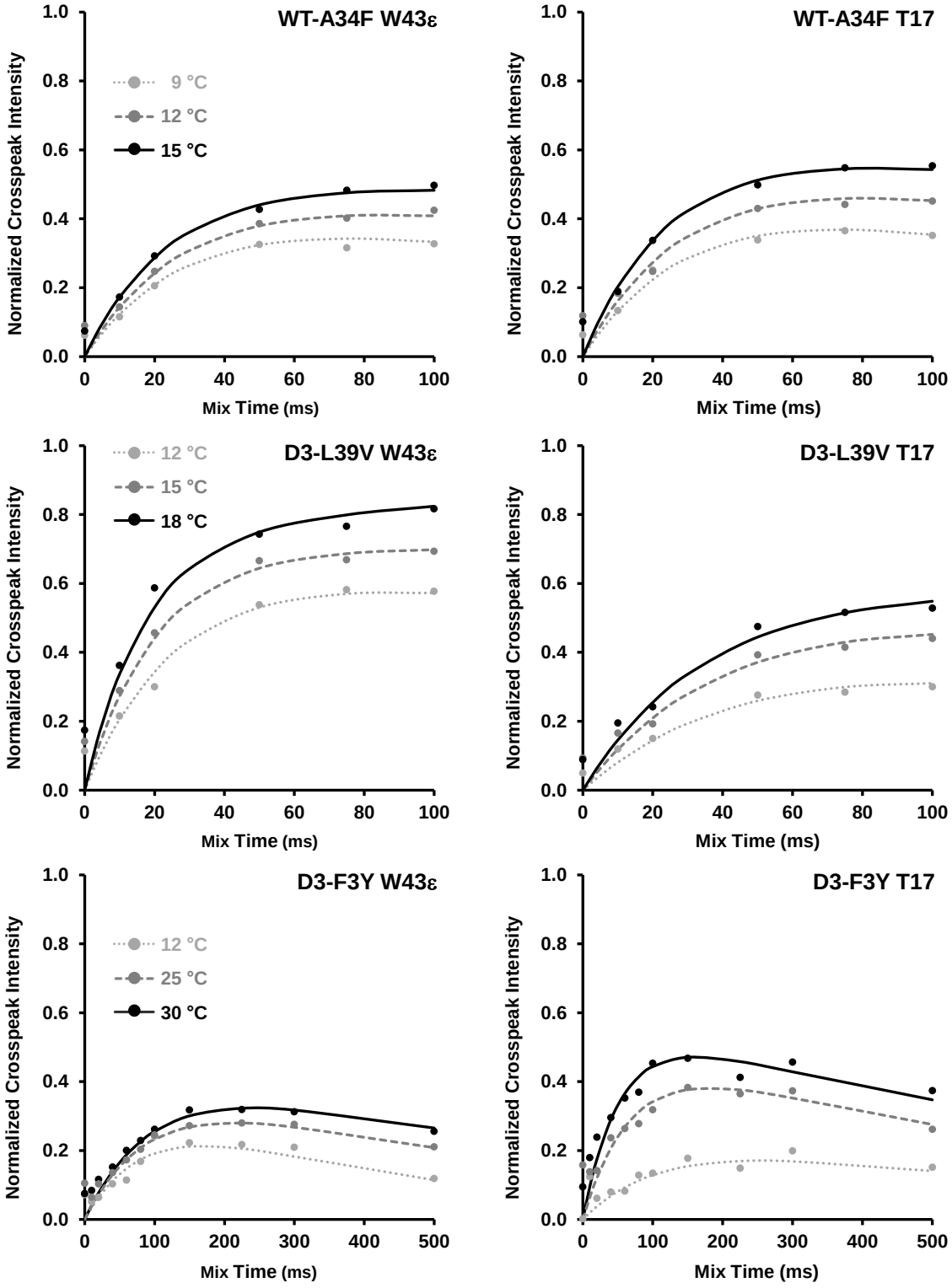


Figure S3.8. ^{1}H-^{15}N HSQC ZZ-Exchange spectra of selected Gβ1 variants. Representative ZZ-exchange spectra are shown in blue and overlaid with 1H-15N HSQCs in black to highlight the presence of exchange peaks. Dotted lines connect exchange correlations with peaks from major and minor used in the analysis of exchange kinetics for each respective variant.


WT-A34F W43ε
9 °C
12 °C
15 °C
WT-A34F T17
D3-L39V W43ε
12 °C
15 °C
18 °C
D3-L39V T17
D3-F3Y W43ε
12 °C
25 °C
30 °C
D3-F3Y T17
Normalized Crosspeak Intensity
Mix Time (ms)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
100
200
300
400
500

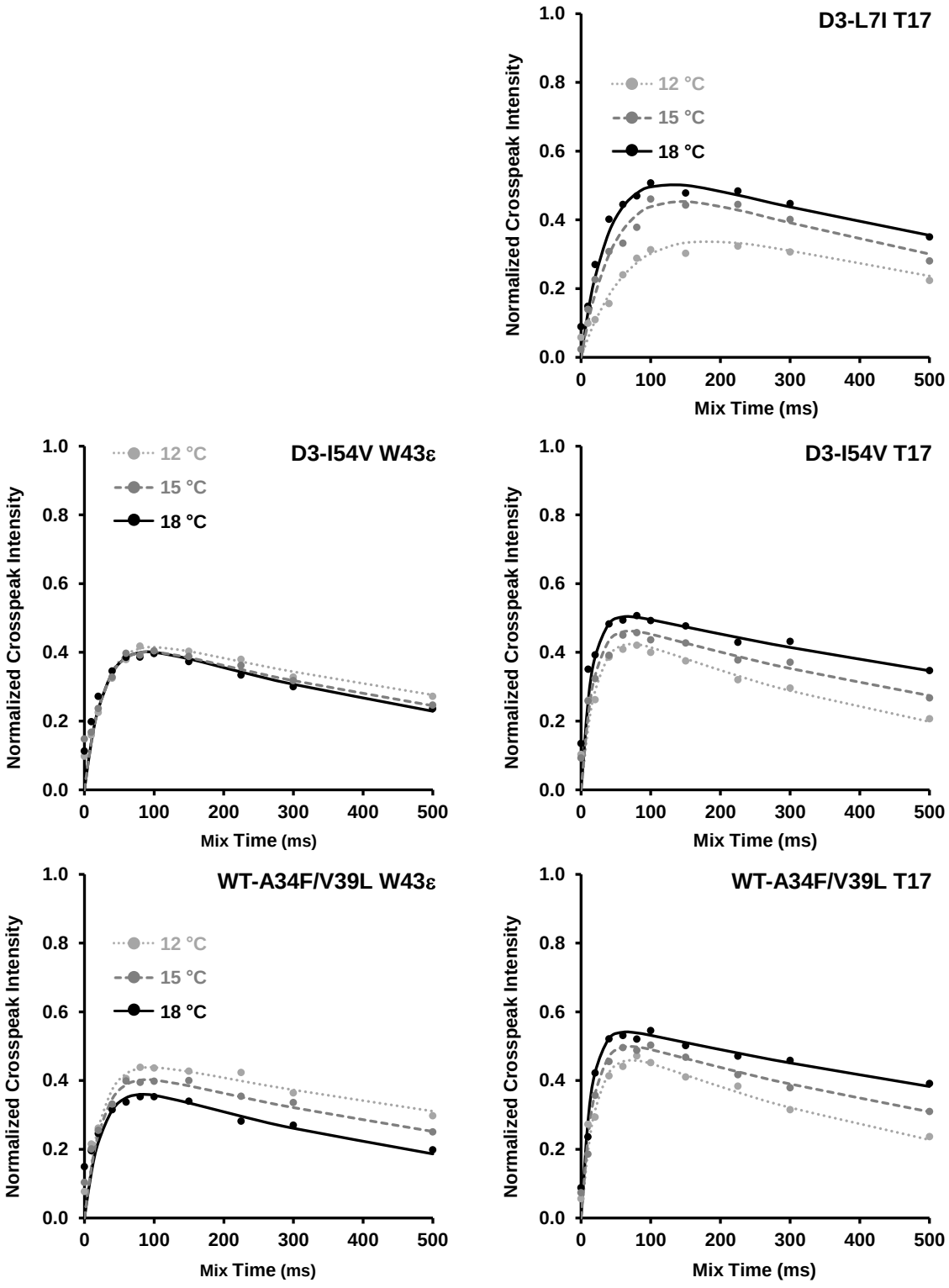


Figure S3.9. ^{1}H-^{15}N HSQC ZZ-Exchange exchange rates as a function of temperature. Crosspeak intensity curves for the Trp43ε minor to major state transition are shown for three temperatures per variant, fit using a four-term exchange and relaxation model.[137]

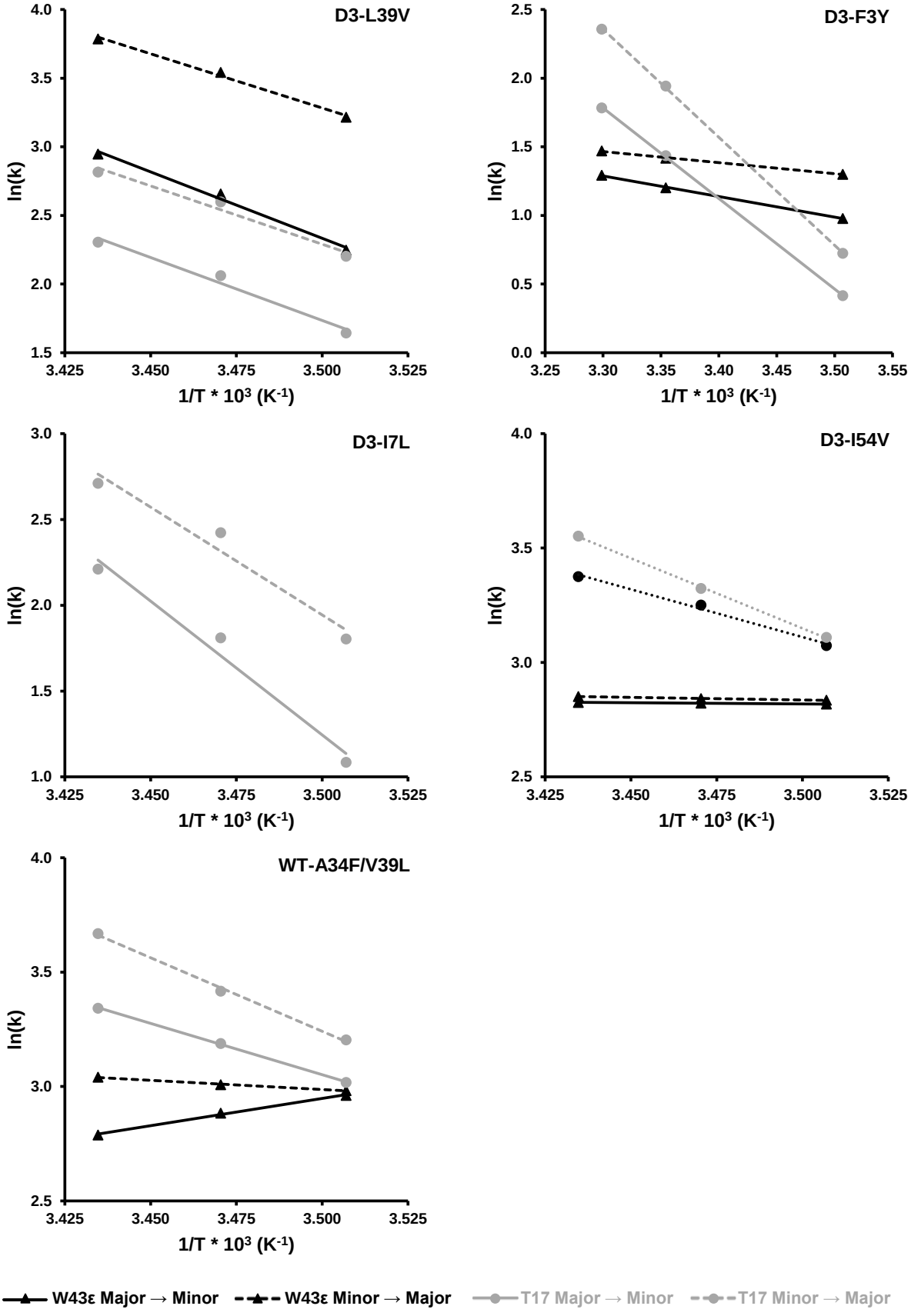

D3-L39V
D3-F3Y
D3-I7L
D3-I54V
WT-A34F/V39L
ln(k)
1/T * 10³ (K⁻¹)
W43ε Major → Minor
W43ε Minor → Major
T17 Major → Minor
T17 Minor → Major

Figure S3.10. Arrhenius plots of selected Gβ1 variants. Arrhenius plots demonstrate Arrhenius behavior for both Trp43ε and Thr17 exchange profiles in D3-L39V, as shown by similar activation energies for all four transitions. D3-F3Y, D3-I54V, and WT-A34F/V39L however demonstrate non-Arrhenius behavior where one or both Trp43ε transitions display a non-physical negative or very weak activation energy indicative that they do not follow a simple two-state Arrhenius model, as opposed to the normal Arrhenius behavior demonstrated by Thr17. D3-I7L W43ε minor state peaks and crosspeaks could not be accurately quantified due to low intensity as opposed to Thr17 (minor state peaks only due to overlap in ^{15}N for crosspeaks) and Thr18 peaks. For all proteins save D3-L39V, the trend observed for Trp43ε transitions is significantly different than that observed for Thr17, suggesting that different modes of exchange are experienced at each position.

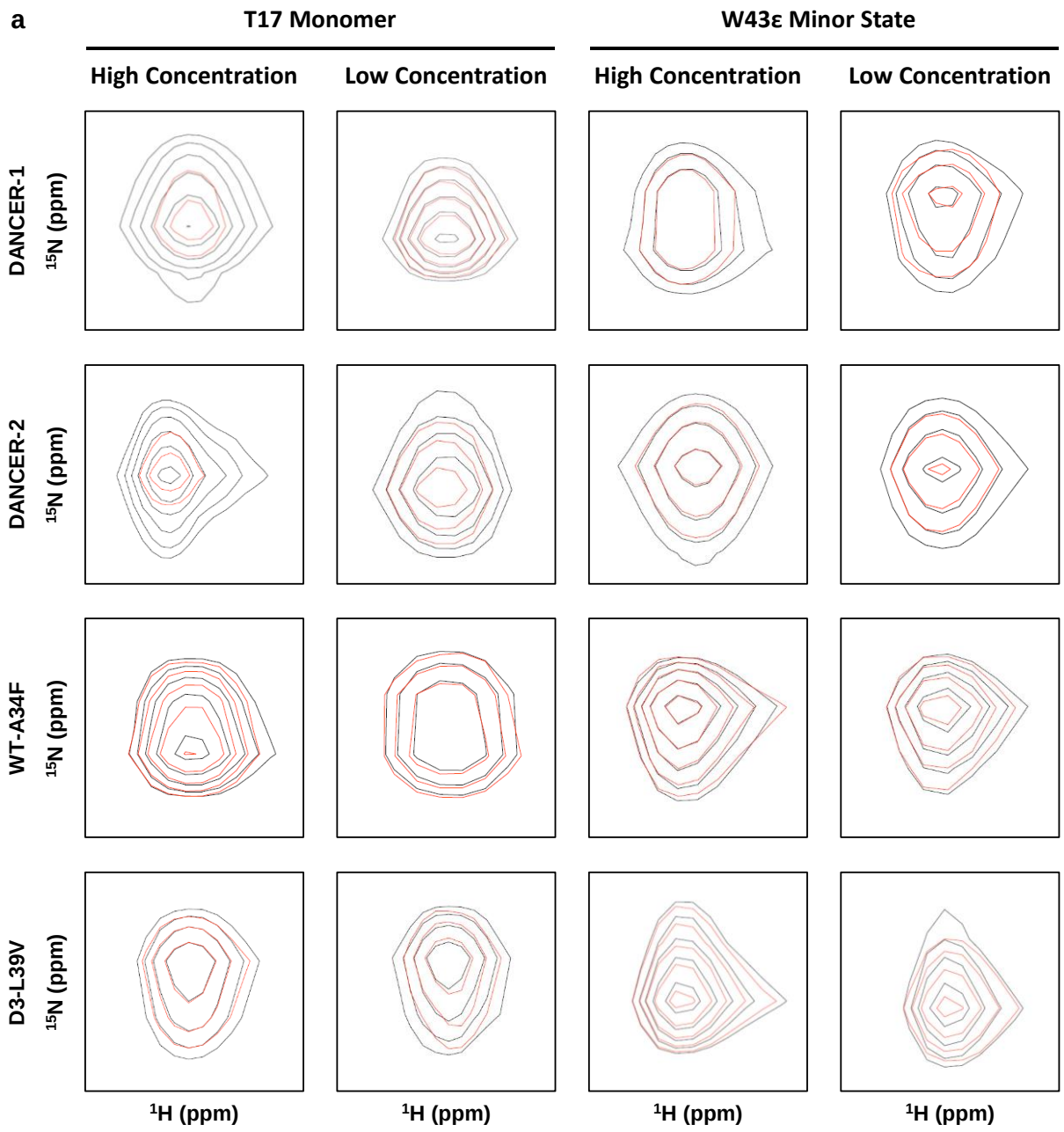


Figure S3.11. ^{1}H-^{15}N HSQC CPMG analysis of Thr17 and Trp43ε dynamics of selected Gβ1 variants. a, Cutouts from CPMG spectra showing peaks for the Thr17 monomeric state and the W43ε minor state. 20 Hz CPMG field spectra are overlaid in red over 1000 Hz CPMG field spectra in black and shown for both high and low protein concentrations. High concentrations are chosen on a protein by protein basis to obtain a well-resolved T17 monomer peak, and low concentrations are at a 1:5 dilution of the high concentration.

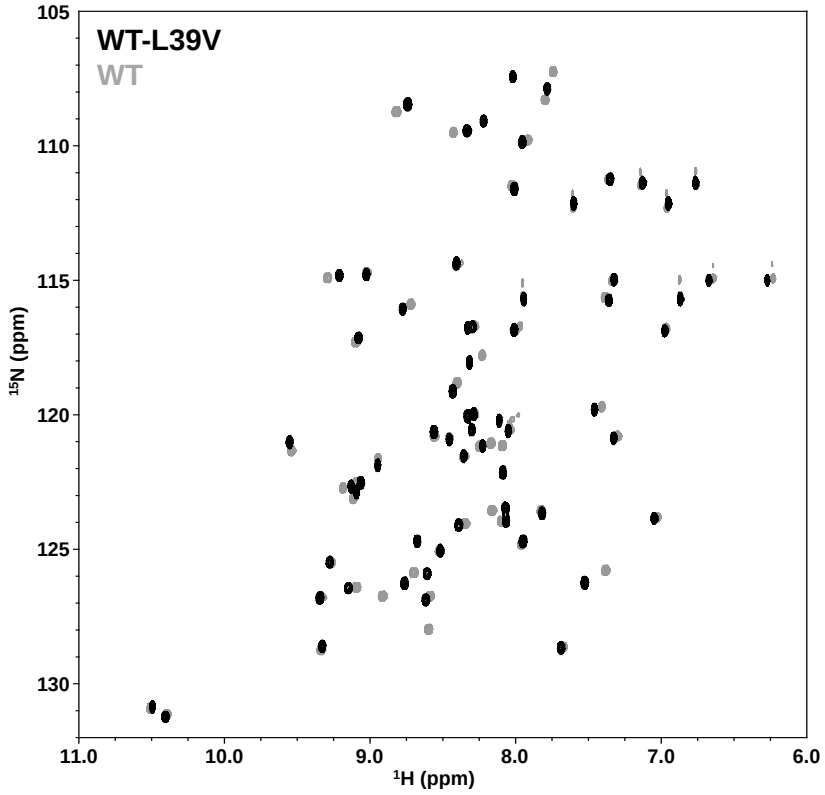


Figure S3.12. Overlay of WT-L39V and wild-type ^{1}H-^{15}N HSQCs. An HSQC spectrum of WT-L39V shown in black is overlaid on a wild-type spectrum in grey, demonstrating spectral similarity that suggests a similar conformation for both proteins.

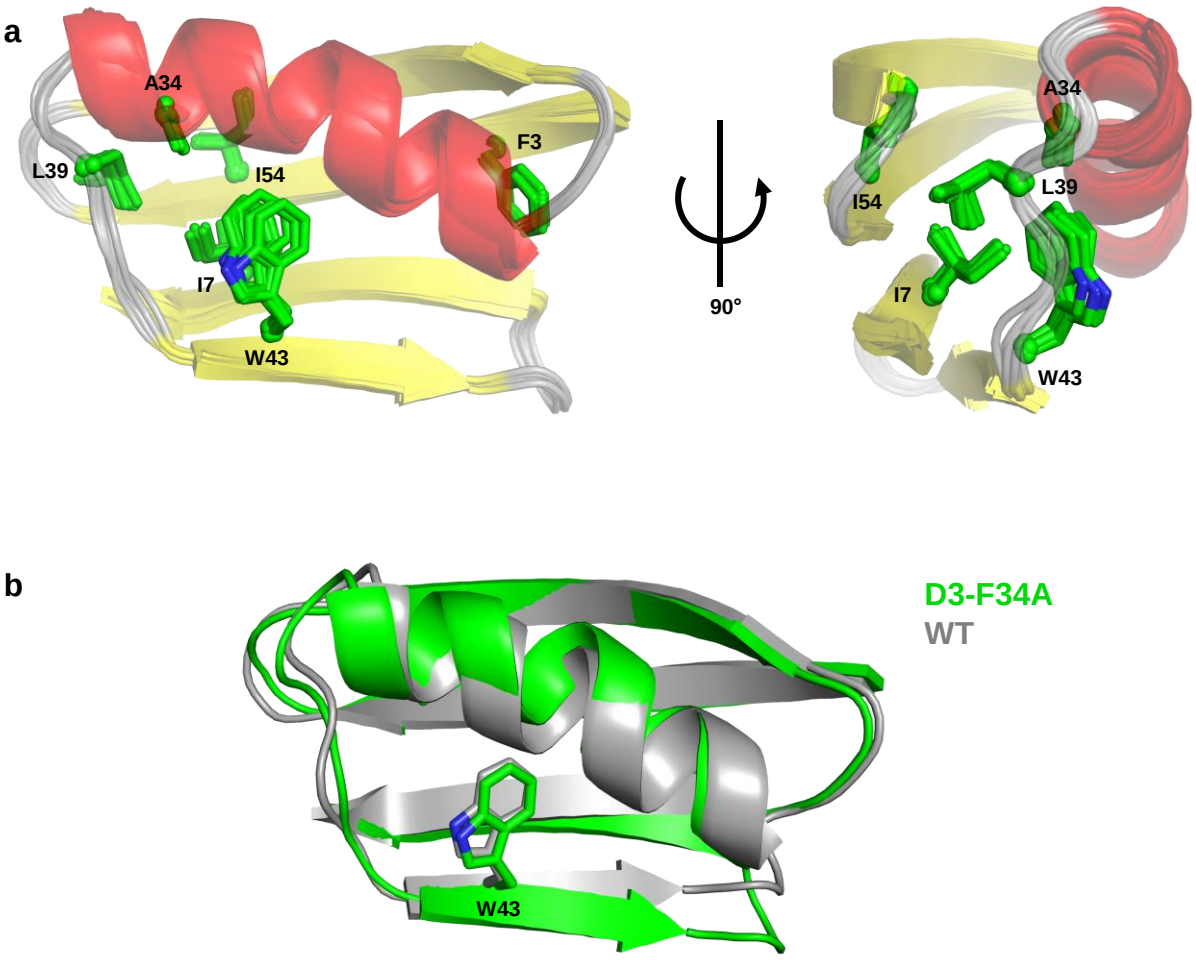


Figure S3.13. NMR solution structure of D3-F34A. a, An NMR ensemble of 10 conformers (lowest energy) is shown for D3-F34A, with mutations from the wild-type, Trp43, and Ala34 shown as green sticks. Secondary structural elements, as assigned by NMR chemical shift indices,[134] are colored in red, yellow, and grey for α-helices, β-sheets, and loops respectively. **b,** An overlay of the D3-F34A cartoon representation is shown in green overlaid over the wild-type crystal structure (PDB ID: 1PGA)[53] in grey. The Trp43 side-chain of each is shown as sticks and adopt nearly identical conformations.

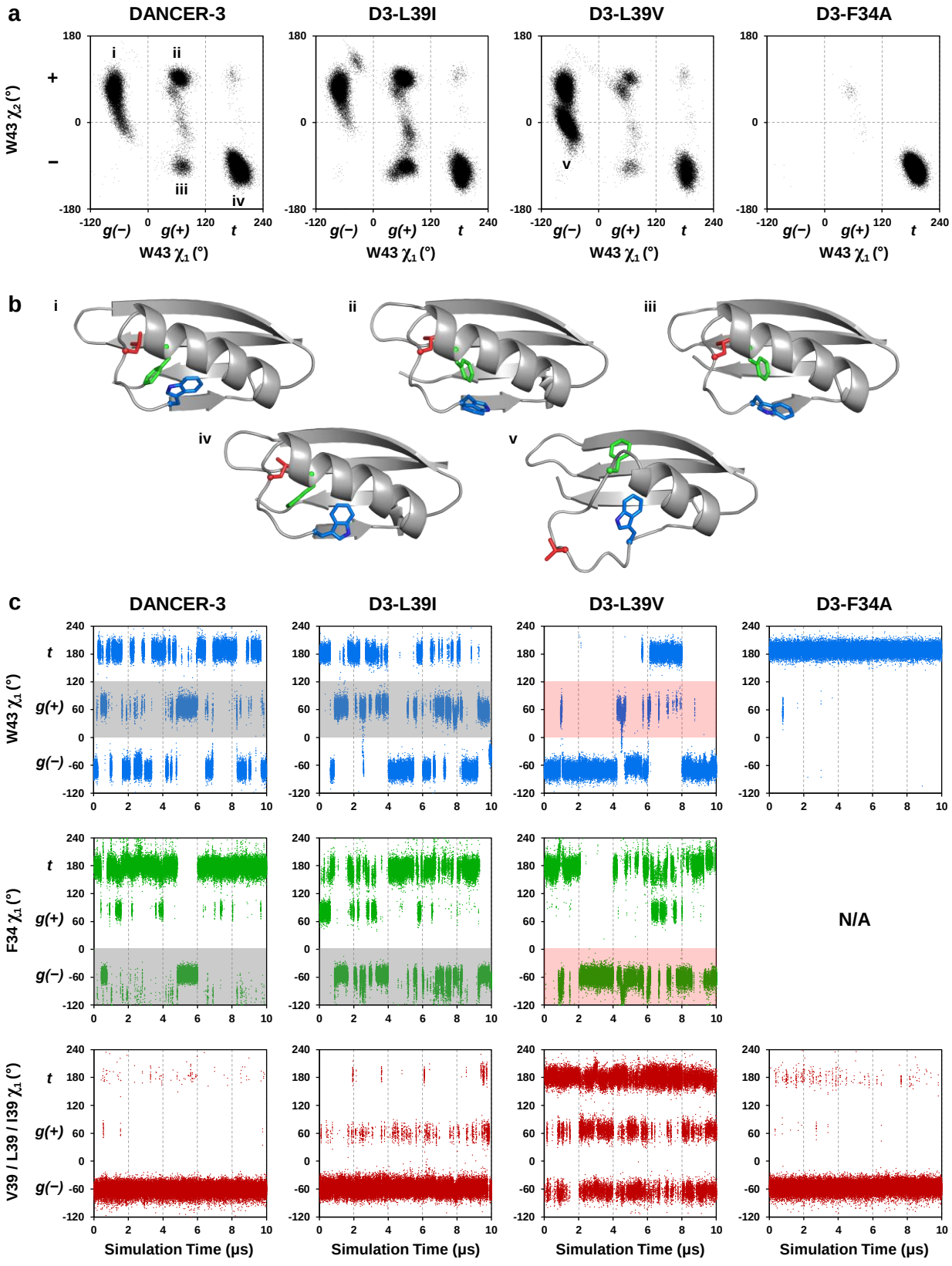

a
DANCER-3
D3-L39I
D3-L39V
D3-F34A
W43 χ2 (°)
W43 χ1 (°)
g(−)
g(+)
t
i
ii
iii
iv
v
b
c
F34 χ1 (°)
V39 / L39 / I39 χ1 (°)
Simulation Time (μs)
N/A

Figure S3.14. Conformational plots from molecular dynamics simulations of selected Gβ1 variants. a, Dihedral angle plots for Trp43 χ_1 and χ_2 dihedral angles over the course of 10 μs of simulation time, sampled every 200 ps. **b,** Representative structures showing conformations of Trp43 (in blue), Phe34 (in green), and Leu39 or Val39 (in red) as well as backbone conformations in each major conformational state adopted over the course of the simulations, identified in (a) using Roman numerals. **c,** χ_1 dihedral angles vs time plots for Trp43, Phe34, and Leu39 or Val39 demonstrate concerted motions between Trp43 and Phe34 in DANCER-3 and D3-L39I, as shown by the simultaneous adoption of a gauche$^-$ conformation in Phe34 when Trp43 adopts a gauche$^+$ conformation (shown as grey shaded regions). These motions become uncoupled in D3-L39V (shown as red shaded regions). No concerted motions to either Trp43 or Phe34 are observed at position 39.

Chapter 4: Brighter red fluorescent proteins display reduced structural dynamics

Adam M. Damry, Serena E. Hunt, Natalie K. Goto, Roberto A. Chica

4.1 Statement of Contribution

Serena E. Hunt and Adam M. Damry performed biophysical characterization experiments. Serena E. Hunt assigned chemical shifts for mPlum-E16P and mRouge. All other NMR experiments were analyzed by Adam M. Damry. Experimental design was devised by Adam M. Damry, Dr. Natalie K. Goto, and Dr. Roberto A. Chica. The chapter was written by Adam M. Damry and edited by Dr. Roberto A. Chica.

4.2 Introduction

Red fluorescent proteins (RFPs) are genetically encodable fluorophores that have found widespread use throughout biological imaging and research fields. They have been used for as varied applications as protein tracking within cells and tissues,[71] as partners with green fluorescent proteins (GFPs) in FRET pairing,[72,73] as signal transducers in biosensors,[74-76] and more. Recently, RFPs have gained popularity due to the advantages that the longer wavelength light they emit presents, notably an increased tissue penetration and a reduced phototoxicity that makes them particularly applicable to deep tissue and whole animal imaging.[62,216] Despite these advantages however, RFPs as a family tend to be dimmer than their green equivalents, and dimmer still as their emission wavelength increases.[62] To counter this disadvantage, RFPs have been extensively engineered through both rational and semi-rational design approaches[67,82,217,218] and directed evolution[83,219] to increase their brightness, though brightness-increasing mutations near the RFP chromophore often come paired with unwanted hypsochromic shifts that result from a perturbation of the chromophore's electronic environment.[220-222] The rational design of second shell residues

and beyond however provides the opportunity to improve RFP brightness without unwanted changes to other spectral properties.

The opportunity for such a design comes from the link between RFP brightness and their structural dynamics. Though a combination of several factors such as maturation and photostability contribute to bulk signal in an RFP sample, RFP brightness is primarily a function of two properties; the chromophore's molar extinction coefficient (ε) representing its ability to absorb incident light, and its quantum yield (φ) representing its probability to re-emit light following excitation.[59] Quantum yield in particular has been linked to dynamics wherein chromophore rigidification has led to reduced non-radiative decay of the chromophore excited state, and by extension to increased quantum yield.[88,223] If we could therefore redesign the structural dynamics in an RFP to favor a rigid chromophore without affecting the chromophore electrostatic environment, we could improve these already useful biosensors. However, though we previously reported a method to computationally design protein dynamics,[213] altering function through the computational design of distal protein dynamics has yet to be achieved. RFPs nonetheless present an excellent system in which to refine our methods for the design of dynamics and extend them to protein function, given the direct link between RFP dynamics and an easily screened function that is their fluorescence. Using current CPD approaches and limited computational resources, it is however not tractable to query the entire protein sequence to locate sites where designed dynamics might trickle down to affect chromophore dynamics. Moreover, predicting which sites would need to be designed to alter chromophore dynamics through intuition and visual approximation is difficult. Therefore, prior to designing dynamics in RFPs, we will need to determine what regions of these RFPs to target *in silico*. Herein, we have addressed this challenge by assessing and comparing structural dynamics throughout a set of related RFPs of various brightness using nuclear

magnetic resonance (NMR) spectroscopy. Through clustering of positions in RFPs where structural dynamics were shown to be correlated to chromophore brightness, we outline regions of the RFP barrel where future designs for increased brightness, both through the rational design of dynamics and directed evolution, should be focused.

4.3 Results

4.3.1 Chemical shift assignment of mCherry family RFPs

One of the most used and studied RFPs is mCherry, a bright monomeric RFP with desirable properties such as rapid maturation and high photostability.[66] Extensive engineering efforts have yielded mCherry variants with quantum yields ranging from 0.70 to 0.02, making this family of RFPs ideal for studying the link between FP structural dynamics and their brightness. For this analysis, we selected four RFPs from the mCherry family with a broad range of quantum yields (Table 4.1); mCherry itself (QY = 0.22),[66] along with mCherry variants mScarlet (QY = 0.7),[83] mPlum-E16P (QY = 0.14),[67] and mRojoA (QY = 0.02).[193] These RFPs each contain between 6 and 30 point mutations relative to the parent mCherry, but all possess the native mCherry chromophore-forming MYG tripeptide (Fig. 4.1, S4.1). With the goal of using solution NMR to assess the dynamic properties of these RFPs on a per residue basis, we first assigned their ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra (Fig 4.2, S4.2). While roughly 90 % of assignable residues in mCherry, mScarlet, and mPlum-E16P could be unambiguously identified, only 69 % of mRojoA assignable residues could be unambiguously assigned due to a poorer spectral quality and a poor room temperature stability stemming from chromophore hydrolysis that limited experiment run time. To improve our dataset, we chose to also include mRouge (QY = 0.02),[193] another low QY RFP derived from mCherry and possessing the native MYG chromophore, for which 82% of assignable residues could be unambiguously identified. Together,

mRojoA and mRouge provide a coverage of 88 % of assignable residues in low QY RFPs, a threshold comparable to that of the medium and high QY RFPs studied. During the chemical shift assignment process, we also observed that spectra from mRojoA and mRouge, the two dimmest RFPs studied, presented several peaks with increased broadness compared to brighter RFPs (Fig. 4.2, S4.2, S4.3). These observations, as well as the greater number of unassignable residues due to missing correlations in both mRojoA and the less hydrolysis-prone mRouge, are in keeping with these dimmer RFPs being more dynamic than their brighter variants.

Table 4.1. RFP spectral properties.

Protein	λex (nm)	λem (nm)	ϕ	ε ($M^{-1}cm^{-1}$)	Brightness [a] ($mM^{-1}cm^{-1}$)	Reference
mScarlet	569	594	0.70	100,000	70.0	83
mCherry	587	610	0.22	72,000	15.8	66
mPlum-E16P	590	630	0.14	29,000	4.06	67
mRojoA	597	633	0.02	48,000	0.96	[193]
mRouge	600	637	0.02	43,000	0.86	[193]

[a] Relative brightness as defined as the product of φ and ε.

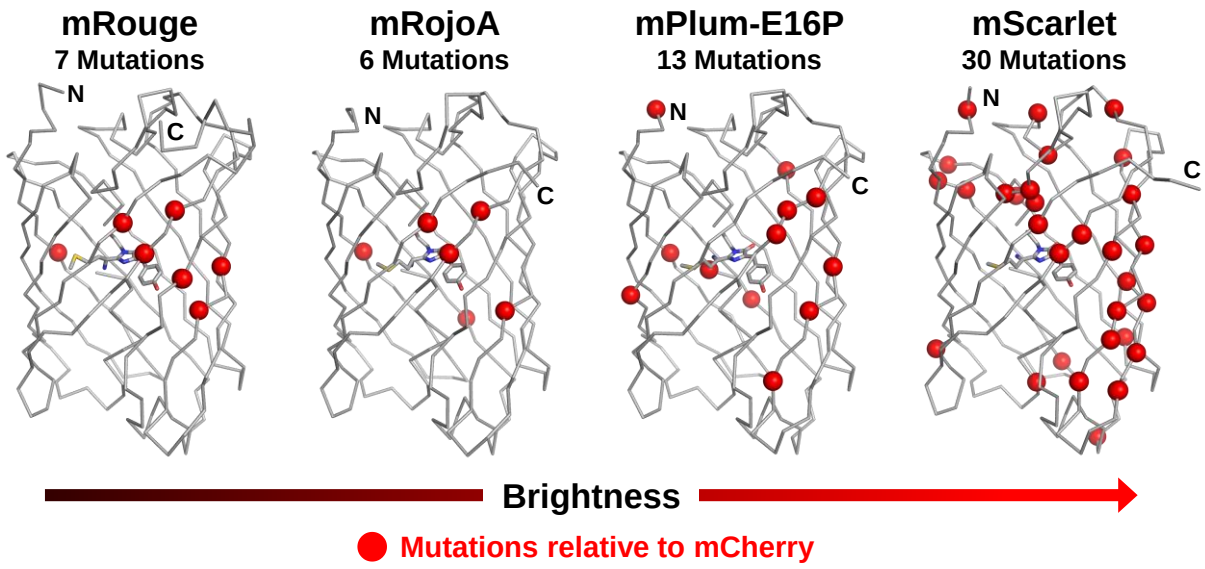


Figure 4.1. Mutated positions in selected mCherry-derived RFPs. Four RFPs were selected for characterization alongside mCherry, shown here in order of increasing brightness as a ribbon representation of their respective crystal structures (mRouge PDB ID: 3NED[193], mRojoA PDB ID: 3NEZ[193], mPlum-E16P PDB ID: 4H3L[67], mScarlet PDB ID: 5LK4[83]). Point mutations relative to the parent mCherry are shown as red spheres, and the chromophore is shown as sticks. mPlum-E16P and mScarlet also contain a four amino acid deletion near the protein N-terminus and mPlum-E16P contains a six amino acid deletion at the protein C-terminus that are not shown, nor included in the count of mutations.

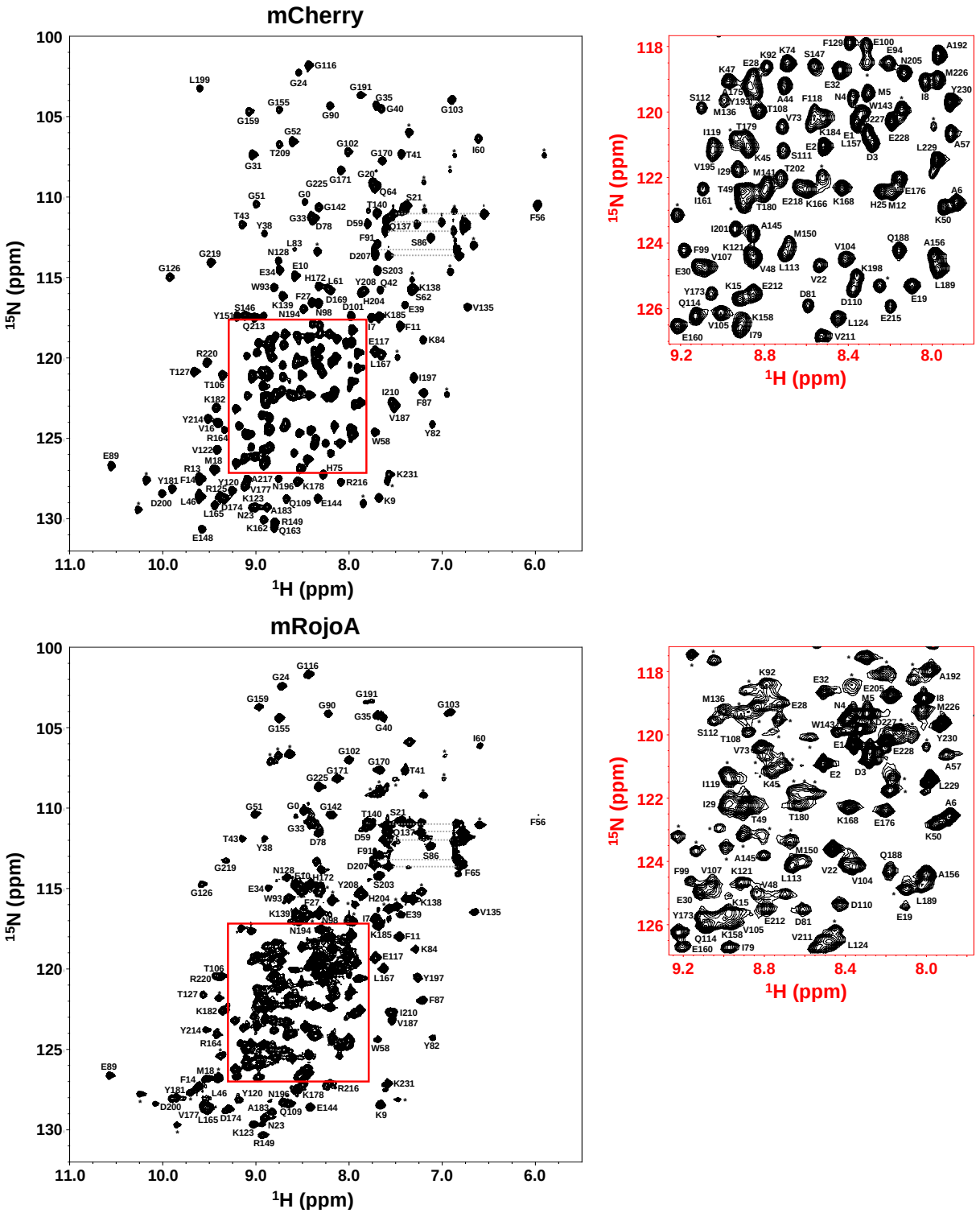


Figure 4.2. Representative assigned ^{1}H-^{15}N HSQC spectra for bright and dim RFPs. HSQC spectra for the bright mCherry and dim mRojoA RFPs are shown labeled with unambiguously assigned residues identified. Side-chain amide resonances from asparagine and glutamine residues are connected by horizontal lines, and unassigned or other side-chain resonances are denoted by an asterisk. The central portion of the spectrum (boxed in red) is shown separately for clarity.

4.3.2 Moderate timescale dynamics dominate correlations to RFP brightness

With all five RFP HSQCs having been assigned, we went on to characterize structural dynamics in the RFP barrel on a per-residue basis. Though RFP fluorescence is a process that occurs on the order of nanoseconds,[89] it is not known whether faster RFP dynamics on the order of fluorescence lifetime or slower dynamics leading to poorly fluorescent protein conformations dominate the relationship between RFP dynamics and brightness. We therefore used three different NMR analysis techniques to probe RFP dynamics on different timescales. CPMG relaxation dispersion examines dynamics on a long microsecond to short millisecond timescale that is characteristic of conformational rearrangements or domain motions in proteins. HSQC peak analysis extends this range to include dynamics on a shorter microsecond to long nanosecond timescale that is characteristic of smaller fluctuations in the global protein structure. Finally, T_1-T_2 relaxation examines dynamics on a picosecond to short nanosecond timescale that is characteristic of protein-solvent interactions and side chain fluctuations.[224-226]

We first performed CPMG relaxation dispersion measurements on mCherry and mRojoA as a preliminary comparison between a moderately bright and a very dim RFP. Neither protein however showed any relaxation dispersion, with almost all residues giving rise to well resolved peaks that showed little change with CPMG field strength (Fig. S4.4). This lack of response to CPMG relaxation-dispersion field strength, observed for most residues throughout both RFPs, suggest that the RFP barrel structure does not undergo significant dynamics on the microsecond to millisecond timescale and range of motion examined by this experiment. By extension, brightness is unlikely to be significantly dependent on dynamics on this timescale.

It is likely that conformational exchange on a slower timescale (e.g. seconds) is also not significant, since each residue in the HSQC could be uniquely assigned to a single residue in the

RFP sequence. Peak linewidths in HSQCs, however, can report on a wider range of timescales than is detected by CPMG experiments. Two different competing dynamic regimes can be observed in HSQC spectra as peak intensities reflect both the effect of local correlation times, which are tied to fast dynamics and result in an increase in peak intensity with increased dynamics, and chemical exchange processes, which are instead tied to slower dynamics and result in a decrease in peak intensity with increased dynamics. A peak with below average intensity thus likely arises from a region of the protein undergoing slower microsecond-timescale dynamics,[125] whereas a peak with above average intensity instead arises from a region of the protein undergoing faster nanosecond-timescale dynamics. Though these processes are the same that contribute to a CPMG signal in the case of chemical exchange, or to a quantifiable change in the local relaxation properties in the case of correlation times, HSQCs can be sensitive to motions that are faster than the optimal 0.1 – 10 ms range examined by CPMG and slower than the optimal < 10 ns range examined by T_1-T_2 relaxation,[124,227] which we have termed moderate-timescale dynamics. Thus, in an HSQC spectrum, increased dynamics can be represented as a deviation from the protein's average peak intensity, whether higher or lower. We calculated this deviation for each residue in each protein and searched for positive correlations between rigidity and RFP brightness (Fig. S4.5, S4.6, Methods), defined as positions where the absolute value of deviation was greatest in dim RFPs and lowest in bright RFPs, thus suggesting that bright RFPs are more rigid at these positions than dim RFPs. Twenty-five such positions, exhibiting moderate-timescale dynamics that were positively correlated to brightness, were identified.

When local dynamics approach or surpass the molecular tumbling rate in solution, which occurs on the scale of tens of nanoseconds for a protein of ~27 kDa, the T_1 / T_2 value begins to reflect local dynamics rather than bulk molecular tumbling. T_1 and T_2 relaxation rates thus allows

us to probe fast protein dynamics on the ns – ps timescale, as represented as a residue-by-residue apparent correlation time ($\tau_c^{app} \propto T_1 / T_2$).[142,143] Variations in τ_c^{app} from residue to residue are representative of local dynamics, and an increased τ_c^{app} corresponds to reduced local dynamics on the ns to ps timescale. Measuring τ_c^{app} for each residue in each protein, we searched for positive correlations between rigidity and RFP brightness (Fig. S4.7, S4.8, Methods), defined as positions where τ_c^{app} was greatest in bright RFPs and lowest in dim RFPs, thus suggesting that bright RFPs are more rigid at these positions than dim RFPs. In this manner, 12 residues exhibiting rigidity on short timescales that positively correlated to brightness were identified. This represents roughly half the number of positive correlations seen for moderate timescale dynamics as detected by HSQC peak intensities. By extension, this suggests that brightness-determining dynamics throughout the RFP barrel are tied primarily to the moderate-timescale dynamics detected by HSQC peak intensities as opposed to the fast dynamics detected by T_1 and T_2 relaxation rates.

4.3.3 Dynamics on the phenolate face of the RFP barrel correlate to brightness

Having now identified positions displaying rigidity correlated to brightness for both moderate and fast timescales, we wanted to verify that these correlations were real effects and not the result of random chance. To do so, we looked at two factors that support the veracity of these observations; first, whether clustering of positively correlating positions was observed, and second, whether these clusters excluded inversely correlating positions where a decrease in rigidity led to increased brightness. For a better overview of the spatial patterning of these correlations, we mapped both positive and negative correlations between rigidity and brightness on the RFP scaffold (Fig. 4.3). Observing both an enrichment of positive correlations and a lack of negative correlations in one region of the RFP scaffold would further support that dynamics in that region are linked to protein brightness. Visually, we pinpointed two such regions in the RFP scaffold. The

first spans the β-sheet formed by β-strands 7-10 (Fig. S4.9) along the face of the protein nearest the chromophore's phenolate moiety, thus we have termed this region the "phenolate face" of the protein. Roughly 58% of moderate timescale and 50% of fast timescale positive correlations are found on this face, which also presents no inverse correlations. Within this region, fast-timescale dynamics detected by T_1-T_2 relaxation experiments are observed near the RFP C-terminus and giving way to moderate-timescale dynamics detected by HSQC peak intensities along the center of the β-strands. Another roughly 16% of moderate timescale positive correlations are located in a tight cluster on the N-terminal portion of the central α-helix that supports the chromophore, again without the presence of negative correlations along this helix. The remaining positive correlations for both timescales are found dispersed throughout the rest of the scaffold without any clear clustering.

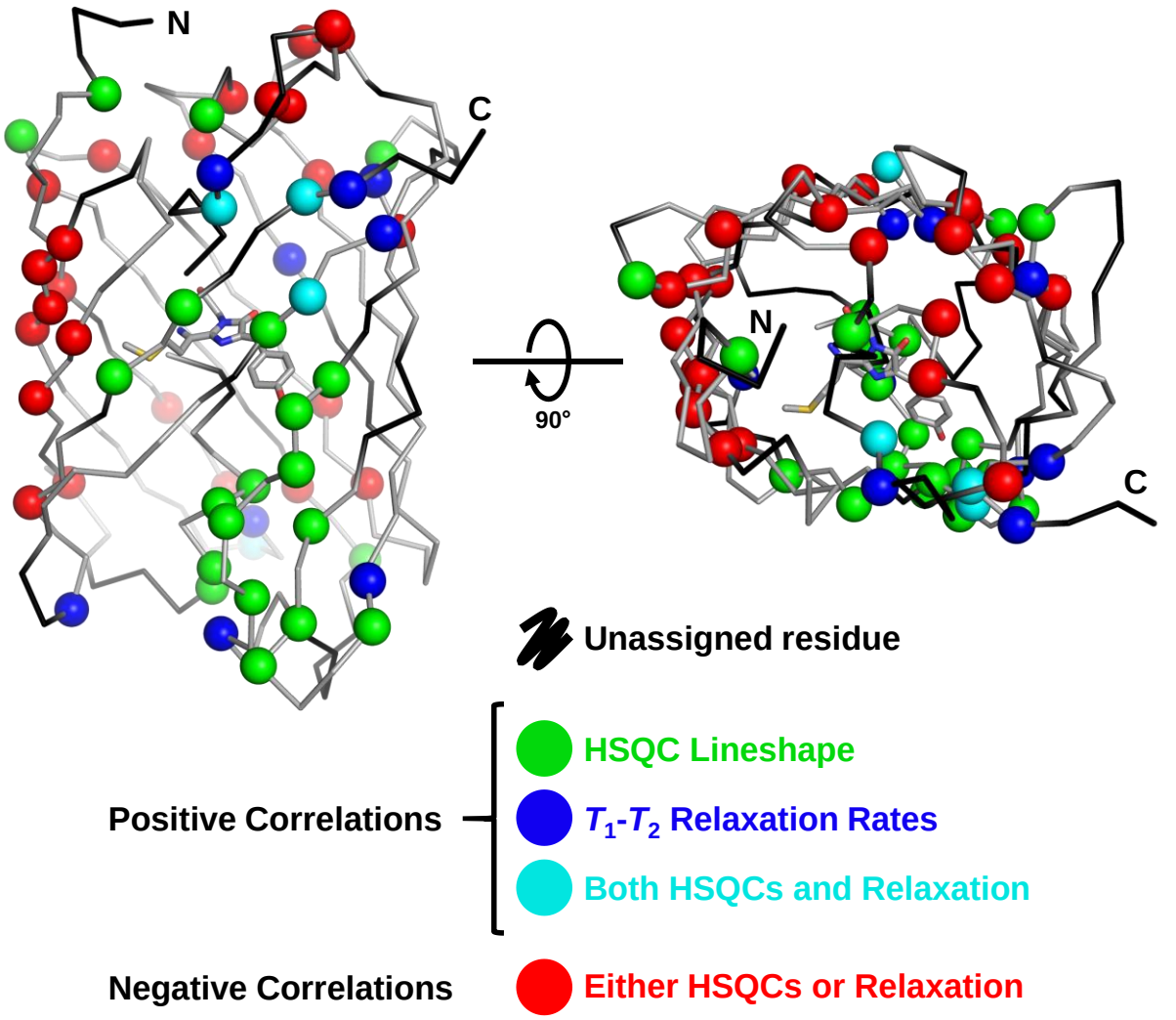


Figure 4.3. Summary of correlations between rigidity and brightness throughout the RFP backbone. Positive correlations between rigidity and brightness observed in HSQC lineshape and T_1-T_2 relaxation rate experiments are mapped as green and blue spheres respectively onto a ribbon representation of an archetypal RFP backbone (mCherry). Positions where positive correlations were observed in both experiments are shown as cyan spheres. Positions where negative correlations between rigidity and brightness were observed in either experiment are shown as black spheres. Positions that could not be unambiguously assigned in both mRouge and mRojoA, or any other single RFP from our dataset are identified as a black coloration on the ribbon representation.

Altogether, the strong prevalence of positive correlations on the phenolate face of RFPs and to a lesser extent along the central σ-helix pinpoints these two regions as key to the link between RFP dynamics and their brightness. In both cases, the regions exhibiting these correlations also stretch far from the chromophore's position at the center of the barrel, reaching the loops at

the ends of the β-barrel and suggesting that distal sites could influence chromophore dynamics and brightness. Along the phenolate face, these distal sites are linked back to the chromophore's surroundings by a series of residues exhibiting positive correlations between rigidity and brightness that run from one extremity of the barrel to the other. Though this was not observed for the residues on the central α-helix, several residues adjacent to the chromophore along this helix could not be unambiguously assigned due to P63 and the chromophore amino acids M66, Y67, and G68 all being undetectable in NH-detected NMR spectra.

4.4 Discussion

RFPs have been extensively engineered through both rational design and directed evolution, including projects aimed at improving their brightness.[67,82,83,217-219] As several of these engineering attempts have introduced mutations throughout the structure of various RFPs, searching for those that led to increased brightness, we could therefore expect to find mutational information supporting that chromophore brightness is linked to dynamics on the phenolate face and central α-helix of RFPs. mScarlet provides an initial assessment of brightness-increasing mutations on the mCherry structure, and with 43% of total mutations or 68% of β-sheet mutations in mScarlet located along β-strands 7-10 (Fig. 4.1), we note concordance between positions mutated in the bright mScarlet RFP and our results, though this is unsurprising given the inclusion of mScarlet in our experimental NMR dataset. Extending this analysis to other engineering projects where RFP brightness was improved, we tabulated the mutations introduced in each (Table S4.1). To reduce noise introduced by artifacts of random mutagenesis, we selected only positions mutated in more than one engineering project and mapped these positions onto the RFP scaffold (Fig. 4.4). Once again, most of these positions fall along β-strands 7-10, with 55% of identified positions being in this region of the protein structure. This suggests that the RFP phenolate face is tightly

linked to chromophore brightness, and now, with our results, that dynamics on the phenolate face are important to brightness. However, though existing brightness-increasing mutations are located on the same β-strands that we pinpointed as important to brightness, mutated residues cluster primarily near the center of strands 8-9 as opposed to clustering nearer the extremities of strands 8-10 and the center of strand 7 as shown by our NMR results.

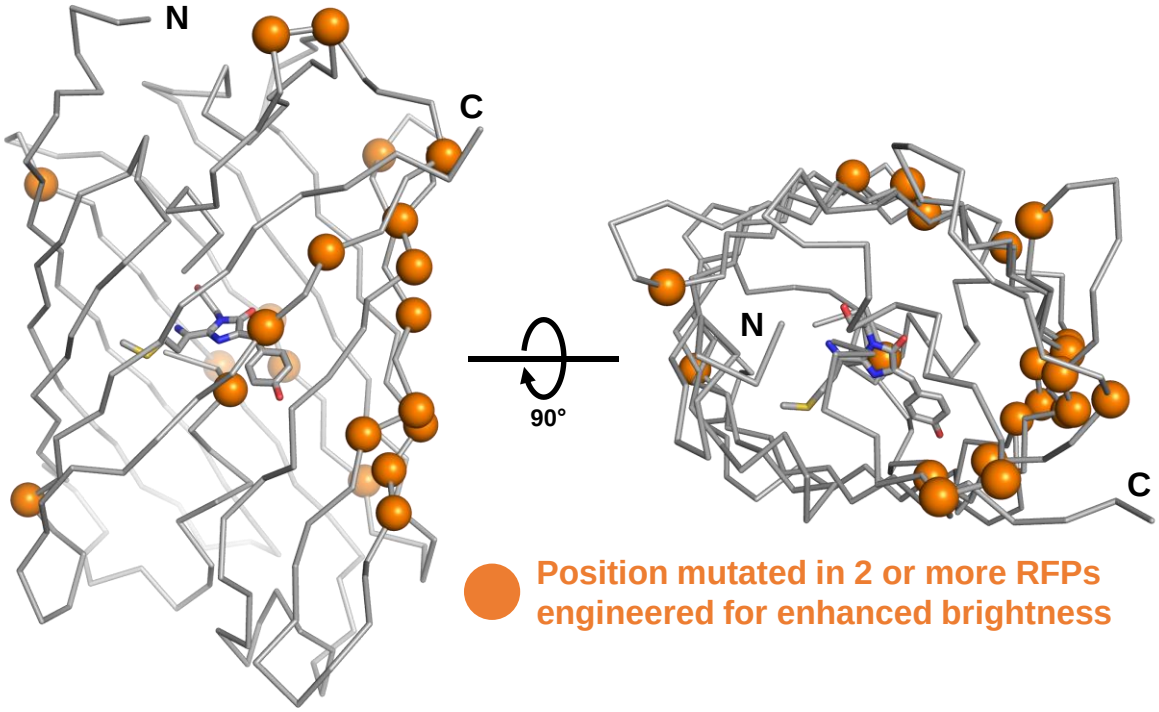


Figure 4.4. Positions mutated in previously engineered RFPs. Positions mutated in 2 or more engineered RFPs with greater quantum yield than the parent protein are shown mapped as orange spheres onto a ribbon representation of an archetypal RFP backbone (mCherry). The RFP chromophore is shown as sticks. In all cases, positions are shown relative to the corresponding mCherry amino acid.

This discrepancy could be explained by the relationship between RFP core packing and chromophore maturation rates. As the RFP chromophore matures into its fluorescent form, two oxidation steps are required.[68] Oxygen access to the chromophore is therefore crucial for maturation to proceed. Previous studies have shown that this access occurs partially through the

channel formed between the extremities β-strands 7 and 10.[228] Thus, mutations to these and nearby β-strand extremities that enhance packing and stabilize this region could slow oxygen access to the protein core and hinder maturation even as they might improve the quantum yield of a mature chromophore. As most RFP directed evolution projects screen only for total brightness of cells, sufficiently maturation-deficient mutants would be of poor fitness, even if mature proteins possess high brightness. Mutating slightly further from the channel, such as near the center of β-strands 8 and 9 as is seen in most directed evolution-driven RFP engineering projects, might provide a better balance between maturation, through continued oxygen access to the chromophore, and brightness, through trickle-down stabilization of critical structural elements. Interestingly, mScarlet was derived from a novel directed evolution-based screening approach for quantum yield that was independent of protein maturation,[83,89,229] and presents mutations at otherwise rarely engineered positions (Fig 4.1, 4.4) on the extremities of β-strands 8 and 9 and throughout the phenolate face. Though these mutations contribute to mScarlet being one of the brightest RFPs designed to date, mScarlet is also maturation-deficient in comparison to mCherry, with a maturation half-time over three times that of mCherry *in vivo*,[230] supporting that though mutations along the extremities of the phenolate face can influence brightness, they can also adversely affect maturation, hence their difficulty to pinpoint and develop using directed evolution. When developing novel RFPs for imaging applications, this brightness-maturation tradeoff will need to be addressed given that slow maturation restricts the utility of an RFP. Our results demonstrate that it may be possible to develop a brighter yet still efficiently maturing RFP by mutating surface positions along the phenolate face to form additional stabilizing polar interactions between β-strands 7-10 without further enhancing core packing that inhibits oxygen access to the chromophore.

Beyond the design of RFPs bearing desirable properties such as enhanced brightness, we are also interested in using RFPs as a model system for the design of functional dynamics. In RFPs, dynamics are intrinsically linked to an easily screened function, in this case brightness, making them an ideal system for tackling this problem. With the phenolate face and central α-helix identified as regions where rigidity correlates to brightness in RFPs, we now also possess the necessary structural information to guide design positions, where a design for reduced dynamics could in turn increase brightness. The difficulty here would lie in selecting proper backbone ensembles representing rigid and dynamic RFPs against which to design, as well as a scoring strategy capable of discriminating between sequences that restrict dynamics given that structures for a low-diversity RFP ensemble would likely be included within the range of movement of a high-diversity ensemble. Alternately, though we here focused on positions where rigidity correlated positively with brightness, negative correlations also provide information about the link between dynamics and brightness. Though chromophore rigidity is directly linked to its brightness, it is reasonable to expect that increased flexibility in certain distal regions of the RFP barrel might stabilize the chromophore's environment, leading to increased brightness. In this vein, we observed a cluster of residues correlated negatively to brightness along the portion of β-strands 2 and 3 nearing the protein's N-terminus (Fig 4.3, S4.9). It is possible that designing increased dynamics in such a region could in turn increase brightness as well, though this design would be further complicated by the possibility of destabilizing the protein when engineering enhanced flexibility.

Despite these yet outstanding barriers to the computational design of RFP brightness through dynamics, our results have shown that dynamics on the phenolate face and central α-helix of RFPs correlate with the protein's brightness. Though directed evolution designs have begun to

mine these regions for brightness-increasing mutations, such as the extremely bright mScarlet, the tradeoff in maturation rates mutations in this region may cause has left it relatively untouched, despite the correlation to brightness revealed here, and future designs for brighter RFPs would benefit from studying this region. With the structural information presented herein in hand, we are also ready to begin computationally designing functional dynamics into proteins, a goal with far reaching applications such as the design of allostery, catalysis, and more.

4.5 Methods

4.5.1 Protein expression and purification

The mScarlet gene was obtained from Addgene (pCytERM_mScarlet_N1 was a gift from Dorus Gadella – Addgene plasmid #85066; http://n2t.net/addgene:85066; RRID:Addgene_85066) and cloned into the pET-11a vector with an added N-terminal His-tag. His-tagged genes cloned into the pET-11a vector for other RFPs tested were available in house. RFPs were expressed using M9 minimal expression medium supplemented with 1 g/L ^{15}N-ammonium chloride and/or 3 g/L ^{13}C-D-glucose for isotopic enrichment. Cultures were grown at 37 °C with shaking to an OD600 of approximately 0.6 after which protein expression was initiated with 1 mM isopropyl β-D-1-thiogalactopyranoside. Following overnight incubation at either 16 °C (mRojoA, mRouge, and mPlum-E16P) or 37 °C (mCherry and mScarlet) with shaking (250 rpm), 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 (Qiagen), followed by gel filtration in 10 mM sodium phosphate buffer (pH 7.4) using an ENrich SEC 650 size-exclusion chromatography column (BioRad). Purified samples were concentrated using Amicon Ultracel-3K centrifugal filter units (EMD Millipore) or Macrosep Advance 10K centrifugal devices (Pall).

4.5.2 NMR spectroscopy

^{15}N- and ^{13}C-labelled Gβ1 samples for NMR consisted of 0.1–0.2 mM uniformly labelled protein in 10 mM sodium phosphate buffer (pH 7.4), 10 μM EDTA, 0.02% sodium azide, 1× cOmplete EDTA-free Protease Inhibitor Cocktail (Roche), and 5% D_2O. All NMR experiments were performed at 25 °C on a Bruker AVANCEIII HD 600 MHz spectrometer equipped with a triple resonance cryoprobe. NMR data sets were processed with the NMRPipe software package[199] and spectra were analyzed with NMRViewJ (One Moon Scientific).[200] Backbone chemical shift assignments were obtained from the standard suite of 3D triple resonance experiments, including HSQC, HNCO, HNCACB and CBCA(CO)NH spectra. Dynamics characterization experiments used included HSQC spectra, as well as T_1 measurement and T_2 measurement spectra,[143,231,232] and CPMG relaxation-dispersion spectra.[138] ^{15}N T_1 values were measured using the hsqct1etf3gpsi3d.2 pulse program, with delay times set to 10, 20, 30, 50, 100, 200, 300, 500, 750, 1000, 1500, and 2000 ms. ^{15}N T_2 values were measured using the hsqct2etf3gpsi3d pulse program, with loop counters for the CPMG pulse train set at 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, and 16. The length of a single CPMG loop was 16.96 ms in the experimental setup.

4.5.3 Dynamics analysis

HSQC lineshapes were deconvoluted and analyzed by fitting a Gaussian model to peaks in distinct clusters using NMRDraw.[199] Peak intensities were converted to a deviation metric by dividing peak intensities for each residue by the standard deviation of peak intensities for that protein, then normalizing to the median peak intensity for that protein. Residues were considered to have rigidity positively correlated to brightness where the absolute value of mScarlet's peak intensity deviation was lowest, followed in order by mCherry, mPlum-E16P, and with either mRouge or mRojoA displaying the highest intensity deviation. Residues were considered to have

rigidity negatively correlated to brightness where the inverse order was observed. Due to poor spectral quality for mRouge and mRojoA spectra, positions were discarded only if neither mRouge nor mRojoA displayed the highest intensity deviation, and the two proteins were not compared to one another due to their similar QY. In addition, a minimum absolute peak intensity deviation threshold for mRouge or mRojoA of at least 0.5 was implemented for a residue to be considered to have rigidity positively correlated to brightness to eliminate potential false positives at positions that were rigid in all RFPs tested.

T_1-*T_2* relaxation experiments were analyzed using the Bruker Dynamics Center software, by fitting a Gaussian model to peaks and fitting peak intensities over 12-point relaxation curves using non-linear regression. Correlation time measurements were derived from *T_1* relaxation and *T_2* relaxation values, calculated using the approximation $\tau_c^{app} \approx \frac{1}{4\pi\nu_N}\sqrt{6\frac{T_1}{T_2} - 7}$ and converted to a normalized correlation time metric by dividing correlation times for each residue by the median correlation time for that protein to remove error introduced by variation in the global protein tumbling rate between proteins. Residues were considered to have rigidity positively correlated to brightness where mScarlet's normalized correlation time was highest, followed in order by mCherry, mPlum-E16P, and with either mRouge or mRojoA displaying the lowest normalized correlation time. Residues were considered to have rigidity negatively correlated to brightness where the inverse order was observed. Due to poor spectral quality for mRouge and mRojoA spectra, positions were discarded only if neither mRouge nor mRojoA displayed the highest intensity deviation, and the two proteins were not compared to one another due to their similar QY.

4.5.4 Data availability

Chemical shift assignments have been deposited in the Biological Magnetic Resonance Data Bank with accession codes 27906 (mCherry), 27907 (mPlum-E16P), 27908 (mScarlet), 27909 (mRouge), and 27910 (mRojoA).

4.6 Supplementary Information

Table S4.1. Positions mutated in RFPs engineered for enhanced brightness. [a]

Protein	Brightness ($mM^{-1}cm^{-1}$)	Parent	Parent Brightness ($mM^{-1}cm^{-1}$)	Mutations [b]	Reference
mScarlet	70.0	mCherry [c]	15.8	I8V / V23M / A58S / K71R / Y73F / V74T / L84Y / L86Q / V105A / S112T / Q115E / E118T / F119L / S132P / S148T / M151L / A157V / E161D / Q164M / R165A / K167R / H173R / D175L / E177D / V178F / L190M / N197D / I198R / I211V / A218S	83
mRojo-VHSV	4.68	mRojoA	0.96	T16V / P63H / W143S / L163V	82
mRojo-VSHVLF	5.41	mRojo-VHSV	4.68	M150L / L165F	– [d]
mCherry-AYC	5.04	mCherry-I197Y	2.30	V195A / A217C	193
mKate2	25.0	mKate	14.8	V45A / M146T / S158A / K231R	233
mNeptune2	21.4	mNeptune	13.4	A104V / I121L / I171H / K207N	217
mRuby3	57.6	mRuby2	42.9	N33R / M36E / T38V / K74A / G75D / M105T / C114E / H118N / Q120K / H159D / M160I / S171H / S173N / I192V / L202I / M209T / F210Y / H216V / F221Y / A222S / G223N	218
mMaroon1	8.80	Maroon0.1	5.50	M11T / M15L / E16T / H23Y / K25E / S28A / G41N / E47R / T73P / Q74P / G75D / F79Y / V93T / T95V / V101T / V104A / K120Q / L121V / V124E / G153S / C158L / N173R / K175E / V195I	234
mCarmine	5.81	mNeptune684	1.17	C61S / T103K / A104V / T105K / H157Y / P159T / I171Q / C172T / N173F	219
mKelly2	7.74	mKelly1	7.04	H72Y / T146Y / V155E / R157T / K192E / Y193H	222

[a] Including RFPs with entries in the fpbase.org FP database and RFPs engineered in-house.
[b] Mutations are indicated relative to the parent and numbering used in the seminal reference for the engineered RFP.
[c] Though mScarlet was evolved from the non-fluorescent synthetic protein mRed7, mRed7 is in turn primarily derived from mCherry.
[d] First published in this work.

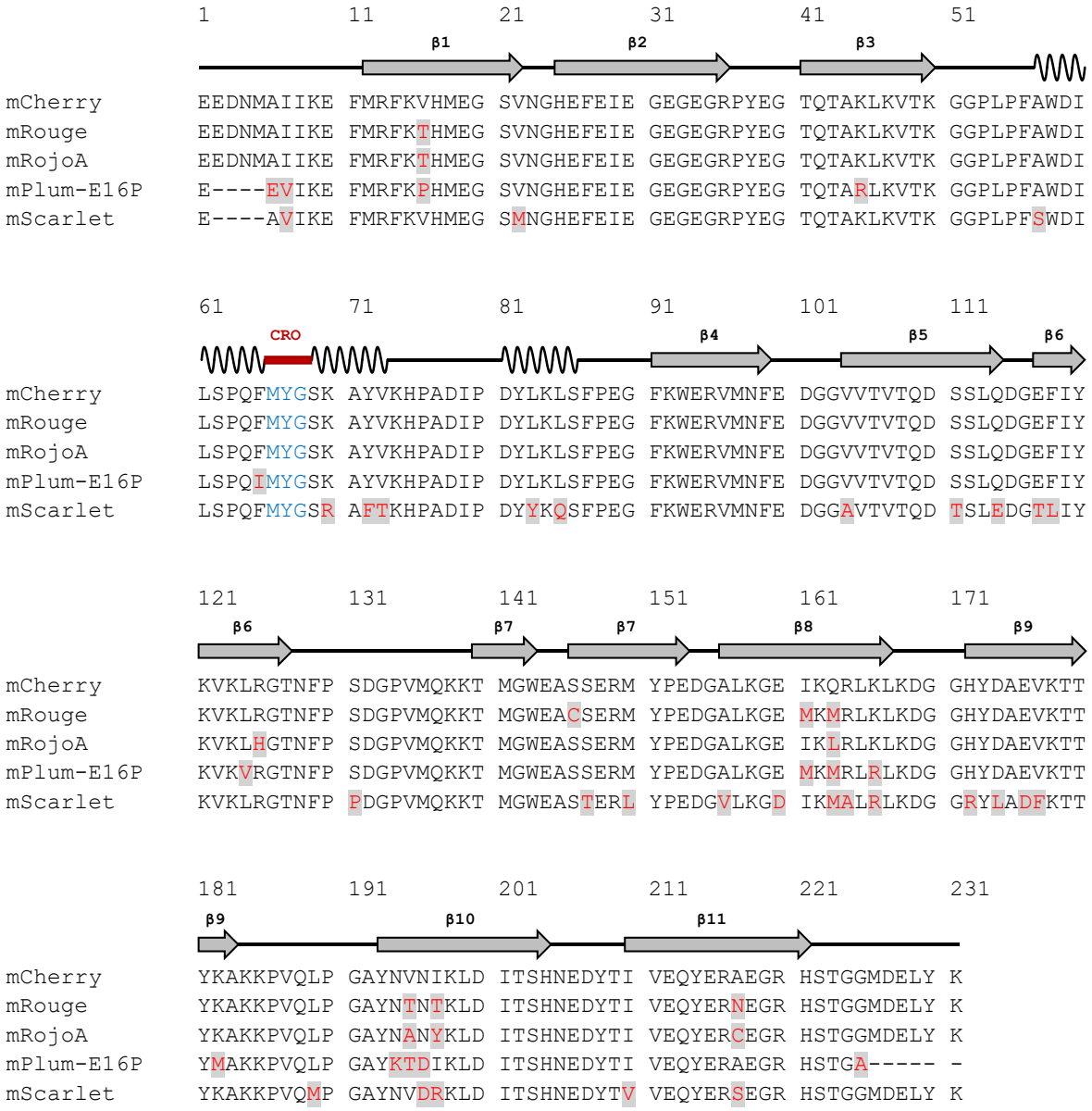
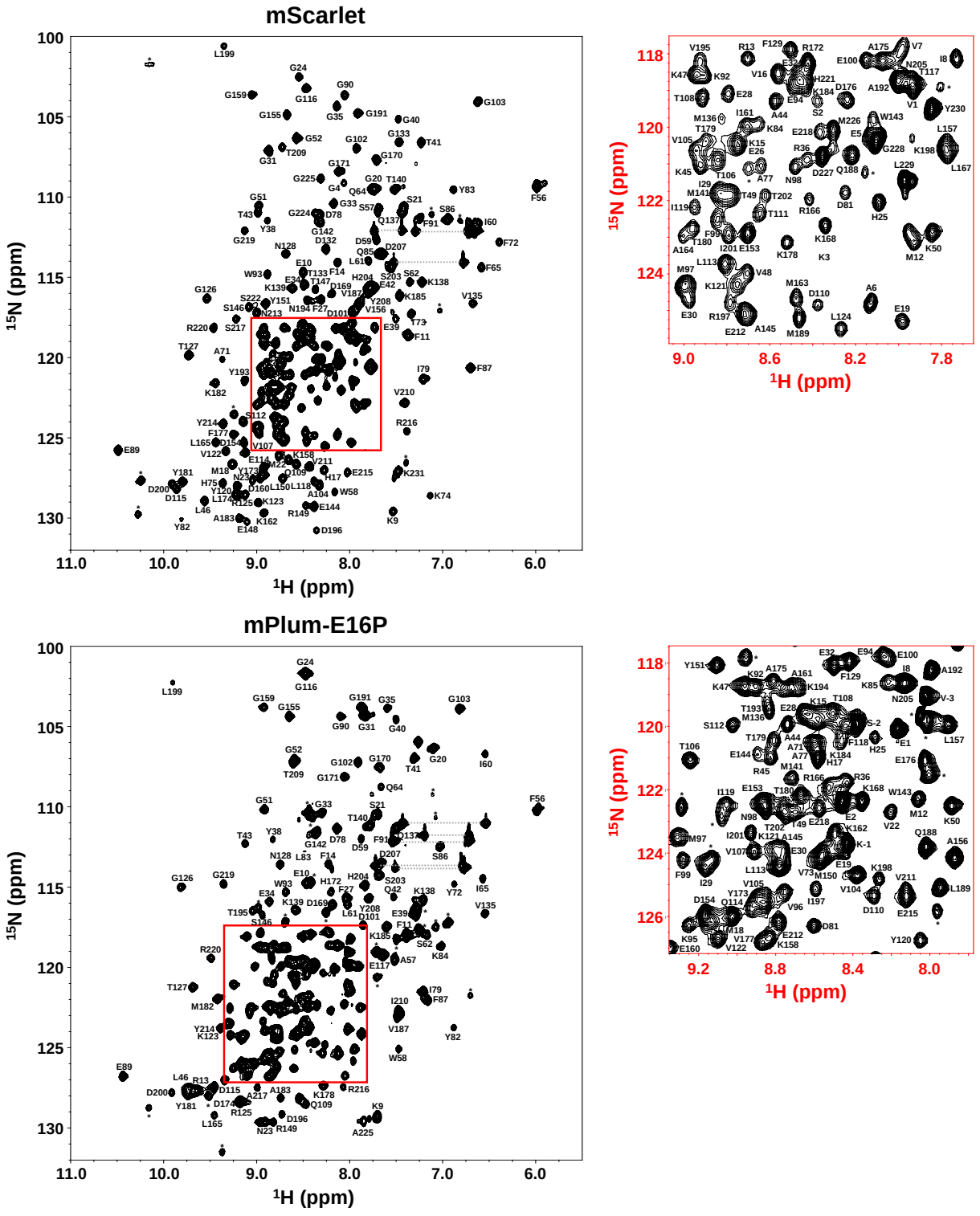


Figure S4.1. Sequence alignment of selected mCherry-derived RFPs. Sequences for selected mCherry derivatives are shown aligned to the mCherry sequence. Amino acid numbering relative to mCherry will be used throughout this work. Point mutations are shown in red with grey shading, and the chromophore-forming amino acids are shown in blue. mCherry secondary structure elements are also shown aligned to the primary sequences.


mScarlet
15N (ppm)
1H (ppm)
mPlum-E16P
15N (ppm)
1H (ppm)

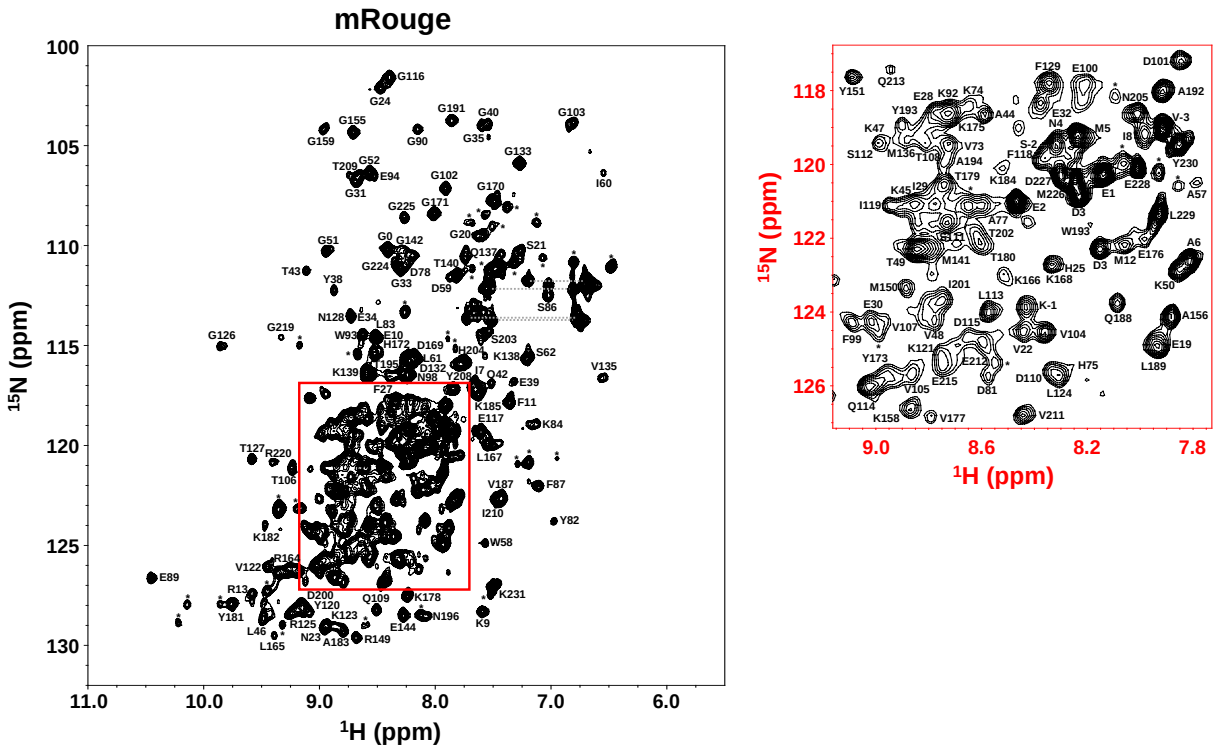


Figure S4.2. Assigned RFP ^{1}H-^{15}N HSQC spectra. HSQC spectra for mScarlet, mPlum-E16P, and mRouge are shown labeled with unambiguously assigned residues identified. Side-chain amide resonances from asparagine and glutamine residues are connected by horizontal lines, and unassigned or other side-chain resonances are denoted by an asterisk. The central portion of the spectrum (boxed in red) is shown separately for clarity.

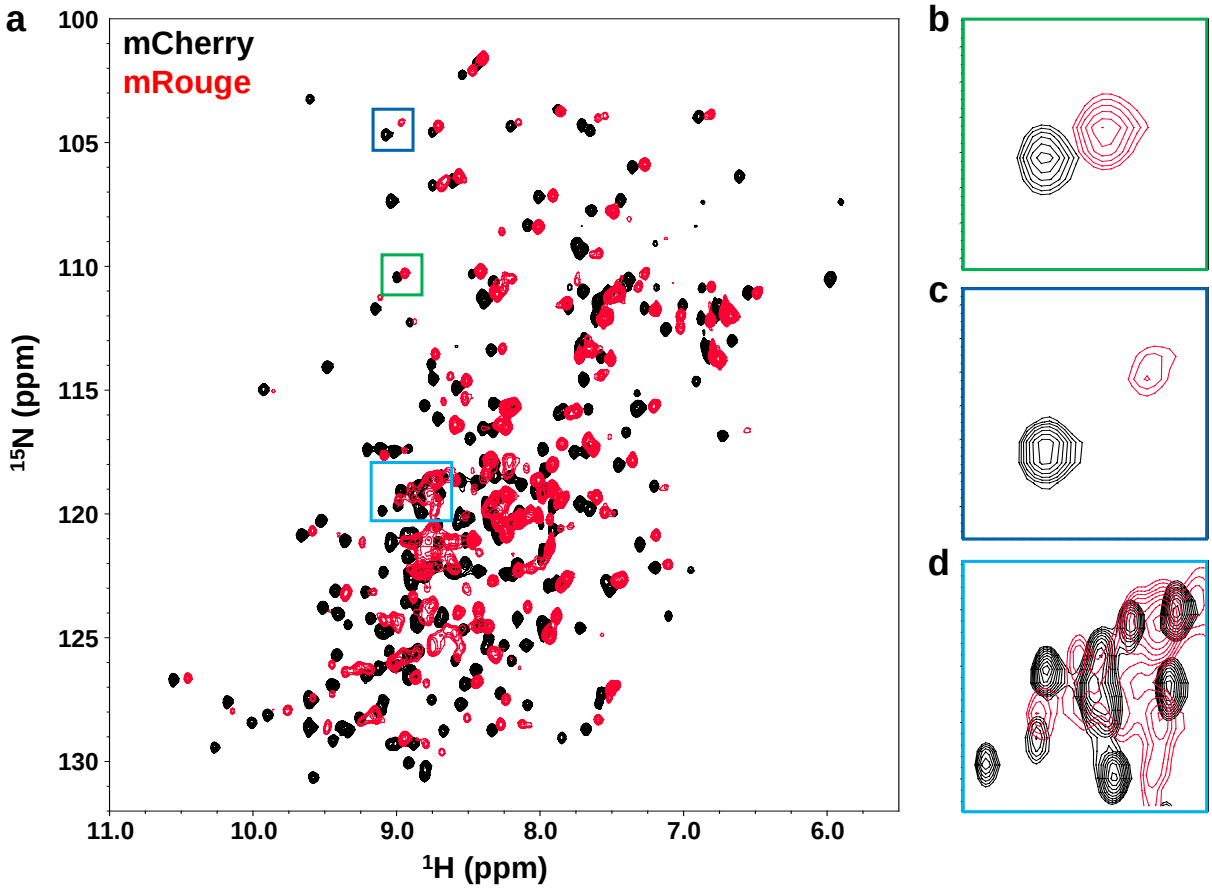


Figure S4.3. Dim RFPs demonstrate peak broadening throughout ^{1}H-^{15}N HSQC spectra. a, HSQC spectra for the bright mCherry (black) and dim mRouge (red) fluorescent proteins are overlaid and concentration normalized to demonstrate systematic peak broadening observed in the mRouge spectrum. **b,** Certain rigid residues are not significantly broadened in mRouge, as seen with the G51 residue shown herein. **c,** Certain residues experience significant peak broadening and loss in intensity in comparison to their mCherry equivalent, as seen with the G159 residue shown herein. **d,** A clear loss of peak resolution due to peak broadening can be observed in the central region of the HSQC spectrum, as seen in the representative cut-out shown herein.

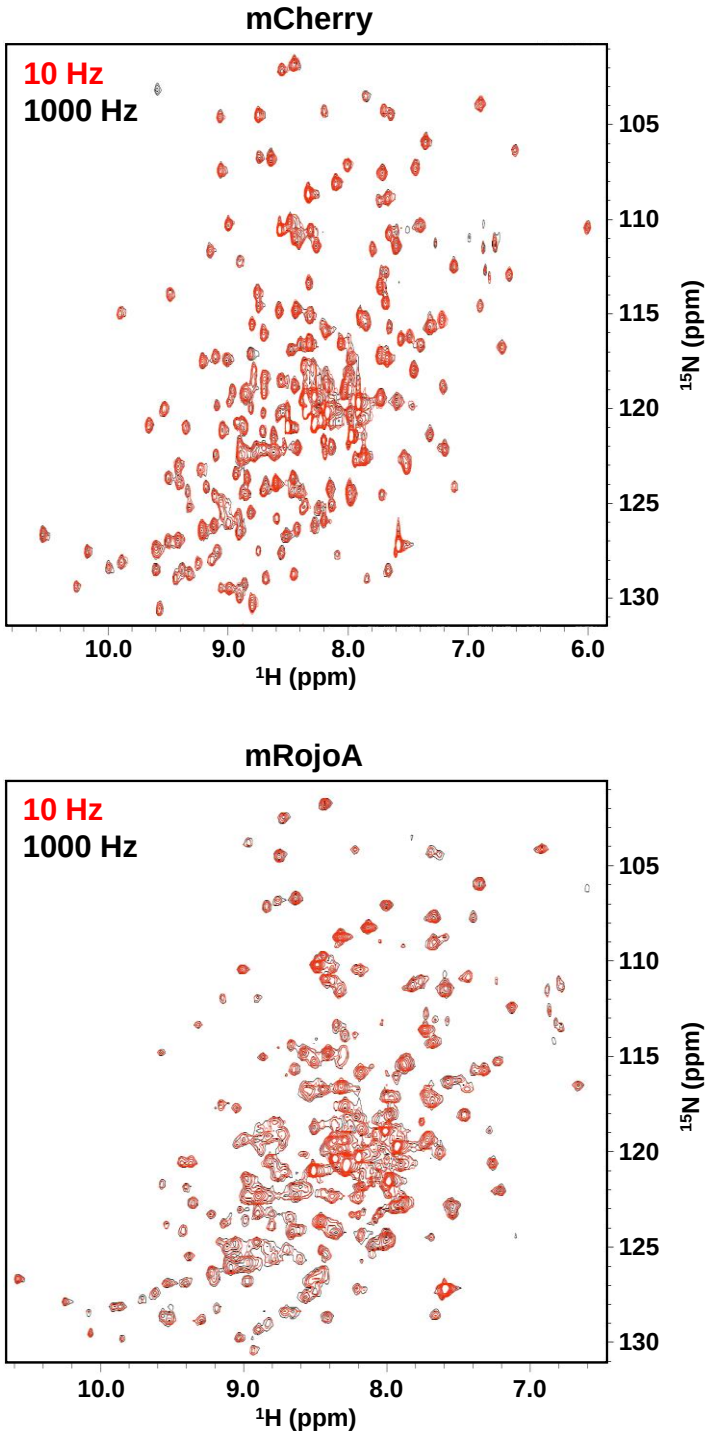


Figure S4.4. RFPs are insensitive to CPMG relaxation-dispersion experiments. CPMG relaxation-dispersion spectra were run with field strengths ranging from 10 Hz to 1000 Hz for the bright mCherry and dim mRojoA RFPs. In both cases, spectra at these extremes were almost identically superimposable, with only minor variations in peak volumes and intensities for most residues. This suggests that RFPs are rigid on the dynamic timescale and/or range of motion probed by CPMG.

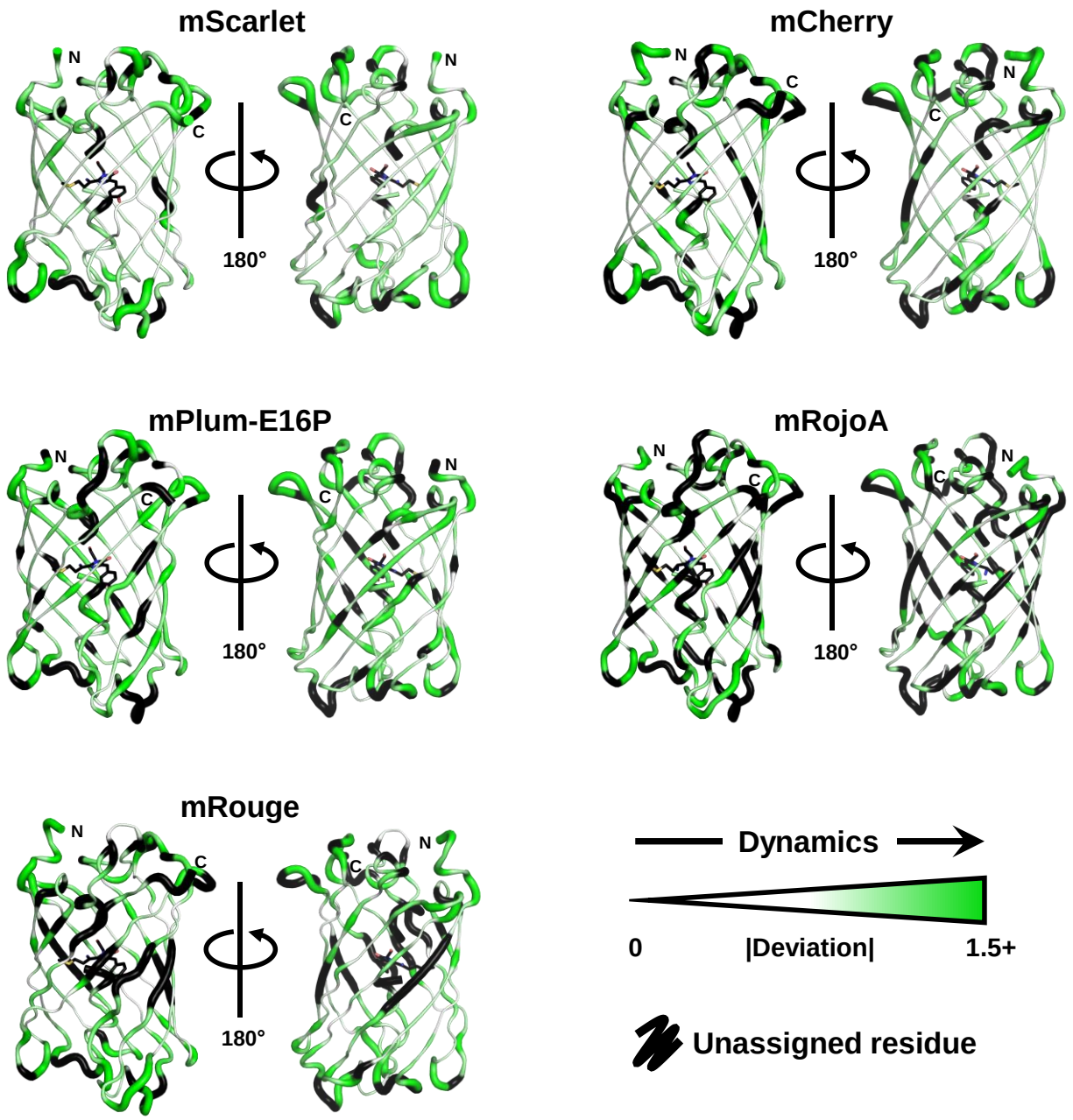


Figure S4.5. RFP dynamics as detected by HSQC peak intensity measurements. Dynamics as represented by deviation from the average peak intensity are mapped for each test protein onto their respective backbone. A deviation of 0, in white and with a thin cartoon sausage, represents a rigid residue on the timescales studied, increasing proportionally to dynamics along a green gradient scale and with increasing cartoon width. Positions which could not be unambiguously assigned are shown in black.

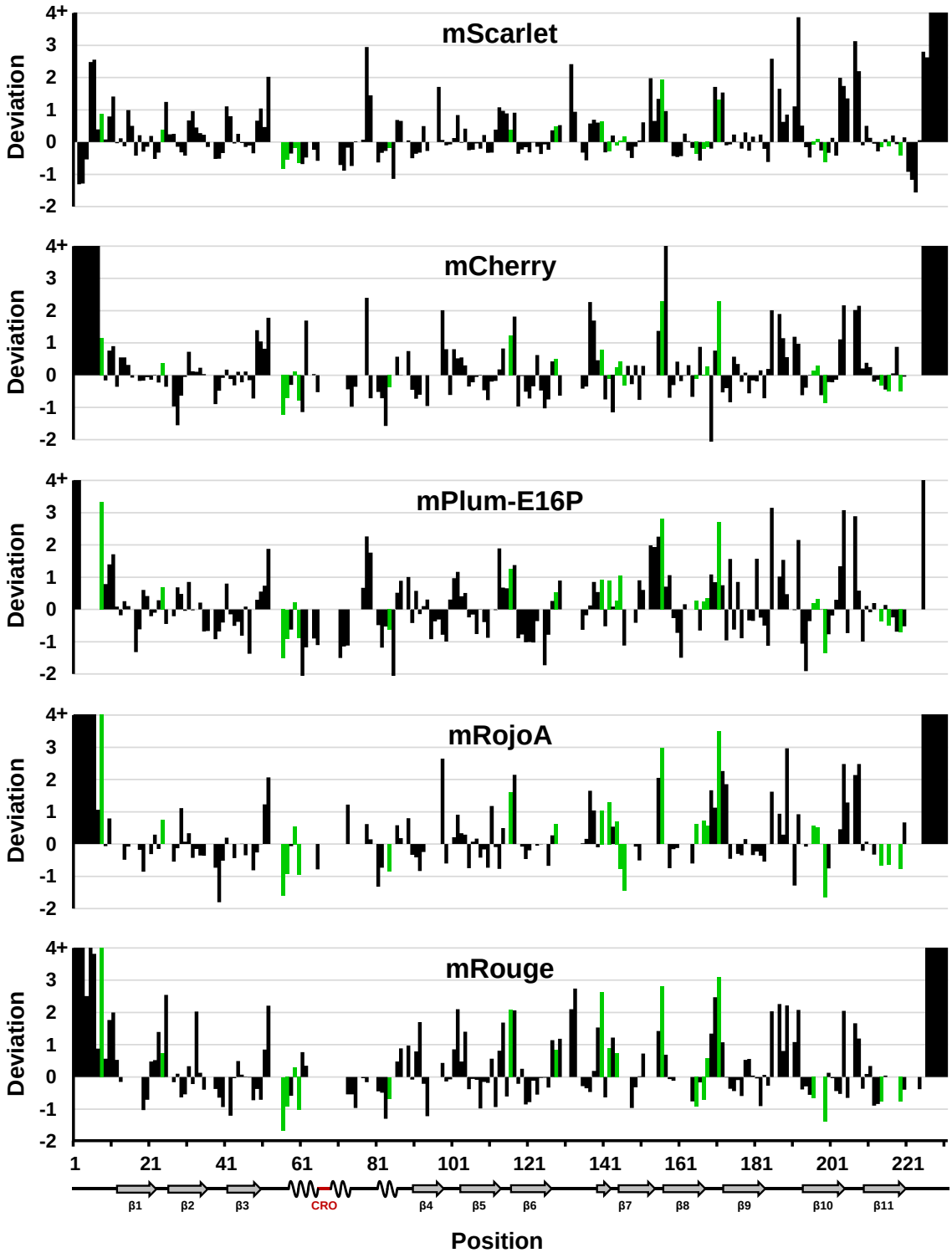


Figure S4.6. HSQC peak intensity deviation by residue. Peak intensity deviation is presented as a histogram for each protein, with green bars identifying positions where a negative correlation between deviation and brightness (representing a positive correlation between rigidity and brightness) is observed. mCherry secondary structure element positions are shown adjacent to the x axis.

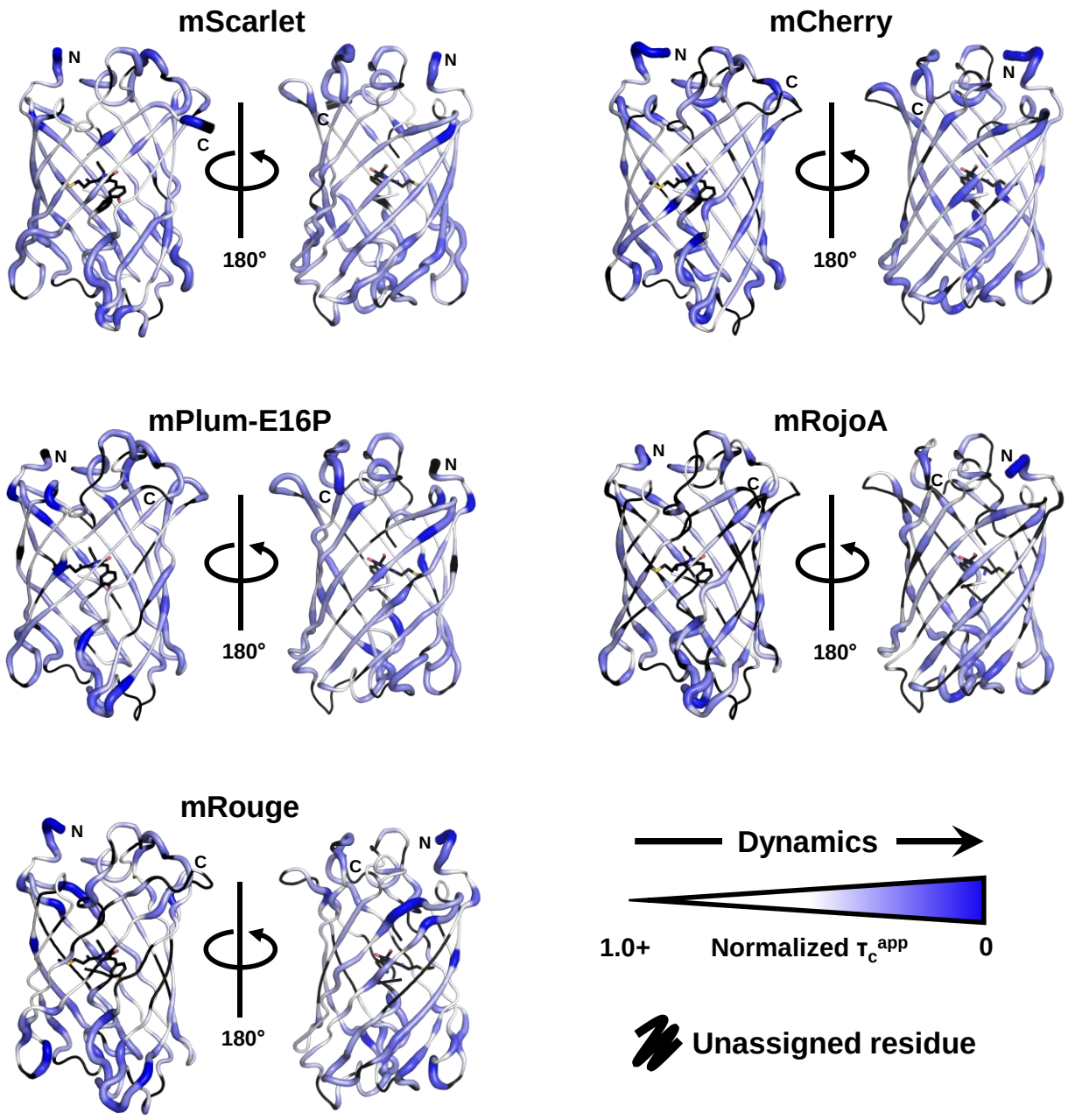


Figure S4.7. RFP dynamics as detected by correlation time measurements. Dynamics as represented by normalized apparent correlation time (τ_c^{app}) are mapped for each test protein onto their respective backbone. A normalized correlation time of 1.0 or greater, in white and with a thin cartoon sausage, represents a rigid residue on the timescales studied, decreasing proportionally to dynamics along a blue gradient scale and with increasing cartoon width. Positions which could not be unambiguously assigned are shown in black.

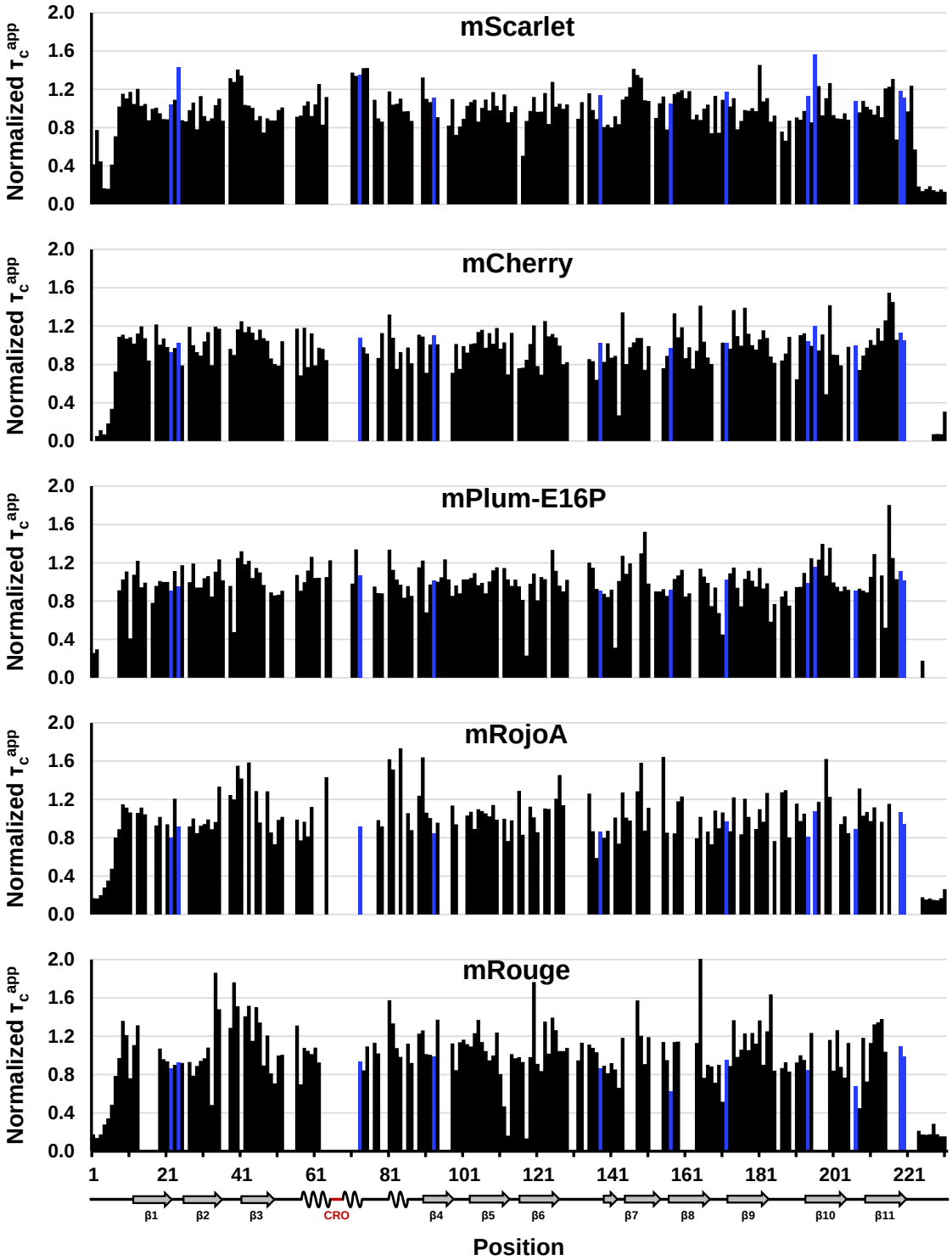


Figure S4.8. Normalized correlation time by residue. Normalized correlation time is presented as a histogram for each protein, with blue bars identifying positions where a positive correlation between correlation time and brightness (representing a positive correlation between rigidity and brightness) is observed. Median correlation times were of 16.3 ns, 16.5 ns, 14.2 ns, 18.4 ns, and 16.6 ns for mScarlet, mCherry, mPlum-E16P, mRojoA, and mRouge respectively. mCherry secondary structure element positions are shown adjacent to the x axis.

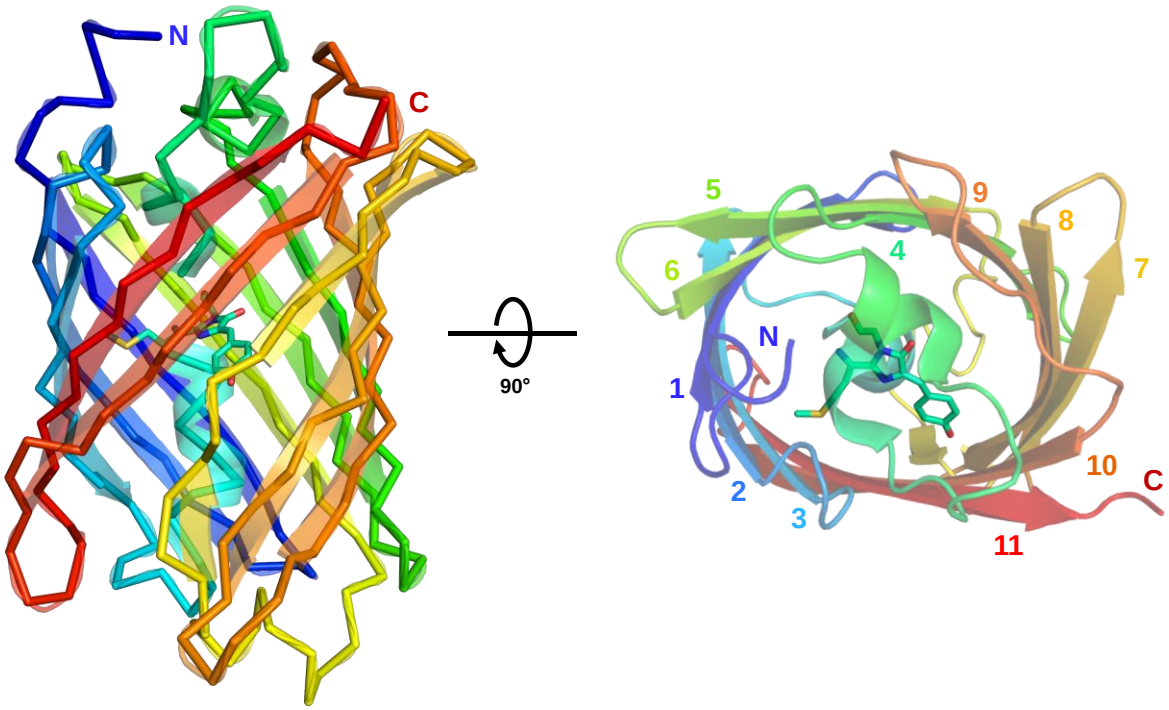


Figure S4.9. RFP β-strand ordering. An archetypal RFP β-barrel structure from the mCherry crystal structure (PDB ID: 2H5Q[70]) is shown colored by a blue-to-red rainbow spectrum from N-terminus to C-terminus to highlight strand ordering. On a top-down view of the barrel, strands are numbered accordingly.

Chapter 5: Discussion and perspectives

5.1 Summary

A better understanding of protein structure-function relationships is necessary to further the field of rational protein engineering and design. Notably, dynamics, which are known to play a major role in the function of a vast number of proteins,[32,166-168] have posed an obstacle to complex computational designs as it is not clear how to model dynamics in a CPD scenario. Standard MSD protocols have been developed that approximate a certain degree of mutability in the protein backbone through the use of backbone ensembles.[54,109,115,116] However, these protocols are not designed to differentiate hypothetical dynamic sequences capable of scoring well on several members of a backbone ensemble from more rigid sequences that satisfy only a few. At a more fundamental level, though it is well documented that protein sequence and dynamics are intrinsically interlinked, we also do not yet fully understand how sequence defines dynamic behavior. Therefore, this thesis presents (1) the development of a new CPD methodology termed *meta*-MSD aimed at overcoming known obstacles to the computational design of dynamics, (2) a characterization of sequence determinants leading to novel dynamics in a stable globular protein, and (3) a comprehensive study of brightness-correlated dynamics in red fluorescent proteins.

Chapter 2. Obstacles to the computational design of protein dynamics include the choice of backbone ensembles to use, and the choice of optimization criteria for the design's search algorithm. Using a combination of ensemble generation strategies and a post-hoc analysis of sequence energies that we termed *meta*-MSD, we designed Gβ1 variants exhibiting a novel mode of millisecond-timescale Trp43 conformational exchange. Three resulting variants, called DANCERs for Dynamic and Native Conformational ExchangeRs, were characterized by NMR

and confirmed to be dynamic as predicted, thus presenting the first successful application of CPD to design a specific mode of exchange into a stable protein fold.

Chapter 3. Following the successful design of DANCERs, we sought to better understand how the mutations included in these DANCERs led to the apparition of their novel dynamic exchange properties. By individually reverting each mutation from DANCER-3, we elucidated their roles in the designed dynamic exchange trajectory through a combination of NMR spectroscopy and MD simulations. In so doing, we determined that the development of dynamics had come about from the combination of a destabilizing effect on the highly stable native Gβ1 conformation, and a second, subtle interaction to stabilize the exchange trajectory that would have been difficult to predict through intuition, highlighting the power of the *meta*-MSD technique developed in Chapter 2.

Chapter 4. Our first design of dynamics introduced novel dynamics into a model protein without considering function. Now, we sought to expand the scope of our designs by altering protein function through the design of dynamics. For this purpose, we chose to work with RFPs, whose brightness is known to depend on chromophore dynamics. In this chapter, we performed a comprehensive study of dynamics in a family of related RFPs, searching for distal positions to design where dynamics were correlated to the protein's brightness. This led to the identification of a cluster of residues on the RFP phenolate face whose dynamics correlated with brightness throughout the series of RFPs studied, giving us a target for future designs and highlighting how dynamics that are distant from the chromophore might still affect its brightness.

Overall, the projects described in this thesis are one more step along the path to a better understanding of the link between protein sequence, structure, and function. In the future, the

knowledge gleaned may help us to engineer and design ever more complex protein functions that rely on dynamics, such as allostery, catalysis, and more.

5.2 Future directions

5.2.1 Design of a brighter RFP through the control of chromophore dynamics

Our study of dynamics in RFPs, presented in Chapter 4 of this thesis, was a precursor project to an application of CPD to functional dynamics. As stated, we chose to work with RFPs due to an easily screened function, fluorescence, which was known to be directly linked to chromophore dynamics.[88,223] With distal positions identified along the RFP phenolate face where dynamics correlate with brightness, we now possess the positional information needed to begin designing brighter RFPs through the control of dynamics. However, there are still several obstacles needing to be addressed for this design to prove successful.

The first obstacle ties in both to the nature of RFPs and to a long-standing weakness of CPD. Though we identified clusters of residues that appear important to brightness near the cleft between β-strands 7 and 10, mutating core residues at this location has been shown to adversely affect RFP maturation.[228] Though this is not in and of itself detrimental to a goal of computationally designing rigid, bright RFPs, long maturation times might reduce the usefulness of any improved RFP created through mutagenesis at these core positions. This might be avoided through the design of surface residues where introducing new strand-to-strand interactions to rigidify the phenolate face proves a promising alternative to improving core packing in this region. Physics-based CPD approaches, however, remain poor at predicting the effects of surface mutations on protein folding and stability due to the complexity of protein-solvent interactions and inaccuracies in the treatment of other non-bonded interactions such as long-range

electrostatics.[98,235-237] Thus, to supplement our structural dynamics dataset and guide future designs, we are in the process of a semi-rational study of the link between the structure of the RFP surface and brightness. Through saturation mutagenesis of surface positions in and adjacent to the brightness-linked portion of the RFP phenolate face identified by NMR (Fig. 5.1), we will identify which amongst them allow mutations, and whether any resulting mutants are brighter than the wild-type protein, in this case mPlum-E16P as it is the brightest mCherry-derived far-red FP for which a high-resolution structure is available. This empirical data will help to further restrict design positions and will provide a training set against which to benchmark computational methods, which will prove critical in accurately assessing the effects of surface mutations that may be necessary to include to meet design goals.

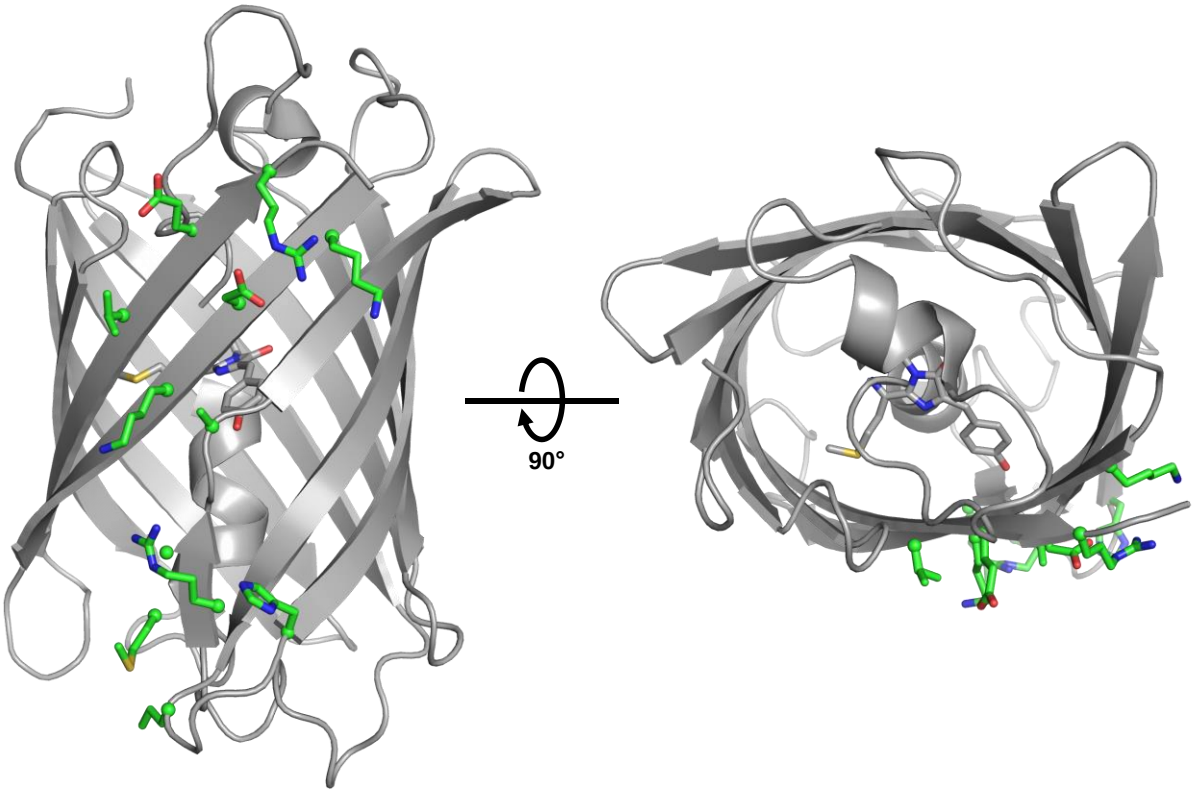


Figure 5.1. mPlum-E16P surface positions targeted for saturation mutagenesis. Positions targeted for mutagenesis are shown as green sticks with the mPlum-E16P crystal structure (PDB ID: 4H3L) shown as a cartoon. A top-down view confirms that all targeted positions correspond to surface residues.

The second obstacle returns to a problem that was addressed in the design of DANCERs; that of optimization criteria for the CPD search algorithm. Previously, our application of *meta*-MSD had been the development of a novel mode of dynamics involving distinct, non-native conformations. With the planned RFP design, however, this is not the case. Rather, in RFPs, flexibility and small motions dominate dynamic differences as evidenced by the moderate-to-fast μs to ps dynamics observed in NMR, rather than slower processes denoting larger motions. Thus, dynamic and rigid RFPs share a strong structural resemblance, raising the question of how to discriminate between them in a CPD protocol. Though molecular dynamics simulations can be used to generate backbone ensembles of varying diversities using different RFPs as starting

templates (Fig. 5.2), backbone configurations from the low-diversity ensemble appear to be included within the range observed in the high-diversity ensemble. A sequence that scores well on the low-diversity ensemble therefore could be expected to also score well on the high-diversity ensemble. As a result, standard positive or negative MSD approaches do not seem promising to tackle this design. Rather, an alternate sequence scoring methodology will be needed to highlight the diversity allowed by a given sequence. *Meta*-MSD would be particularly well suited to this task, given its decoupling of scoring from sequence optimization, where an RMSD for all ensemble members scoring within a certain energy threshold from the most stable member could be used to predict sequence dynamicity in a post-hoc analysis (Fig. 5.3a). As *meta*-MSD is very computationally demanding, this approach limits the scope of the sequence space that could feasibly be tested. However, the *meta*-MSD approach could be coupled to standard MSD to first eliminate sequences that are not sufficiently stable on any ensemble member. Other potential alternatives could study the energy gap between the most and n^{th} most stable ensemble members in a single, high diversity ensemble to seek sequences that are highly stabilizing on few members in an intra-ensemble negative design (Fig. 5.3b), or to perform negative MSD comparing a high-diversity ensemble to a low-diversity ensemble using the mathematical average fitness of the n lowest energy ensemble members rather than a Boltzmann-weighted average (Fig. 5.3c). We could then expect that while both rigid and dynamic sequences would score well on the low-diversity ensemble, dynamic sequence would score better than rigid sequences on a high-diversity ensemble due to the increased dispersiveness-sensitivity of a mathematical average over the Boltzmann-weighted average that instead heavily weights the most stable energies. Despite these three hypothetical methodologies approaching the problem at hand from different angles, they are all based on the concept of studying dispersiveness of energies in a backbone ensemble, or how many

unique members of an ensemble a sequence stabilizes, which we hypothesize to be a key metric in designing dynamics in RFPs. Whichever approach proves most successful, it remains that we are delving into hitherto unexplored territory in CPD, and significant optimization will be needed for any of these methods to become predictive. However, with our NMR data and soon mutagenesis results in hand, we are well placed to begin tackling this project.

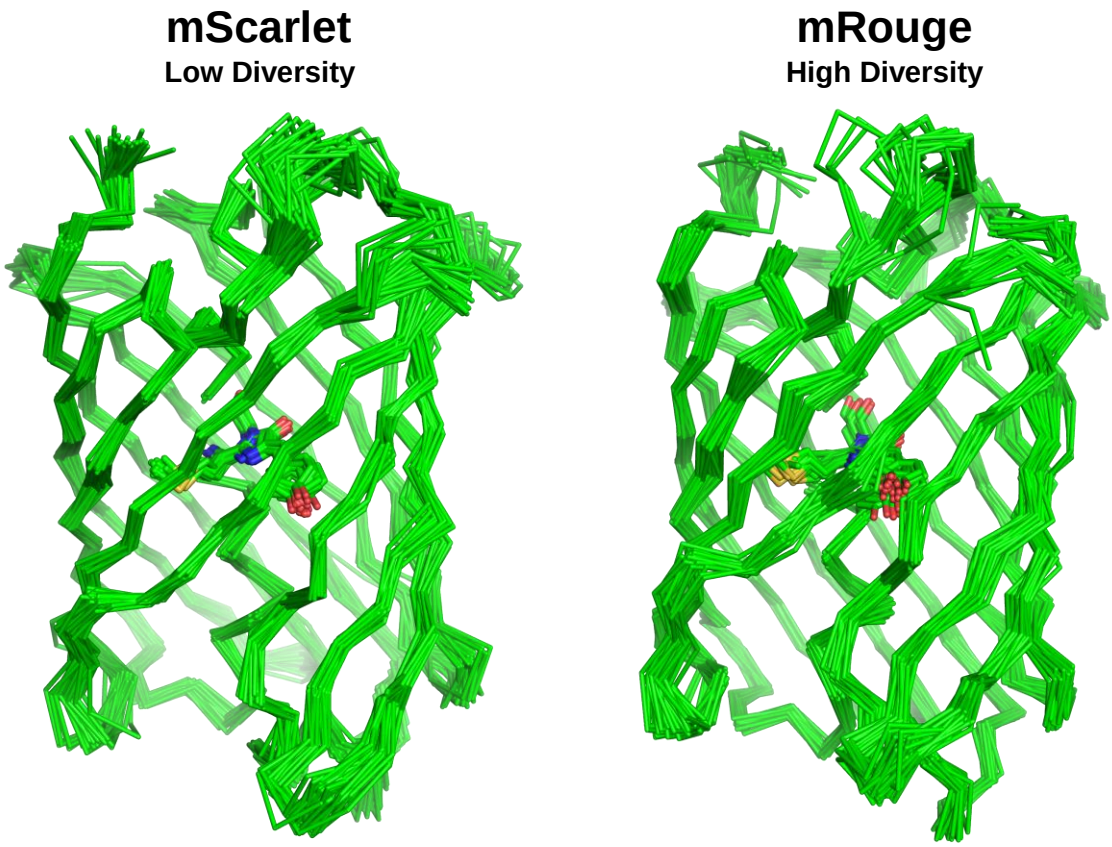


Figure 5.2. Representative high-diversity and low-diversity RFP MD ensembles. Ensembles of thirty representative conformers taken at regular intervals over a 1.5 microsecond Molecular Dynamics simulation are shown as ribbons, with the RFP chromophore shown as sticks. The ensemble generated from the mScarlet crystal structure is of visually-discernable lower diversity along the phenolate face, as well as β-strand 11 and the chromophore, in comparison to the mRouge ensemble. These RFP backbone ensembles thus serve as examples of high- and low- diversity ensembles for future designs.

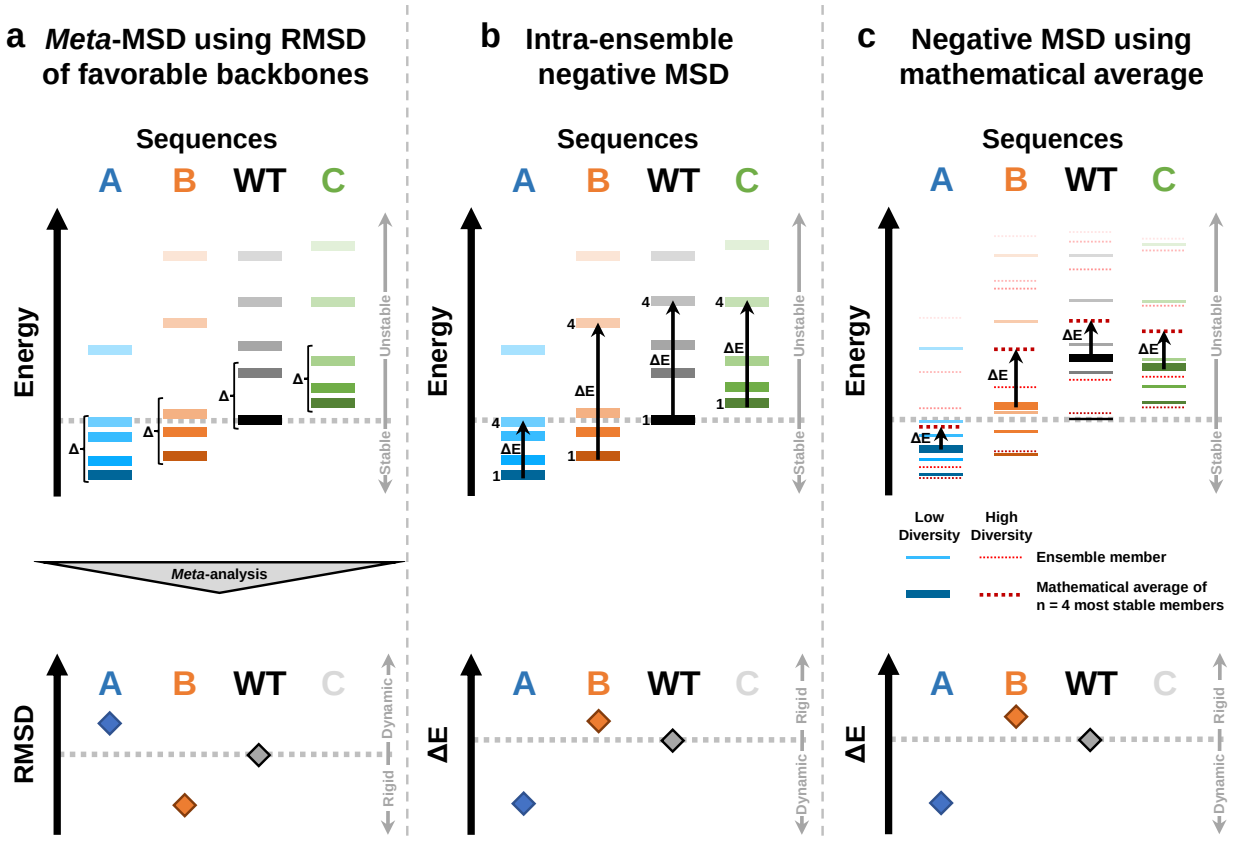


Figure 5.3. Hypothetical CPD strategies for the design of dynamics in RFPs. Three schemes for the design of dynamics in RFPs are presented as toy examples. **a.** In a *meta*-MSD strategy using the RMSD of favorable backbones as scoring criterion, a standard positive MSD approach is carried out resulting in a discrete energy being calculated for each sequence on each backbone, and sequences predicted to be less stable than the WT are discarded. Following energy calculations, in a *meta*-analysis, for each sequence, all backbones scoring within a certain ΔE, shown here as a bracket, are collected and their RMSD is calculated. For rigid sequences, which stabilize only a low-diversity set of backbones, a low RMSD is expected. **b.** In an intra-ensemble negative MSD strategy, a single backbone ensemble is used, and a ΔE between two members of the ensemble is used as score. Here, we show the ΔE between the most stable ensemble member and the arbitrarily chosen n = 4th most stable member for each sequence. We expect that rigid sequences would stabilize fewer backbones in a high-diversity ensemble than dynamic sequences, thus possess a greater ΔE for a properly optimized value of n. **c.** In a negative MSD strategy using mathematical averages, a standard negative MSD approach is carried out with the exception that the mathematical average of energies is used as a score rather than a Boltzmann-weighted average. As the mathematical average is more representative of dispersiveness in a dataset, when performing negative MSD with a low-diversity ensemble as positive ensemble and a high-diversity ensemble as negative ensemble, we expect a greater ΔE for rigid sequences that are capable of stabilizing fewer members of the high-diversity ensemble. However, to avoid artifacts of fixed backbones leading to non-physical high-energy scores, we expect that an average of only the n most stable members of the ensemble (shown here for an arbitrary n = 4) for each sequence will be more predictive than the average of energies over the entire ensemble.

5.2.2 Design of complex protein function

Beyond the immediate next steps for our forays into the computational design of protein dynamics, we anticipate applying a *meta*-MSD methodology to the design of complex protein functions such as catalysis or allostery. For catalysis, though several examples of CPD-based enzyme design exist,[97,238-240] our lab has worked on the redesign of specificity and equilibrium in multi-substrate enzymes.[117,241] Though an MSD approach comparing sequence scores on several backbone ensembles proved effective in modifying substrate specificity in an aminotransferase, the design focused on substrate binding rather than catalytic steps in the enzymatic reaction and thus led to a catalytic efficiency that was orders of magnitude lower than that of natural enzymes as is typical with enzymes designed using traditional CPD methodologies.[242,243] A more complete approach could consider not only substrate binding, but also intermediate states throughout the catalytic cycle if enough information about the mechanism is known to reconstruct them, much like how a putative intermediate was utilized in the development of Gβ1 DANCERs, as well as potential alternate conformations of the unbound enzyme, as many enzymes are highly dynamic molecules that can sample non-productive conformations.[18,244,245] For such a purpose, *meta*-MSD proves particularly useful due to its ability to consider a potentially unlimited number of target states and pick out sequences capable of stabilizing a desired conformational and catalytic pathway that might otherwise be lost in the context of certain conformational states when sequence optimization is carried out separately on each, while also excluding undesired trajectories. In this manner, we could in the future tackle the computational design of even more complex enzymes.

The design of allostery provides another potential application for *meta*-MSD, following a similar process to that used in the design of Gβ1 DANCERs. In DANCERs, we showed that Trp43 conformational exchange is predicated on the concerted motion of Phe34, which moves to occupy

the cavity vacated by the bulky tryptophan side-chain. In the case of the design of allostery, a similar concerted motion would be designed, though with the inclusion of a ligand-bound state in the design that alters the equilibrium of the motion. As we would again be designing a pathway for conformational exchange, with the added consideration of ligand binding, the number of conformational states to be considered and the need for specific energy differences between states to push conformational equilibrium upon ligand binding make *meta*-MSD particularly well suited to approach this problem.

5.3 Perspective on *meta*-MSD and CPD – Designing custom protein energy landscapes

Alongside efforts to develop new CPD methodologies to tackle new types of projects such as the design of dynamics, the CPD and biophysics communities have levied great effort to improve the predictiveness of both statistical and physics-based forcefields.[235,246-251] As *in silico* tools have become more and more faithful to the true behavior of protein, the predictiveness of associated CPD tools has also improved, leading to more reliable results. However, though inaccuracies still remain, especially in the consideration of long-range electrostatics, solvation, and solvent properties,[252-254] as evidenced by the often weak correlation between predicted and experimental stabilities and kinetic parameters,[115,184,255,256] CPD forcefields have become sufficiently predictive to allow for the precise design of *de novo* proteins, enzymes, and more.[97,105,240,257] That we were capable of designing DANCERs, despite the demonstrated subtlety of the interactions that were needed to allow Trp43 conformational exchange, and moreover that DANCER conformational exchange could also be recapitulated by MD simulations further suggests that computational forcefields have become optimized enough to accurately emulate the major determinants of protein energetics.

The success of *meta*-MSD and the design of DANCERs did not come from a novel energetics term, or indeed any modification to the PHOENIX forcefield used for these designs. Rather, the power of the *meta*-MSD methodology lies with the vast structure space searched, with the complete set of ensembles used totaling over 12,000 backbones, several orders of magnitude more than is typically used in an MSD approach.[54,109,115,116] Such an approach however does not come without its flaws. Due to limited computational resources, the current *meta*-MSD methodology more exhaustively searches local structure space for specific energetics patterns but is forced to conversely search a more limited sequence space than standard MSD due to restrictions on sequence optimization, given that each sequence to score must be tested on every *meta*-MSD node. At present, this highlights the tradeoff between search spaces, where a design evaluating energies on a larger set of ensembles will need to evaluate fewer sequences on the set of ensembles, as was the case with the design of DANCERs where only 1296 sequences were evaluated, this time several orders of magnitude lower than is typically used in CPD. Thus, the more ambitious the design and the more states that need considering, the more we will need to supplement our *meta*-MSD approach with empirical mutational data to aid us in selecting a search space that is likely to produce hits, much like how the design of DANCERs included only core mutations known to produce folded variants, albeit not in a combinatorial manner.[54] This quandary stands at odds with a push to also consider distant mutations in designs, which have been shown to have potentially large impacts on the function of many proteins.[258-260] Nonetheless, as our computational resources improve with new developments in computing technology, we will be able to enhance the scope of our designs in kind, and in the meantime, protein structural studies such as the NMR-based evaluation of RFP dynamics and brightness presented herein can aid us to refine our designs down to a feasible scope.

Our work, both in what is discussed in this thesis and through other projects, has shown that beyond the forcefield and specific design parameters we use, it is most critical to carefully choose states and ensembles that are physically representative of the pathways we are attempting to create. The *meta*-MSD approach outlined in this thesis has proven to be one more step down this trajectory, utilizing a large set of ensembles that models a rich and diverse segment of the local protein energy landscape we wished to reengineer. The development of *meta*-MSD and the need to model such an expansive set of conformational states as well as our characterization of DANCER dynamics however highlight the complexity and ruggedness of the protein energy landscape. Though this complexity indeed hinders the predictiveness of our computational tools, it is also the wellspring from which springs the incredible mutability and functional wealth that proteins offer. As CPD moves to tackle increasingly complex problems, we will in turn need to model larger swaths of this energy landscape to find those sequences that follow the specific trajectory we are seeking. And ultimately, as our understanding of the protein energy landscape deepens, so too will our ability to readily design such complex protein functions as allostery, catalysis, and more.

References

1 Agarwal, P. K. Enzymes: An integrated view of structure, dynamics and function. *Microb Cell Fact* **5**, doi:10.1186/1475-2859-5-2 (2006).

2 Smith, V. L., Kaetzel, M. A. & Dedman, J. R. Stimulus-response coupling: the search for intracellular calcium mediator proteins. *Cell Regul* **1**, 165-172 (1990).

3 Nishizuka, Y. Perspectives on the Role of Protein Kinase C in Stimulus-Response Coupling. *J Natl Cancer Inst* **76**, 363-370, doi:10.1093/jnci/76.3.363 (1986).

4 Meyer, A. J., Almendrala, D. K., Go, M. M. & Krauss, S. W. Structural protein 4.1R is integrally involved in nuclear envelope protein localization, centrosome–nucleus association and transcriptional signaling. *J Cell Sci* **124**, 1433-1444, doi:10.1242/jcs.077883 (2011).

5 Fosket, D. E. & Morejohn, L. C. Structural and functional organization of tubulin. *Annu Rev Plant Physiol Plant Mol Biol* **43**, 201-240 (1992).

6 Quizon, P. M. *et al.* Molecular mechanism: the human dopamine transporter histidine 547 regulates basal and HIV-1 Tat protein-inhibited dopamine transport. *Sci Rep* **6**, Article number: 39048, doi:10.1038/srep39048 (2016).

7 Horlacher, R. *et al.* Archaeal Binding Protein-Dependent ABC Transporter: Molecular and Biochemical Analysis of the Trehalose/Maltose Transport System of the Hyperthermophilic Archaeon Thermococcus litoralis. *J Bacteriol* **180**, 680-689 (1998).

8 Orengo, C. A., Todd, A. E. & Thornton, J. M. From protein structure to function. *Curr Opin Struct Biol* **9**, 374-382 (1999).

9 Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. From structure to function: approaches and limitations. *Nat Struct Biol* **7**, 991-994 (2000).

10 Ouzonis, C. A., Coulson, R. M., Enright, A. J., Kunin, V. & Pereira-Leal, J. B. Classification schemes for protein structure and function. *Nat Rev Genet* **4**, 508-519 (2003).

11 Lee, D., Redfern, O. & Orengo, C. A. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995-1005 (2007).

12 Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**, 147-164, doi:10.1006/jmbi.1999.2661 (1999).

13 Kinoshita, K. & Nakamura, H. Protein informatics towards function identification. *Curr Opin Struct Biol* **13**, 396-400 (2003).

14 Hvidsten, T. R. *et al.* A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity. *PLoS One* **4**, e6266, doi:10.1371/journal.pone.0006266 (2009).

15 Orozco, M. A theoretical view of protein dynamics. *Chem Soc Rev* **43**, 5051, doi:10.1039/c3cs60474h (2014).

16 Tokuriki, N. & S., T. D. Protein dynamism and evolvability. *Science* **324**, 203-207 (2009).

17 Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638-1642 (2006).

18 Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol* **12**, 944-950, doi:10.1038/nchembio.2175 (2016).

19 Janežič, D., Venable, R. M. & Brooks, B. R. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *J Comput Chem* **16**, 1554-1566, doi:10.1002/jcc.540161211 (1995).

20 Li, C., Tang, C. & Liu, M. Protein dynamics elucidated by NMR technique. *Protein Cell* **4**, 726-730, doi:10.1007/s13238-013-3912-1 (2013).

21 Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964-972, doi:10.1038/nature06522 (2007).
22 Ma, J. & Karplus, M. The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. *Proc Natl Acad Sci U S A* **95**, 8502-8507, doi:10.1073/pnas.95.15.8502 (1998).
23 Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* **102**, 6679-6685, doi:10.1073/pnas.0408930102 (2005).
24 Wolf-Watz, M. *et al.* Linkage Between Dynamics and Catalysis in a Thermophilic-Mesophilic Enzyme Pair. *Nat Struct Mol Biol* **11**, 945-949 (2004).
25 Guo, J. & Zhou, H.-X. Protein Allostery and Conformational Dynamics. *Chem Rev* **116**, 6503-6515, doi:10.1021/acs.chemrev.5b00590 (2016).
26 Dunker, A. K., Silman, I., Uversky, V. N. & Sussman, J. L. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18**, 756-764, doi:10.1016/j.sbi.2008.10.002 (2008).
27 Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**, 161-171 (2000).
28 Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197-208 (2005).
29 Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**, 573-584 (2002).
30 Islam, M. M. *et al.* Binding of C5-dicarboxylic substrate to aspartate aminotransferase: implications for the conformational change at the transaldimination step. *Biochemistry* **44**, 8218-8229, doi:10.1021/bi050071g (2005).
31 Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838-U813, doi:10.1038/nature06410 (2007).
32 Kerns, S. J. *et al.* The energy landscape of adenylate kinase during catalysis. *Nat Struct Mol Biol* **22**, 124-131, doi:10.1038/nsmb.2941 (2015).
33 D'rozario, R. S. G. & Sansom, M. S. P. Helix dynamics in a membrane transport protein: comparative simulations of the glycerol-3-phosphate transporter and its constituent helices. *Mol Membr Biol* **25**, 571-583, doi:10.1080/09687680802549113 (2009).
34 Pan, Y., Piyadasa, H., O'Neil, J. D. & Konermann, L. Conformational dynamics of a membrane transport protein probed by H/D exchange and covalent labeling: the glycerol facilitator. *J Mol Biol* **416**, 400-413, doi:10.1016/j.jmb.2011.12.052 (2012).
35 Campbell, E. C. *et al.* Laboratory evolution of protein conformational dynamics. *Curr Opin Struct Biol* **50**, 49-57 (2018).
36 Muller, C. W. & Schulz, G. E. Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state. *J Mol Biol* **224**, 159-177 (1992).
37 Muller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147-156 (1996).
38 Okamoto, A., Higuchi, T., Hirotsu, K., Kuramitsu, S. & Kagamiyama, H. X-ray crystallographic study of pyridoxal 5'-phosphate-type aspartate aminotransferases from Escherichia coli in open and closed form. *J Biochem* **116**, 95-107 (1994).
39 Heo, L. & Feig, M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proc Natl Acad Sci U S A* **115**, 13276-13281, doi:10.1073/pnas.1811364115 (2018).
40 Piana, S., Klepeis, J. L. & Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol* **24**, 98-105, doi:10.1016/j.sbi.2013.12.006 (2014).

41 Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **115**, E4758-4766, doi:10.1073/pnas.1800690115 (2018).
42 Marcos, E. & Silva, D.-A. Essentials of de novo protein design: Methods and applications. *WIREs Comput Mol Sci* **8**, doi:10.1002/wcms.1374 (2018).
43 Bhardwaj, G. *et al.* Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329-335, doi:10.1038/nature19791 (2016).
44 Sjöbring, U., Björk, L. & Kastern, W. Streptococcal Protein G Gene Structure and Protein Binding Properties. *J Biol Chem* **266**, 399-405 (1991).
45 Björk, L. & Kronvall, G. Purification and some properties of streptococcal protein G, a novel IgG-binding reagent. *J Immunol* **133**, 969-974 (1984).
46 Fahnestock, S. R., Alexander, P., Nagle, J. & Filpula, D. Gene for an Immunoglobulin-Binding Protein from a Group G Streptococcus. *J Bacteriol* **167**, 870-880 (1986).
47 Negron, C. & Keating, A. Multistate protein design using CLEVER and CLASSY. *Methods Enzymol* **531**, 171-190, doi:10.1016/B978-0-12-394292-0.00008-4 (2013).
48 Gronenborn, A. M. *et al.* A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657-661 (1991).
49 Creighton, T. E. Disulfide bonds as probes of protein folding pathways. *Methods Enzymol* **131**, 83-106 (1986).
50 Cheng, Y. & Patel, D. J. An efficient system for small protein expression and refolding. *Biochem Biophys Res Commun* **317**, 401-405, doi:10.1016/j.bbrc.2004.03.068 (2004).
51 Zhou, P., Lugovskoy, A. A. & Wagner, G. A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. *J Biomol NMR* **20**, 11-14 (1997).
52 Gronenborn, A. M. & Clore, G. M. Rapid screening for structural integrity of expressed proteins by heteronuclear NMR spectroscopy. *Protein Sci* **5**, 174-177 (1996).
53 Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721-4729 (1994).
54 Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci U S A* **107**, 19838-19843, doi:10.1073/pnas.1012985107 (2010).
55 Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**, 10172-10177 (1997).
56 Jee, J., Byeon, I. J., Louis, J. M. & Gronenborn, A. M. The point mutation A34F causes dimerization of GB1. *Proteins* **71**, 1420-1431, doi:10.1002/prot.21831 (2008).
57 Byeon, I. J., Louis, J. M. & Gronenborn, A. M. A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. *J Mol Biol* **333**, 141-152 (2003).
58 Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A* **104**, 11963-11968, doi:10.1073/pnas.0700922104 (2007).
59 Matz, M. V. *et al.* Fluorescent proteins from nonbioluminescent Anthozoa species. *Nat Biotechnol* **17**, 969-973, doi:10.1038/13657 (1999).
60 Yarbrough, D., Wachter, R. M., Kallio, K., Matz, M. V. & Remington, S. J. Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-A resolution. *Proc Natl Acad Sci U S A* **98**, 462-467, doi:10.1073/pnas.98.2.462 (2001).
61 Merzlyak, E. M. *et al.* Bright monomeric red fluorescent protein with an extended fluorescence lifetime. *Nat Methods* **4**, 555-557, doi:10.1038/nmeth1062 (2007).

62 Chudakov, D. M., Matz, M. V., Lukyanov, S. & Lukyanov, K. A. Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol Rev* **90**, 1103-1163, doi:10.1152/physrev.00038.2009 (2010).
63 Nienhaus, K. & Nienhaus, G. U. Fluorescent proteins for live-cell imaging with super-resolution. *Chem Soc Rev* **43**, 1088-1106, doi:10.1039/c3cs60171d (2014).
64 Fradkov, A. *et al.* Far-red fluorescent tag for protein labelling. *Biochem J* **368**, 17-21, doi:10.1042/BJ20021191 (2002).
65 Campbell, R. E. *et al.* A monomeric red fluorescent protein. *Proc Natl Acad Sci U S A* **99**, 7877-7882, doi:10.1073/pnas.082243699 (2002).
66 Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotechnol* **22**, 1567-1572, doi:10.1038/nbt1037 (2004).
67 Moore, M. M., Oteng-Pabi, S. K., Pandelieva, A. T., Mayo, S. L. & Chica, R. A. Recovery of red fluorescent protein chromophore maturation deficiency through rational design. *PLoS One* **7**, e52463, doi:10.1371/journal.pone.0052463 (2012).
68 Strack, R. L., Strongin, D. E., Mets, L., Glick, B. S. & Keenan, R. J. Chromophore formation in DsRed occurs by a branched pathway. *J Am Chem Soc* **132**, 8496-8505, doi:10.1021/ja1030084 (2010).
69 Verkhusha, V. V., Chudakov, D. M., Gurskaya, N. G., Lukyanov, S. & Lukyanov, K. A. Common pathway for the red chromophore formation in fluorescent proteins and chromoproteins. *Chem Biol* **11**, 845-854, doi:10.1016/j.chembiol.2004.04.007 (2004).
70 Shu, X., Shaner, N. C., Yarbrough, C. A., Tsien, R. Y. & Remington, S. J. Novel chromophores and buried charges control color in mFruits. *Biochemistry* **45**, 9639-9647, doi:10.1021/bi060773l (2006).
71 Young, C. L., Raden, D. L., Caplan, J. L., Czymmek, K. J. & Robinson, A. S. Cassette series designed for live-cell imaging of proteins and high-resolution techniques in yeast. *Yeast* **29**, 119-136, doi:10.1002/yea.2895 (2012).
72 Lam, A. J. *et al.* Improving FRET dynamic range with bright green and red fluorescent proteins. *Nat Methods* **9**, 1005-1012, doi:10.1038/nmeth.2171 (2012).
73 Hochreiter, B., Pardo-Garcia, A. & Schmid, J. A. Fluorescent Proteins as Genetically Encoded FRET Biosensors in Life Sciences. *Sensors* **15**, 26281-26314 (2015).
74 Zhao, Y. *et al.* An expanded palette of genetically encoded Ca(2)(+) indicators. *Science* **333**, 1888-1891, doi:10.1126/science.1208592 (2011).
75 Tantama, M., Hung, Y. P. & Yellen, G. Imaging intracellular pH in live cells with a genetically encoded red fluorescent protein sensor. *J Am Chem Soc* **133**, 10034-10037, doi:10.1021/ja202902d (2011).
76 Wu, J. *et al.* Improved orange and red Ca(2)+/- indicators and photophysical considerations for optogenetic applications. *ACS Chem Neurosci* **4**, 963-972, doi:10.1021/cn400012b (2013).
77 Shcherbo, D. *et al.* Bright far-red fluorescent protein for whole-body imaging. *Nat Methods* **4**, 741-746, doi:10.1038/nmeth1083 (2007).
78 Parker, C. A. & Rees, W. T. Correction of fluorescence spectra and measurement of fluorescence quantum efficiency. *Analyst* **85**, 587-600, doi:10.1039/an9608500587 (1960).
79 Bowen, E. J. Fluorescence quenching in solution and in the vapour state. *J Chem Soc Faraday Trans* **50**, 97-102, doi:10.1039/tf9545000097 (1954).
80 Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33-38, doi:10.1016/0378-1119(95)00685-0 (1996).
81 Glick, B. S., Strongin, D. E., Keenan, R., Strack, R. L. & Bhattacharyya, D. (Google Patents, 2014).

82 Pandelieva, A. T. *et al.* Brighter Red Fluorescent Proteins by Rational Design of Triple-Decker Motif. *ACS Chem Biol* **11**, 508-517, doi:10.1021/acschembio.5b00774 (2016).
83 Bindels, D. S. *et al.* mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nat Methods*, doi:10.1038/nmeth.4074 (2016).
84 Lambert, T. tlambert03/FPbase. *Zenodo*, doi:10.5281/zenodo.1244328 (2018).
85 Haebig, J. E. Vibrational effects in radiationless transitions in aromatic molecules. *J Phys Chem* **71**, 4203-4209, doi:10.1021/j100872a007 (1967).
86 Henderson, J. N. & Remington, S. J. The kindling fluorescent protein: A transient photoswitchable marker. *Physiology* **21**, 162-170, doi:10.1152/physiol.00056.2005 (2006).
87 Pletnev, S. *et al.* A crystallographic study of bright far-red fluorescent protein mKate reveals pH-induced cis-trans isomerization of the chromophore. *J Biol Chem* **283**, 28980-28987, doi:10.1074/jbc.M800599200 (2008).
88 Laurent, A. D., Mironov, V. A., Chapagain, P. P., Nemukhin, A. V. & Krylov, A. I. Exploring structural and optical properties of fluorescent proteins by squeezing: modeling high-pressure effects on the mStrawberry and mCherry red fluorescent proteins. *J Phys Chem B* **116**, 12426-12440, doi:10.1021/jp3060944 (2012).
89 Goedhart, J. *et al.* Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%. *Nat Commun* **3**, 751, doi:10.1038/ncomms1738 (2012).
90 Helms, V., Straatsma, T. P. & McCammon, J. A. Internal dynamics of green fluorescent protein. *J Phys Chem B* **103**, 3263-3269, doi:10.1021/jp983120q (1999).
91 Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* **5**, 470-475 (1998).
92 Altman, M. D., Nalivaika, E. A., Prabu-Jeyabalan, M., Schiffer, C. A. & Tidor, B. Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. *Proteins* **70**, 678-694, doi:10.1002/prot.21514 (2008).
93 Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656-659, doi:10.1038/nature04818 (2006).
94 Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **268**, 209-225 (1997).
95 Simons, K. T. *et al.* Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct Funct Bioinf* **34**, 82-95 (1999).
96 Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93 (2004).
97 Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc Natl Acad Sci U S A* **109**, 3790-3795, doi:10.1073/pnas.1118082108 (2012).
98 Mayo, S. L., Olafson, B. D. & Goddard, W. A. Dreiding - a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **94**, 8897-8909, doi:Doi 10.1021/J100389a010 (1990).
99 Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins-Structure Function and Genetics* **35**, 133-152, doi:Doi 10.1002/(Sici)1097-0134(19990501)35:2<133::Aid-Prot1>3.0.Co;2-N (1999).
100 Street, A. G. & Mayo, S. L. Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design* **3**, 253-258, doi:Doi 10.1016/S1359-0278(98)00036-4 (1998).
101 Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542 (1992).

102 Desmet, J., Spriet, J. & Lasters, I. Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. *Proteins-Structure Function and Genetics* **48**, 31-43, doi:Doi 10.1002/Prot.10131 (2002).
103 Pierce, N. A. & Winfree, E. Protein Design is NP-hard. *Protein Eng Des Sel* **15**, 779-782 (2002).
104 Voigt, C. A., Gordon, D. B. & Mayo, S. L. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**, 789-803 (2000).
105 Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368, doi:10.1126/science.1089427 (2003).
106 Shah, P. S. *et al.* Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol* **372**, 1-6, doi:10.1016/j.jmb.2007.06.032 (2007).
107 Nauli, S., Kuhlman, B. & Baker, D. Computer-based redesign of a protein folding pathway. *Nat Struct Biol* **8**, 602-605 (2001).
108 Choi, E. J., Guntas, G. & Kuhlman, B. in *Protein Engineering and Design* (eds S. J. Park & J. R. Cochran) (CRC Press, 2009).
109 Davey, J. A. & Chica, R. A. Multistate approaches in computational protein design. *Protein Sci* **21**, 1241-1252, doi:10.1002/pro.2128 (2012).
110 Murphy, G. S. *et al.* Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure* **20**, 1086-1096, doi:10.1016/j.str.2012.03.026 (2012).
111 Smith, C. A. & Kortemme, T. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One* **6**, e20451, doi:10.1371/journal.pone.0020451 (2011).
112 Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proc Natl Acad Sci U S A* **109**, 1494-1499, doi:10.1073/pnas.1115172109 (2012).
113 Gainza, P., Roberts, K. E. & Donald, B. R. Protein design using continuous rotamers. *PLoS Comput Biol* **8**, e1002335, doi:10.1371/journal.pcbi.1002335 (2012).
114 Lanouette, S. *et al.* Discovery of substrates for a SET domain lysine methyltransferase predicted by multistate computational protein design. *Structure* **23**, 206-215, doi:10.1016/j.str.2014.11.004 (2015).
115 Davey, J. A., Damry, A. M., Euler, C. K., Goto, N. K. & Chica, R. A. Prediction of Stable Globular Proteins Using Negative Design with Non-native Backbone Ensembles. *Structure* **23**, 2011-2021, doi:10.1016/j.str.2015.07.021 (2015).
116 Davey, J. A. & Chica, R. A. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins* **82**, 771-784, doi:10.1002/prot.24457 (2014).
117 St-Jacques, A. D. *Engineering of Multi-Substrate Enzyme Specificity and Conformational Equilibrium Using Multistate Computational Protein Design* Ph. D. thesis, University of Ottawa, (2018).
118 Keeler, J. *Understanding NMR Spectroscopy*. 2 edn, (John Wiley & Sons, 2011).
119 Levitt, M. H. *Spin Dynamics*. (John Wiley & Sons, 2001).
120 Bloch, F. Nuclear Induction. *Phys Rev* **70**, 460-474 (1946).
121 Abragam, A. *Principles of Nuclear Magnetism*. (Oxford University Press, 1961).
122 Kay, L. E., Keifer, P. & Saarinen, T. Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J Am Chem Soc* **114**, 10663-10665 (1992).
123 Marion, D., Tarbouriech, N., Ruigrok, R. W., Burmeister, W. P. & Blanchard, L. Letter to the Editor: Assignment of the 1H, 15N and 13C resonances of the nucleocapsid-binding domain of the Sendai virus Phosphoprotein. *J Biomol NMR* **21**, 75-76 (2001).

124 Kleckner, I. R. & Foster, M. P. An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta* **1814**, 942-968, doi:10.1016/j.bbapap.2010.10.012 (2011).
125 Mittermaier, A. K. & Kay, L. E. Observing biological dynamics at atomic resolution using NMR. *Trends Biochem Sci* **34**, 601-611, doi:10.1016/j.tibs.2009.07.004 (2009).
126 Wittekind, M. & Mueller, L. HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. *J Magn Reson* **101**, 201-205, doi:10.1006/jmrb.1993.1033 (1993).
127 Grzesiek, S. & Bax, A. Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J Biomol NMR* **3**, 185-204 (1993).
128 Muhandiram, D. R. & Kay, L. E. Gradient-Enhanced Triple-Resonance Three-Dimensional NMR Experiments with Improved Sensitivity. *J Magn Reson* **103**, 203-216, doi:10.1006/jmrb.1994.1032 (1994).
129 Kay, L. E., Xu, G. Y., Singer, A. U., Muhandiram, D. R. & Forman-Kay, J. D. A Gradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain 1H and 13C Correlations in H 2O Samples of Proteins. *J Magn Reson* **101**, 333-337, doi:10.1006/jmrb.1993.1053 (1993).
130 Davis, A. L., Keeler, J., Laue, E. D. & Moskau, D. Experiments for recording pure-absorption heteronuclear correlation spectra using pulsed field gradients. *J Magn Reson* **98**, 207-216, doi:10.1016/0022-2364(92)90126-R (1992).
131 Markwick, P. R. L., Malliavin, T. & Nilges, M. Structural Biology by NMR: Structure, Dynamics, and Interactions. *PloS Comp Biol*, doi:10.1371/journal.pcbi.1000168 (2008).
132 Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **44**, 213-223, doi:10.1007/s10858-009-9333-z (2009).
133 Berjanskii, M. V., Neal, S. & Wishart, D. S. PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res* **1**, W63-69 (2006).
134 Wishart, D., Sykes, B. & Richards, F. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647-1651 (1992).
135 Güntert, P. in *Protein NMR Techniques* (ed A. Kristina Downing) 353-378 (Humana Press, 2004).
136 Marion, D. An Introduction to Biological NMR Spectroscopy. *Mol Cell Proteomics* **12**, 3006-3025, doi:10.1074/mcp.O113.030239 (2013).
137 Farrow, N. A., Zhang, O., Forman-Kay, J. D. & Kay, L. E. A heteronuclear correlation experiment for simultaneous determination of 15N longitudinal decay and chemical exchange rates of systems in slow equilibrium. *Journal of Biomolecular NMR* **4**, 727-734, doi:10.1007/bf00404280 (1994).
138 Tollinger, M., Skrynnikov, N. R., Mulder, F. A., Forman-Kay, J. D. & Kay, L. E. Slow dynamics in folded and unfolded states of an SH3 domain. *J Am Chem Soc* **123**, 11341-11352 (2001).
139 Ishima, R. in *Protein Dynamics: Methods in Molecular Biology* Vol. 1084 (ed D. Livesay) (Humana Press, 2014).
140 Waudby, C. A., Ramos, A., Cabrita, L. D. & Christodoulou, J. Two-Dimensional NMR Lineshape Analysis. *Sci Rep* **6**, doi:10.1038/srep24826 (2016).
141 Baxter, N. J. & Williamson, M. P. Temperature dependence of 1H chemical shifts in proteins. *J Biomol NMR* **9**, 359-369 (1997).
142 Mittermaier, A. K. & Kay, L. E. New Tools Provide New Insights in NMR Studies of Protein Dynamics. *Science* **312**, 224-228, doi:10.1126/science.1124964 (2006).

143 Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28**, 8972-8979 (1989).
144 Kelly, S. M., Jess, T. J. & Price, N. C. How to study proteins by circular dichroism. *Biochim Biophys Acta* **1751**, 119-139 (2005).
145 Ptitsyn, O. B. Molten globule and protein folding. *Adv Prot Chem* **47**, 83-229 (1995).
146 Price, N. E., Price, N. C., Kelly, S. M. & McDonnell, J. M. The key role ofprotein flexibility in modulating IgE interactions. *J Biol Chem* **280**, 2324-2330 (2005).
147 Goodman, M., Verdini, A. S., Toniolo, C., Phillips, W. D. & Bovey, F. A. Sensitive Criteria for the Critical Size for Helix Formation in Oligopeptides. *Proc Natl Acad Sci U S A* **64**, 444-450 (1969).
148 Whitmore, L. & Wallace, B. A. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers* **89**, 392-400 (2008).
149 Woody, R. W. & Sreerama, N. On the analysis of membrane protein circular dichroism spectra. *Protein Sci* **13**, 100-112 (2004).
150 Sreerama, N. & Woody, R. W. A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem* **209**, 32-44 (1993).
151 Manavalan, P. & Johnson, W. C. J. Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal Biochem* **167**, 76-85 (1987).
152 Provencher, S. W. & Glöckner, J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33-37 (1981).
153 Krittanai, C. & Johnson, W. C. J. Correcting the circular dichroismspectra of peptides for contributions of absorbing side chains. *Anal Biochem* **253**, 57-64 (1997).
154 Sreerama, N. & Woody, R. W. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem* **287**, 252-260 (2000).
155 van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grondelle, R. & Groen, F. C. A. Estimation of Protein Secondary Structure and Error Analysis from Circular Dichroism Spectra *Anal Biochem* **191**, 110-118 (1990).
156 Boxer, D. H. *et al.* Sensing of remote oxyanion binding at the DNA binding domain of the molybdate-dependent transcriptional regulator, ModE. *Org Biomol Chem* **2**, 2829-2837 (2004).
157 Hope, J. *et al.* Cytotoxicity of prion protein peptide (PrP106-126) differs in mechanism from the cytotoxic activity of the Alzheimer's disease amyloid peptide, Ab 25-35. *Neuredegeneration* **5**, 1-11 (1996).
158 Pandya, M. J. *et al.* Sequence and structural duality: designing peptides to adopt two stable conformation. *J Am Chem Soc* **126**, 17016-17024 (2004).
159 Kelly, S. M. & Price, N. C. The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Curr Protein Pep Sci* **1**, 349-384 (2000).
160 Greenfield, N. J. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat Protoc* **1**, 2527-2535, doi:10.1038/nprot.2006.204 (2006).
161 Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82-87 (1997).
162 Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222-227, doi:10.1038/nature11600 (2012).
163 Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved beta sheets. *Science* **355**, 201-206, doi:10.1126/science.aah7389 (2017).

164 Ambroggio, X. I. & Kuhlman, B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **128**, 1154-1161, doi:10.1021/ja054718w (2006).
165 Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391, doi:10.1126/science.1152692 (2008).
166 Bhabha, G. *et al.* A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **332**, 234-238, doi:10.1126/science.1198542 (2011).
167 Tzeng, S. R. & Kalodimos, C. G. Dynamic activation of an allosteric regulatory protein. *Nature* **462**, 368-372, doi:10.1038/nature08560 (2009).
168 Tuinstra, R. L. *et al.* Interconversion between two unrelated protein folds in the lymphotactin native state. *Proceedings of the National Academy of Sciences* **105**, 5057-5062, doi:10.1073/pnas.0709518105 (2008).
169 Crowhurst, K. A. & Mayo, S. L. NMR-detected conformational exchange observed in a computationally designed variant of protein Gbeta1. *Protein Eng Des Sel* **21**, 577-587, doi:10.1093/protein/gzn035 (2008).
170 Bouvignies, G. *et al.* Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13885-13890, doi:10.1073/pnas.0505129102 (2005).
171 Derrick, J. P. & Wigley, D. B. 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. *J Mol Biol* **243**, 906-918, doi:10.1006/jmbi.1994.1691 (1994).
172 Wylie, B. J. *et al.* Ultrahigh resolution protein structures using NMR chemical shift tensors. *Proc Natl Acad Sci U S A* **108**, 16974-16979, doi:10.1073/pnas.1103728108 (2011).
173 Tomlinson, J. H., Green, V. L., Baker, P. J. & Williamson, M. P. Structural origins of pH-dependent chemical shifts in the B1 domain of protein G. *Proteins* **78**, 3000-3016, doi:10.1002/prot.22825 (2010).
174 Wilton, D. J., Tunnicliffe, R. B., Kamatari, Y. O., Akasaka, K. & Williamson, M. P. Pressure-induced changes in the solution structure of the GB1 domain of protein G. *Proteins* **71**, 1432-1440, doi:10.1002/prot.21832 (2008).
175 Strop, P., Marinescu, A. M. & Mayo, S. L. Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci* **9**, 1391-1394, doi:10.1110/ps.9.7.1391 (2000).
176 Saio, T., Ogura, K., Yokochi, M., Kobashigawa, Y. & Inagaki, F. Two-point anchoring of a lanthanide-binding peptide to a target protein enhances the paramagnetic anisotropic effect. *J Biomol NMR* **44**, 157-166, doi:10.1007/s10858-009-9325-z (2009).
177 Jee, J., Ishima, R. & Gronenborn, A. M. Characterization of specific protein association by 15N CPMG relaxation dispersion NMR: the GB1(A34F) monomer-dimer equilibrium. *J Phys Chem B* **112**, 6008-6012, doi:10.1021/jp076094h (2008).
178 Kuszewski, J., Gronenborn, A. M. & Clore, G. M. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. *Journal of the American Chemical Society* **121**, 2337-2338 (1999).
179 Wei, G. H., Xi, W. H., Nussinov, R. & Ma, B. Y. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews* **116**, 6516-6551, doi:10.1021/acs.chemrev.5b00562 (2016).
180 Davey, J. A. & Chica, R. A. Optimization of rotamers prior to template minimization improves stability predictions made by computational protein design. *Protein Sci* **24**, 545-560, doi:10.1002/pro.2618 (2015).

181 Davey, J. A. & Chica, R. A. Multistate Computational Protein Design with Backbone Ensembles. *Methods Mol Biol* **1529**, 161-179, doi:10.1007/978-1-4939-6637-0_7 (2017).
182 Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **4**, 2138-2148, doi:10.1002/pro.5560041020 (1995).
183 Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710, doi:10.1002/prot.20264 (2004).
184 Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838, doi:10.1002/prot.22921 (2011).
185 Reeve, S. M. *et al.* Protein design algorithms predict viable resistance to an experimental antifolate. *Proc Natl Acad Sci U S A* **112**, 749-754, doi:10.1073/pnas.1411548112 (2015).
186 Roberts, K. E., Cushing, P. R., Boisguerin, P., Madden, D. R. & Donald, B. R. Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. *PLoS Comput Biol* **8**, e1002477, doi:10.1371/journal.pcbi.1002477 (2012).
187 Bouvignies, G. *et al.* Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **477**, 111-U134, doi:10.1038/nature10349 (2011).
188 Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins-Structure Function and Bioinformatics* **75**, 187-205, doi:Doi 10.1002/Prot.22234 (2009).
189 Davis, I. W., Arendall, W. B., 3rd, Richardson, D. C. & Richardson, J. S. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* **14**, 265-274, doi:10.1016/j.str.2005.10.007 (2006).
190 Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L. & Kortemme, T. RosettaBackrub-a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Research* **38**, W569-W575, doi:Doi 10.1093/Nar/Gkq369 (2010).
191 Nash, S. G. A survey of truncated-Newton methods. *Journal of Computational and Applied Mathematics* **124**, 45-59, doi:Doi 10.1016/S0377-0427(00)00426-X (2000).
192 Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049-1074, doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F (2000).
193 Chica, R. A., Moore, M. M., Allen, B. D. & Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc Natl Acad Sci U S A* **107**, 20257-20262, doi:10.1073/pnas.1013910107 (2010).
194 Allen, B. D. & Mayo, S. L. Dramatic performance enhancements for the FASTER optimization algorithm. *J Comput Chem* **27**, 1071-1075, doi:10.1002/jcc.20420 (2006).
195 Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J Comput Chem* **31**, 904-916, doi:10.1002/jcc.21375 (2010).
196 Dunbrack, R. L. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* **6**, 1661-1681 (1997).
197 Lazaridis, T. & Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *Journal of Molecular Biology* **288**, 477-487, doi:DOI 10.1006/jmbi.1999.2685 (1999).
198 Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel beta-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Science : A Publication of the Protein Society* **8**, 841-853 (1999).

199 Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**, 277-293, doi:10.1007/bf00197809 (1995).
200 Johnson, B. & Blevins, R. NMR View: A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR* **4**, 603-614, doi:10.1007/BF00404272 (1994).
201 Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375-383, doi:10.1093/nar/gkm216 (2007).
202 Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure, Function, and Bioinformatics* **66**, 778-795, doi:10.1002/prot.21165 (2007).
203 Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): Structure Quality Assessment Measures Based on Information Retrieval Statistics. *Journal of the American Chemical Society* **127**, 1665-1674, doi:10.1021/ja047109h (2005).
204 Grant, B. J., Gorfe, A. A. & McCammon, J. A. Large conformational changes in proteins: signalling and other functions. *Curr Opin Struct Biol* **20**, 142-147 (2010).
205 Bhabha, G. *et al.* Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat Struct Mol Biol* **20**, 1243-1249 (2013).
206 Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669-673, doi:10.1038/nature08615 (2009).
207 Narayanan, C. *et al.* Conservation of Dynamics Associated with Biological Function in an Enzyme Superfamily. *Structure* **26**, 426-436, doi:10.1016/j.str.2018.01.015 (2018).
208 Hensen, U. *et al.* Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS One* **7**, e33931 (2012).
209 Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci* **25**, 1260-1272, doi:10.1002/pro.2876 (2016).
210 Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat Chem* **9**, 50-56 (2017).
211 Risso, V. A. *et al. De novo* active sites for resurrected Precambrian enzymes. *Nat Commun* **8** (2017).
212 Clifton, B. E. & Jackson, C. J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem Biol* **23**, 236-245 (2016).
213 Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A. Rational design of proteins that exchange on functional timescales. *Nat Chem Biol* **13**, 1280-1285, doi:10.1038/nchembio.2503 (2017).
214 Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscape and motions of proteins. *Science* **254**, 1598-1603, doi:10.1126/science.1749933 (1991).
215 Loewe, L. & Hill, W. G. The population genetics of mutations: good, bad and indifferent. *Philos Trans R Soc Lond B Biol Sci* **365**, 1153-1167 (2010).
216 Konig, K. Multiphoton microscopy in life sciences. *J Microsc* **200**, 83-104 (2000).
217 Chu, J. *et al.* Non-invasive intravital imaging of cellular differentiation with a bright red-excitable fluorescent protein. *Nat Methods* **11**, 572-578, doi:10.1038/nmeth.2888 (2014).
218 Bajar, B. T. *et al.* Improving brightness and photostability of green and red fluorescent proteins for live cell imaging and FRET reporting. *Sci Rep* **6**, 20889, doi:10.1038/srep20889 (2016).
219 Fabritius, A. *et al.* Imaging-Based Screening Platform Assists Protein Engineering. *Cell Chem Biol* **25**, 1554-1561, doi:10.1016/j.chembiol.2018.08.008 (2018).
220 Kredel, S. *et al.* mRuby, a bright monomeric red fluorescent protein for labeling of subcellular structures. *PLoS One* **4**, e4391, doi:10.1371/journal.pone.0004391 (2009).
221 Shemiakina, II *et al.* A monomeric red fluorescent protein with low cytotoxicity. *Nat Commun* **3**, 1204, doi:10.1038/ncomms2208 (2012).

222 Wannier, T. M. *et al.* Monomerization of far-red fluorescent proteins. *Proc Natl Acad Sci U S A* **115**, E11294-E11301, doi:10.1073/pnas.1807449115 (2018).
223 Megley, C. M., Dickson, L. A., Maddalo, S. L., Chandler, G. J. & Zimmer, M. Photophysics and dihedral freedom of the chromophore in yellow, blue, and green fluorescent protein. *J Phys Chem B* **113**, 302-308, doi:10.1021/jp806285s (2009).
224 Henzler-Wildman, K. A. *et al.* A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450**, 913-U927, doi:10.1038/nature06407 (2007).
225 Xu, Y. & Havenith, M. Perspective: Watching low-frequency vibrations of water in biomolecular recognition by THz spectroscopy. *J Chem Phys* **143**, 170901-170901 - 170901-170907, doi:10.1063/1.4934504 (2015).
226 Fisette, O., P., L., Gagné, S. & Morin, S. Synergistic Applications of MD and NMR for the Study of Biological Systems. *J Biomed Biotechnol* **7**, doi:10.1155/2012/254208 (2012).
227 Ortega, G., Pons, M. & Millet, O. in *Advances in Protein Chemistry and Structural Biology* Vol. 92 Ch. 6, 219-251 (Elsevier, 2013).
228 Chapagain, P. P., Regmi, C. K. & Castillo, W. Fluorescent protein barrel fluctuations and oxygen diffusion pathways in mCherry. *J Chem Phys* **135**, 235101, doi:10.1063/1.3660197 (2011).
229 Goedhart, J. *et al.* Bright cyan fluorescent protein variants identified by fluorescence lifetime screening. *Nat Methods* **7**, 137-139, doi:10.1038/nmeth.1415 (2010).
230 Balleza, E., Kim, J. M. & Cluzel, P. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat Methods* **15**, 47-51, doi:10.1038/nmeth.4509 (2018).
231 Farrow, N. A. *et al.* Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33**, 5984-6003 (1994).
232 Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone dyanamics of calmodulin studied by ^{15}N relaxation using inverse detected two-dimensional NMR spectroscopy. *Biochemistry* **31**, 5269-5278 (1992).
233 Shcherbo, D. *et al.* Far-red fluorescent tags for protein imaging in living tissues. *Biochemical Journal* **418**, 567-574, doi:10.1042/BJ20081949 (2009).
234 Bajar, B. T. *et al.* Fluorescent indicators for simultaneous reporting of all four cell cycle phases. *Nat Methods* **13**, 993-996, doi:10.1038/nmeth.4045 (2016).
235 Alvizo, O. & Mayo, S. L. Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proc Natl Acad Sci U S A* **105**, 12242-12247 (2008).
236 Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J. & Baker, D. Structure-guided forcefield optimization. *Proteins* **79**, 1898-1909 (2011).
237 Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* **523**, 109-143 (2013).
238 Gordon, S. R. *et al.* Computational design of an alpha-gliadin peptidase. *J Am Chem Soc* **134**, 20513-20520 (2012).
239 Khare, S. D. *et al.* Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat Chem Biol* **8**, 294-300 (2012).
240 Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309-313, doi:10.1126/science.1190239 (2010).
241 Walton, C. J. W. *et al.* Engineered Aminotransferase for the Production of D-Phenylalanine Derivatives using Biocatalytic Cascades. *Chem Cat Chem* **10**, 470-474 (2018).
242 Wolfenden, R. & Snider, M. J. The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res* **34**, 938-945 (2001).
243 Mak, W. S. & Siegel, J. B. Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr Opin Struct Biol* **27**, 87-94 (2014).

244 Keedy, D. A. *et al.* Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **4**, doi:10.7554/eLife.07574 (2015).
245 Jimenez-Oses, G. *et al.* The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat Chem Biol* **10**, 431-436, doi:10.1038/nchembio.1503 (2014).
246 Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *J Phys Chem B* **114**, 2549-2564 (2010).
247 LuCore, S. D. *et al.* Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures. *Biophys J* **109**, 816-826 (2015).
248 O'Meara, M. J. *et al.* Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput* **11**, 609-622 (2015).
249 Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J Chem Theory Comput* **14**, 2751-2760, doi:10.1021/acs.jctc.8b00033 (2018).
250 Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048, doi:10.1021/acs.jctc.7b00125 (2017).
251 Topham, C. M., Barbe, S. & Andre, I. An Atomistic Statistically Effective Energy Function for Computational Protein Design. *J Chem Theory Comput* **12**, 4146-4168, doi:10.1021/acs.jctc.6b00090 (2016).
252 Fumagalli, L. *et al.* Anomalously low dielectric constant of confined water. *Science* **360**, 1339-1342, doi:10.1126/science.aat4191 (2018).
253 Jaramillo, A. & Wodak, S. J. Computational Protein Design Is a Challenge for Implicit Solvation Models. *Biophys J* **88**, 156-171, doi:10.1529/biophysj.104.042044 (2005).
254 Vanommeslaeghe, K. & MacKerell, A. D. J. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim Biophys Acta* **1850**, 861-871 (2015).
255 Carlin, D. A. *et al.* Kinetic Characterization of 100 Glycoside Hydrolase Mutants Enables the Discovery of Structural Features Correlated with Kinetic Constants. *PLoS One* **11**, e0147596 (2014).
256 Yin, S., Ding, F. & Dokholyan, N. V. Modeling backbone flexibility improves protein stability estimation. *Structure* **15**, 1567-1576, doi:10.1016/j.str.2007.09.024 (2007).
257 Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74-79 (2017).
258 Gagne, D., Narayanan, C. & Doucet, N. Network of long-range concerted chemical shift displacements upon ligand binding to human angiogenin. *Protein Sci*, doi:10.1002/pro.2613 (2014).
259 Gagne, D. *et al.* Perturbation of the Conformational Dynamics of an Active-Site Loop Alters Enzyme Activity. *Structure* **23**, 2256-2266 (2015).
260 Gagne, D. & Doucet, N. Structural and functional importance of local and global conformational fluctuations in the RNase A superfamily. *FEBS J* **280**, 5596-5607 (2013).