Protein Design and Structure Determination at High-Precision

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Mohammad ElGamacy

aus Kairo, Ägypten

Tübingen

2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	20.11.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Andrei Lupas
2. Berichterstatter:	Prof. Dr. Dirk Schwarzer
3. Berichterstatter:	Dr. Arnout Voet

Protein Design and Structure Determination at High-Precision

Abstract

Due to the complementarity of the protein design and folding problems, progress on either front has consistently advanced the other. Although both problems remain major challenges, computational protein design has benefited amply from protein structure prediction methods. Likewise, the fields of structure prediction and structural biology have widely adopted techniques from the protein design field. The work I present here aims to put forward new protein design as well as structure determination strategies with the objective of achieving maximum precision. Both strategies capitalise on two posits: the first is that localising the sampling problem allows for exhaustive and finer granularity solution searching, while the second is that accelerated temporal dynamics can allow for directed and accurate exploration of energy landscapes. In the presented protein design projects, the level of precision was evaluated by comparing the coordinates from the experimental structures of the designs to their *in silico* models. Whereas in the structure determination projects, the precision was evaluated by how well a determined structure ensemble reproduces various experimental observables.

Since all of the previous design work utilising conserved supersecondary structures has aimed at constructing repeat proteins from amplifying a single fragment, my first project aims at designing an asymmetric globular (i.e. non-repetitive) fold from two unrelated supersecondary structures. I thereby conceive an interface-driven strategy aiming at constructing a viable intramolecular interface across the participating supersecondary structures. I report the successful design of the target fold that agrees with the experimental NMR structure at atomic precision (backbone RMSD of 0.9 Å), where the designed protein was substantially more stable than its closest natural counterpart.

Through the second project I aim to demonstrate the capacity of this interface-driven strategy to tackle the more difficult problem of novel fold design. The computational design of novel folds persists as a profound challenge, as in this case the association between structural and sequence features is absent *a priori*. This has kept most of the previous design efforts within the known fold space. I accordingly have expanded my interface-design methods, with the goal of achieving efficient sampling at maximum topological control. As a demonstration I conceive and design a novel *corrugated* protein architecture that does not exist in nature. The resulting NMR and X-ray structures for two different designs agree with the *in silico* models at atomic precision.

On the third project I develop a new generalised method for mapping protein conformational populations from NMR data by unravelling the distribution of states that underlie the

experimentally acquired average quantities. The CoMAND method does not only provide a quantitative mapping of the probabilities of the constituent microstates, but is also capable of extracting previously untapped structural information and solving structures *de novo* from a single NOESY experiment. I further present a detailed protocol that produces highly refined, dynamics-based ensembles without any recourse to heuristics or knowledge-based scoring. Finally, I validate the method's precision by using the refined ensemble to quantitatively predict NMR observables that are orthogonal to the NOESY data.

Proteinstrukturdesign und Proteinstrukturvorhersage sind zwei zueinander komplementäre Felder, deren Fortschritte sich stets gegenseitig voranbringen. Obwohl beide noch immer große Herausforderungen darstellen, hat Proteindesign von vielen Methoden der rechnergestützten Proteinstrukturvorhersage profitiert. Umgekehrt haben die Strukturbiologie und Strukturvorhersage viele Techniken aus dem Feld des Proteindesigns aufgegriffen. In dieser Arbeit präsentiere ich neue Strategien für das Design und die Aufklärung von Proteinstrukturen, welche maximale Genauigkeit anstreben. Diese Strategien folgen zwei Prinzipien: Erstens, dass eine verstärkte Betrachtung lokaler Lösungen eine vollständigere und feiner aufgelöste Abdeckung des Suchraums erlaubt; und zweitens, dass zeitliche Simulation von molekularer Dynamik eine genaue und zielgerichtete Abtastung der Energielandschaft ermöglicht. In den hier beschriebenen Design-Projekten wurde die Genauigkeit der designten Strukturen durch Vergleich mit deren experimentell bestimmten Koordinaten festgestellt. In den Strukturbestimmungs-Projekten wurde die Genauigkeit hingegen daran gemessen, wie gut ein experimentell bestimmtes Stukturensemble verschiedene experimentelle Beobachtungen reproduziert.

Die meisten bisherigen Proteindesigns auf Basis konservierter Supersekundärstrukturen sind repetetive Proteine die durch Wiederholung eines einzelnen Fragments erzeugt wurden. Vor diesem Hintergrund zielt das erste Projekt dieser Dissertation darauf ab, eine asymmetrische globuläre Struktur zu designen, die aus zwei unverwandten Supersekundärstrukturen zusammensetzt ist. Hierzu wurde eine Methode zur Konstruktion einer intramolekularen Kontaktfläche entlang der beteiligten Supersekundärstrukturen entwickelt. Die damit erfolgreich designte Proteinfaltung stimmt mit der experimentell bestimmten NMR-Struktur mit atomarer Genauigkeit überein (RMSD 0.9 Å). Das designte Protein ist zudem maßgeblich stabliler als die natürliche Struktur, die ihr am ähnlichsten ist.

Im zweiten Projekt dieser Disseratation veranschauliche ich die Leistungsfähigkeit dieser Kontaktflächen-Strategie durch deren Anwendung auf das weitaus schwierigere Problem eine neue Proteinfaltung zu designen. Dies stellt noch immer eine komplexe Herausforderung dar, da bei neuartigen Faltungen keine *a priori* Information zur Beziehung zwischen Sequenz und Struktur verfügbar ist; aus diesem Grund lagen bishere Designs zumeist innerhalb des Raums bereits bekannter Faltungen. Entsprechend habe ich die Kontaktflächen-Methode erweitert, um effizient Strukturen zu samplen und gleichzeitig die maximale Kontrolle über die Ziel-Sruktur zu behalten. Mit einer sich gegenläufig windenden Struktur ist mir das Design einer Protein-Architektur gelungen, die nicht in der Natur vorkommt. Die entsprechenden NMR- und Kristallstrukturen stimmen mit den designten Modell mit atomarer Genauigkeit überein.

In dem dritten Projekt habe ich eine allgemeine Methode entwickelt, um Protein-Konformationen anhand von NMR-Daten zu charakterisieren. Dabei wird die Verteilung der Konformationszustände errechnet, die den exprimentell bestimmten Daten zu Grunde liegen. Die CoMAND-Methode ermöglicht nicht nur eine quantitative Zuweisung der Wahrscheinlichkeiten einzelner Mikrozustände, sondern ermöglicht zuvor nicht auswertbare strukturelle Information zu bestimmen und Strukturen mithilfe eines einzigen NOESY-Experiments zu lösen. Zudem stelle ich ein detailliertes Protokoll zur Herstellung verfeinerter und dynamischer Strukturensembles vor, welches ohne Heuristik oder Vorkenntnissen von ähnlichen Strukturen

auskommt. Die Genauigkeit dieser Methode wird bestimmt, indem das verfeinerte Ensemble genutzt wird, um zu den NOESY-Daten orthogonale NMR-Beobachtungen vorherzusagen.

Contents

1 Chapter 1: Introduction **1**

1.1 Background 1

1.1.1 Protein folding 2

1.1.1.1 Folding mechanisms 2

1.1.1.2 Folding thermodynamics and kinetics 4

1.1.1.3 The fold space 7

1.1.2 Protein design 8

1.1.2.1 The sampling problem 8

1.1.2.2 The scoring problem 12

1.1.3 Protein dynamics 15

1.2 Research focus 19

1.2.1 Asymmetric globular domain design 19

1.2.2 Design of a novel protein architecture 23

1.2.3 Elucidating protein conformational landscapes in solution 28

2 Chapter 2: Asymmetric Protein Design from Conserved Supersecondary Structures **37**

3 Chapter 3: An Interface-Driven Design Strategy Yields a Novel Corrugated Protein Architecture **58**

4 Chapter 4: Mapping Local Conformational Landscapes of Proteins in Solution **81**

5 Conclusions **114**

References **121**

Acknowledgments

I would like to start by expressing my deepest gratitude for my advisor, Andrei Lupas. He has encouraged me from the start to undertake several approaches to protein design problems. Andrei always offered me his time when I needed it, and had a sustained faith in me and helped me to commit to my research even when getting complete results was very difficult at the beginning. After my first project had concluded successfully, he stimulated me to keep aiming for more difficult goals and to constantly improve my approaches. This support was further reinforced and corroborated by Murray Coles's supervision that started by my second year. Murray has not only guided me and taught me a lot of in-depth structural biology concepts, but also he has participated in expanding my view angle towards many concepts of protein dynamics and theoretical chemistry.

My thesis advisory committee (TAC) members have also offered me helpful directions and were very supportive of my progress throughout. Starting by Dirk Schwarzer who always has supported me when I needed help, Birte Höcker, from whose experience I benefited a lot, Birte Hernandez, who has not just offered me help through the TAC, but helped me with detailed scientific input whenever I faced difficulties in the laboratory. Lastly, thanks go to Hongbo Zhu, who has both provided me input through TAC recommendations, and detailed scientific advice with molecular modelling.

I would also like to thank the group leaders who offered me very fruitful collaborations. Particularly, Pak-Lee Chau, for all of the chemical physics research we did together, but also for all of his advice and teaching that has substantially broadened my knowledge, Andreas Plückthun and Marcus Hartmann, for their collaboration on solving the BRIC structures, Birte Hernandez, for her collaboration on designing active leads, and Patrick Müller for his collaboration on the exciting optogenetics research. I am also very grateful for coworking on one or more projects with Hongbo Zhu, Patrick Ernst, Vincent Truffault.

Many other friends and colleagues have offered me very significant, direct and indirect scientific support, and helped shape and prune my ideas throughout my doctoral studies. By a temporal order, these would be Andre Noll, Jörg Martin, Amit Kumar, Laura Weidmann, Reinhard Albrecht, Silvie Deiss, Eva Hertle and Astrid Ursinus.

Chapter 1: Introduction

Background

The deduction of the set of rules that governs protein sequence-structure relationship remains a major challenge in the field of biochemistry. This challenge has been principally categorised into two fundamental problems, depending on the mapping directionality; the protein folding problem and the protein design problem. The folding problem represents the sequence-to-structure mapping, where the sequence is predetermined and the structure is unknown. Conversely, the design problem represents the structure-to-sequence mapping, where the structural blueprint is defined and the sequence required to achieve that structure is unknown. Although these problems are intimately related, there is no exact symmetry between them due to the degeneracy of the design problem; whereas the sequence of a folded protein usually maps to a unique structure, a target structural template can map to many sequences.

The research presented here focuses on achieving efficient, high-accuracy approaches to protein design. In particular, protein design poses a hyper-dimensional problem given the associated compositional and configurational degrees of freedom that must be enumerated and evaluated *in silico*. In this work I evaluate alternative strategies to protein design, with the aim of simplifying the sampling problem while relying on more rigorous scoring schemes. To validate the resulting designs, structure determination is the major step. Here, objectivity is a major consideration, given that the design itself constitutes an expectation bias. While a quantitative measure of validation exists for X-ray crystallography (i.e. the R-factor), no equivalent exists for NMR-based methods. For this reason, I also aimed at developing a solution structure determination method that intrinsically quantifies the match between the solved structure and the input spectra. Thus, throughout my research I have aimed at testing the utility of strategies based on exhaustive localised sampling and accelerated temporal dynamics to achieve high-precision design and structure determination outcomes.

This section proceeds to briefly lay out the advances and challenges in fields of protein folding, protein design and protein dynamics from computational chemistry and structural biology standpoints.

Protein folding

Folding mechanisms

Proteins are the main molecules responsible for information processing, catalysis and mechanical roles in living cells. While a protein is synthesised in the form of one-dimensional peptide chain, for it to assume its biological roles, it mostly adopts a unique three-dimensional native structure. Early protein refolding studies have led to the Thermodynamic Hypothesis[1], which postulates that under physiological conditions, the native structure is the unique, kinetically accessible, and the most thermodynamically favoured configuration as dictated by the protein's sequence.

Four primary physical effects drive the folding of a linear protein chain into in a three-

dimensional structure, overcoming the large loss of conformational entropy in the process[2]: The first is hydrogen bonding, which is a directional, primarily - though not exclusively[3] - electrostatic interaction, which takes place between two electronegative atoms through an interstitial hydrogen atom. The backbone hydrogen bonding patterns are particularly important for folding as they define secondary structure types. The second is the hydrophobic effect, which is an entropic effect stemming from the cost of disruption of the dynamic hydrogen bonding of water by non-polar residues. The third is the Van der Waals force, which results in weak and very short-ranged interactions of induced-dipole nature. The fourth is electrostatic interaction, which occurs between formally or partially charged atoms, and can have magnitudes heavily dependent on the chemical environment.

A newly synthesised protein chain has to navigate a very rugged potential energy landscape under the influence of these forces to reach its native state. Assuming a random search of the configurational space, an average-sized protein would need billions of years to fold. However folding kinetics studies show that most proteins fold within seconds and some on even sub-millisecond timescales; this has been described as the Levinthal paradox[4]. This faster than expected folding implies that proteins fold according to biased pathways, which has led to the emergence of several hypotheses. For example, the *nucleation-growth* model proposes the formation of an initial folding nucleus within a group of adjacent residues, followed by the sequential folding of the rest of the protein. In contrast, the *framework* model proposes that secondary structures form first and then dock against each other into the native tertiary structure, possibly by a diffusion–collision mechanism. Lastly, in the *hydrophobic collapse* model, hydrophobic residues collapse together into a molten globule that forms a conformationally restricted intermediate state on the pathway to the native state[5].

Several experimental techniques have been applied to monitor the folding process, however the problem remains severely under-determined. Ideally, a structure determination technique with sufficiently high sensitivity (i.e. signal magnitude per substance concentration) and time-resolution (i.e. *shutter speed*) could provide a temporal monitoring of the folding process under native conditions. Although this is far from practical using present technologies, several spectroscopic techniques have been successfully deployed to monitor protein folding events in nanosecond and sub-nanosecond regimes[6]. Methods like Fourier-transform infrared

(FTIR), circular dichroism (CD), or fluorescence spectroscopy have particularly benefited from ultra-fast conformational triggers such as pressure and temperature jumps[7]. The fundamental drawback is that these methods lack any spatial resolution, as the resulting spectra describe the entire protein collectively, in addition to being ensemble averaged. To the end of acquiring fragment-wise spatial resolution down to millisecond time-scales, the highest possible sensitivity has been achieved through hydrogen-exchange mass spectrometry (HX-MS)[8].

A complementary approach would be computational; atomistic molecular dynamics (MD) simulations have been able to recapitulate the folding of small-sized proteins to millisecond spans, either through equilibrium MD[9], or Markov state model (MSM) MD methods[10]. Despite having relied on purpose-built supercomputers, or large GPU clusters, respectively, the accessible computing time constitutes the first major limitation. The second major limitation stems from the accuracy of the force field itself, as errors build-up with the simulation scale[10]. This is particularly the case for current force fields that employ parameters that stay constant throughout the simulation, despite the drastic changes in chemical environment along the protein's folding pathway.

Folding thermodynamics and kinetics

The thermal stability of the folded state (i.e. its folding thermodynamics) and how fast a nascent peptide chain folds (i.e. its folding kinetics) are governed by the folding landscape. This landscape can be described as an energy surface where the ordinate (the dependent variable) is the potential energy of the system, and the remaining dimensions represent the conformational degrees of freedom, perhaps also expressed by more abstract reaction coordinates (e.g. topological descriptors like secondary structure content or number of native contacts[9]). Defining every conformation as a microstate in a canonical ensemble, the probability of occurrence P of the unique microstate s_i is defined by the Gibbs distribution as:

$$P(s_i) = \frac{e^{-\beta E(s_i)}}{Z_\beta} \tag{1.1}$$

Where $\beta = 1/k_B T$, k_B is the Boltzmann's constant, T is the temperature, $E(s_i)$ is the potential energy of the microstate s_i, and Z_β is the partition function that serves as the normalisation constant over all of the possible microstates, whereby $Z_\beta = \sum_j e^{-\beta E(s_j)}$. This function enables the calculation of absolute free energies and the derivation of absolute entropy as the densities of associated microstates[11].

The thermodynamic stability of the folded state relates to its potential energy value and its associated density of states. The kinetic stability of the folded state, on the other hand, relates to the landscape ruggedness - e.g. the presence of kinetic traps that stabilise partially or misfolded species - plus the steepness and breadth of paths to the global minimum. Practically, however, there is no experimental means of detecting and evaluating the probabilities associated with exact microstates, hence a broader definition of configurations that encompass closely related microstates (e.g. folded *vs.* unfolded, or monomer *vs.* dimer) becomes useful. In a typical equilibrium unfolding experiment, for example, a macroscopic quantity related to "foldedness" is measured (e.g. CD ellipticity or fluorescence). Here the main goal is to determine the equilibrium constant of the unfolding reaction:

$$K_U = \frac{[P_U]}{[P_F]} \tag{1.2}$$

Where $[P_U]$ and $[P_F]$ are the concentrations of the folded and unfolded protein species, respectively. The free energy of unfolding can then be evaluated according to Eq. 1.3, and a small, single-domain protein would generally follow a first-order rate as in differential Eq. 1.4 :

$$\Delta G_U = -RT \ln K_U \tag{1.3}$$

$$\frac{dF_U(t)}{dt} = -k_f F_U(t) \tag{1.4}$$

Where ΔG_U is the observed free energy of unfolding, R is the gas constant, and T is the temperature. While $F_U(t)$ is the fraction of unfolded protein at time t, and k_f is the effective folding rate constant. Where the latter is related to the equilibrium constant K_U through the unfolding rate constant k_u according to $K_U = \frac{k_u}{k_f}$.

Under folding conditions, however, the fraction of unfolded species is vanishingly small and cannot be experimentally detected. Therefore, in equilibrium unfolding experiments chemical denaturants or temperature are commonly used to favour the denatured species, and the unfolding reaction is measured as a function of denaturant concentration or temperature. Likewise, with folding kinetics experiments the protein is initially unfolded using denaturants or high temperature, and rapid mixing or steep temperature drops are used to instigate the folding process. The measured observable in these experiments, which can be a single data point or a full spectrum, is at best an ensemble average in kinetics experiments, or an ensemble and time average in equilibrium unfolding experiments. In addition to averaging, the mathematical modelling over-reduces the vast underlying conformational heterogeneity into a binary categorisation of fraction folded F_F and fraction unfolded F_U. This demonstrates the fundamental lack of atomistic details on protein folding and unfolding mechanisms. Perhaps the most data-intensive research on this front has been the development of nanosecond-resolved 2D-FTIR[12] and its combination with MSM-MD simulations. This has so far only succeeded in providing semi-quantitative agreement between massively parallel folding simulations and time resolved spectra[13]. The main challenge facing this approach is the computational cost associated with the deployment of more accurate force fields that better account for bond stretches and their local electrostatic environments. To date though, there is no model that can accurately predict the folding rate or folding free energy of a protein from its structure. Moreover, absolute free energy and entropy estimations are impossible without accurately accounting for the partition function Z_β, which appears to be practically impossible given the implied conformational degrees of freedom. This proves to have far-reaching consequences for the protein design problem.

The fold space

According to differing terminology, a *fold* (according to SCOP[14]) or a *topology* (according to CATH[15]) can be defined as the three-dimensional arrangement of secondary structures in space and their connection order in a protein domain. Under this definition, a pertinent question is how many natural folds exist. There is an inherent difficulty in estimating this number depending on how strictly or loosely a fold is defined, given the existence of some continuity in the natural fold space[16]. Nevertheless, various estimates have placed this number between 1000 to 10000[17]. Although a very small figure relative to the number of theoretically possible folds, these estimates of the number of unique natural folds are not expected to grow significantly, despite the growing number of determined protein structures[18].

The limited number of known natural folds has led to the assumption that nature has historically sampled very narrow and clustered regions of fold space, presumably due to parsimonious evolutionary mechanisms[19]. This has raised the question on how large the "dark matter" (as Taylor calls it[20]) of the fold space may be; i.e. protein folds that are potentially viable, but not yet sampled by nature. Predictive models have shown these novel folds would massively outnumber the existing repertoire[20]. For example, the possibilities for arranging secondary structures (that can be either an α-helix or a β-strand) in one dimensional space has the ordering complexity of $2^{n_{ss}}$, where n_{ss} is the number of secondary structures. If these secondary structures are allowed to assume one of two orientations (only up or down), these possibilities become $2^{2(n_{ss}-1)}$. Trying then to account for all of the possible ways these secondary structures can be connected by loops would mean a complexity of $(n_{ss}-1)! \times 2^{2(n_{ss}-1)}$. The vastness of this space encourages the idea that the accessible fold space is more continuous than that presently observed[21]. The notion of tapping into unexplored regions of fold space then becomes very appealing; it also provides the major motivation for computational protein design as a means of selectively determining a target fold and constructing it *de novo*, ideally at atomic precision. This paves the way for generating novel scaffolds that can support the design and engineering of functional proteins.

Protein design

Computational protein design ultimately is aimed at finding the sequence that would achieve the most negative free energy change upon folding into a predefined blueprint fold, while also possessing the lowest possible configurational degeneracy (i.e. does not easily misfold into undesired configurations). Even though a viable solution can be reached for the former (i.e. finding a solution through a fixed-backbone design), the latter is not guaranteed, as the evaluation of the partition function is practically impossible for an average sized protein.

A straight forward *in silico* representation considers every constituting atom in the protein and its environment as an independent particle that is coupled by various forces to other atoms. Modelling such systems falls into the class of N-body simulations, where N is the number of atoms of the system. Given such a representation, the protein design problem can be decomposed in to two problems: the sampling problem (enumerating possible conformers and sequences) and the scoring problem (evaluating them). The sampling problem can be further decomposed into sequence sampling and conformer sampling, although that distinction between sampling types may be dismissed in Monte Carlo methods.

The sampling problem

Monte Carlo (MC) sampling encompasses a group of sampling algorithms with stochastic and statistical mechanical underpinnings. The most commonly used MC technique in protein modelling is the Markov Chain Monte Carlo approach (MCMC), where conformational sampling steps are performed in probabilistic, sequential moves. In MCMC, every conformation is considered as a state, every state is randomly perturbed every trial move, and every such trial move may be accepted or rejected. The central requirements for the validity of the sampling are the phase space accessibility and the ergodicity requirement (Eq. 1.5), where the relative probability of transition from an old state o to a new state n is equal to its relative

probability after a finite number of sampling steps:

$$\frac{P_{acc}(o \to n)}{P_{acc}(n \to o)} = \frac{P_B(n)}{P_B(o)} \tag{1.5}$$

Where $P_{acc}(x)$ is the acceptance probability of state x, and $P_B(x)$ is the Boltzmann weight of state x. This condition of *detailed balance* is necessary for an MCMC simulation to reach equilibrium. The acceptance probability of a move is 100% if the new state is more probable than the old state (i.e. $P_B(n) \geq P_B(o)$), otherwise, the acceptance probability is the ratio $\frac{P_B(n)}{P_B(o)}$, as in Eq. 1.6.

$$P_{acc}(o \to n) = min\left\{1, \frac{P_B(n)}{P_B(o)}\right\} \tag{1.6}$$

The main advantages of MC sampling include the possibility of making relatively large conformational jumps, which allows it to explore larger phase spaces than MD. In protein design applications, these conformational moves are commonly sampled from an empirical probability distribution derived from a library of known structures. The discrete sampling allows for trivial barrier crossing through arbitrary basin hopping to avoid entrapment. Also, since the MC ensembles are atemporal, a differentiable potential is not necessary for scoring, allowing for liberal scoring schemes that are often employed in protein design algorithms. Finally, in the context of design, the nature of MC sampling allows for combining conformational and sequence sampling in the same simulation, as the system is no longer strictly physical. On the downside, optimising for a more likely state through MCMC totally dismisses the associated entropy and thus cannot yield a true energy estimate. In addition, MCMC techniques cannot be practically deployed in explicit solvent without a steep drop in performance, to account for to complicated solvent rearrangements associated with solute moves[22]. Also, a key finding in the research represented here is that MCMC sampling schemes also tend to under-sample large-scale backbone rearrangements in comparison with MD.

In contrast to MC, MD aims to deterministically solve an N-body problem. The meaning

of *deterministic* here is that the underlying equations of motion can be integrated (forwards and backwards); accurately predicting the system's future state, given its present state. The ultimate goal of MD is to faithfully simulate the time evolution of a system, thus creating a temporal ensemble according to a predefined *thermodynamic state*. For example an isobaric-isothermal state, would imply that the macroscopic temperature and pressure are fixed. The premise of MD is that every particle (i.e. atom) possesses six degrees of freedom (rotational and translational) and its motion is modelled as a function of time under a potential V. The starting point for a simulation is a set of initial coordinates and velocities for every atom. The initial velocities are typically generated by a single thermal perturbation by assigning randomly distributed velocities (which can also be predefined), while correcting for zero net momentum. The force vectors - and hence the acceleration vectors - are calculated from the potential function at every time step, as shown in Equation 1.7. This differentiation step obligates that all of the components of potential V be differentiable.

$$F_i = -\nabla_i V = -\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{1.7}$$

Where F_i is the force acting at atom i, solving for the potential energy's gradient $\nabla_i V$, with respect to the atomic position vector r_i, and atomic mass m_i at time t.

Once initial coordinate, velocity, and acceleration vector sets are known, numerical integration of the equations of motion can performed. The system should time-evolve faithfully provided that the integrator preserves the conservation of energy and momentum. Different integration methods vary in computational efficiency and precision. For example, Equations 1.8 and 1.9 show how the leap-frog algorithm updates the velocities and coordinates every time step δt.

$$v_i\left(t + \frac{1}{2}\delta t\right) = v_i\left(t - \frac{1}{2}\delta t\right) + a_i(t)\delta t \tag{1.8}$$

$$r_i(t+\delta t) = r_i(t) + v_i\left(t + \frac{1}{2}\delta t\right)\delta t \tag{1.9}$$

Where $v_i(t)$ is the velocity vector for atom i at time t, $a_i(t)$ is its corresponding the acceleration vector, and $r_i(t)$ is the corresponding position vector. The leap-frog algorithm begins by solving for the next velocity midway between two time steps; $v_i\left(t + \frac{1}{2}\delta t\right)$, then, the coordinates are updated for the next time step; $r_i(t + \delta t)$. Because of this inter-step *leap* of velocity over the position, solving for the velocity at $v_i(t)$ can be achieved through the average: $\frac{1}{2}\left[v_i\left(t - \frac{1}{2}\delta t\right) + v_i\left(t + \frac{1}{2}\delta t\right)\right]$.

The goal of MD simulations is to uncover the microstates underlying a macroscopic observable. Thus, MD is based on the assumption that, given sufficient temporal sampling, the time-averaged properties of a system are equivalent to the ensemble-averaged properties and can accordingly reproduce macroscopic observables. Time scales accessible through equilibrium MD simulations, however, are extremely short - even when carried out under purely molecular mechanical potentials that largely dismiss electronic arrangements and electron dynamics. Even then, to date, only sub-millisecond time scales have been achieved in explicit solvent[10]. This expensive sampling, explicitly accounting for the time functions in molecular systems, is the cost that comes with the higher accuracy that MD offers. Several successful acceleration techniques, such as non-equilibrium and steered MD (SMD), are widely used for atomistic MD to enhance the sampling efficiency and reproduce observables at much lighter computational footprints. These include methods like replica exchange[23], MSM[10], umbrella sampling[24] and adaptive tempering[25].

Since atomically-resolved forces constitute a central quantity in MD calculation cycles, external forces can be modelled and exerted to accelerate configurational transitions that are otherwise inaccessible on natural time scales. So-called *Steering* can thus follow very advanced and adaptive schemes, where the system's condition is reassessed and the externally applied force vectors are adjusted on-the-fly through a programmatic framework. I demonstrate here that this external interference offers an invaluable tool for *focally* accelerating localised conformational events without disturbing the rest of the system and with no need

for applying restraints. This locally enhanced sampling benefits greatly from space reduction strategies, like fragment-based design and decomposing the sampling into separate successive problems (e.g. core-directed *vs.* loop-directed sampling).

The scoring problem

Proteins at physiological conditions constitute condensed phase regimes where the mechanics are non-relativistic and primarily classical in nature (i.e. The particles move at speeds much slower than 10^8 m/s, with masses effectively larger than 10^{-27} kg). Of course the classical mechanics assumption is only valid on the condition that no chemical bond creation or destruction is taking place. This allows for modelling such systems using a purely Newtonian description of the involved atoms motions, with disregard to the electronic structures of these atoms.

Since MC methods are based on sampling *moves* through randomised trials, they do not rely on computing forces. This lends their scoring functions to account for "energy" terms that do not have to be differentiable (i.e. they need not be smooth functions). Given the MC framework, these would not represent true energies anyway, which allows a more liberal interpretation of the energy terms included and their derivation origin. In practice, MC trajectories are aimed at optimising for the best scoring sequence-conformer combination, where the state-associated score is directly considered as "energy". For example, the Rosetta MC scoring function combines a set of scoring terms E_{term} and their respective tunable weights w_{term} [26] as follows:

$$\begin{aligned} E_{tot} = w_{attr}E_{attr} + w_{rep}E_{rep} + w_{sol}E_{sol} + w_{elec}E_{elec} + w_{hbond}E_{hbond} \\ +w_{paapp}E_{paapp} + w_{rama}E_{rama} + w_{dun}E_{dun} + w_{ref}E_{ref} + w_{pro}E_{pro} + w_{dslf}E_{dslf} \end{aligned} \tag{1.10}$$

Where the total energy score E_{tot} is a weighted mixture of: A Lennard-Jones attractive component E_{attr}, a Lennard-Jones repulsive component E_{rep}, a Lazaridis-Karplus implicit

solvation energy term E_{sol}, a short-ranged knowledge-based electrostatic term E_{elec}, a statistical, orientation dependent hydrogen bonding term E_{hbond}, a PDB-derived amino acid probability score given its backbone dihedrals E_{paapp}, a PDB-derived Ramachandran probability E_{rama}, a Dunbrack-library probability of side chain rotamer E_{dun}, a proline ring closure score E_{pro} and a disulfide bond geometry score E_{dslf}.

In contrast, the notion of "energy" in MD is more formal and not readily obtainable from the sampled states. Broadly, free energy methods in MD can be categorised into two main categories: density-of-states methods or work-based methods. Density-of-states methods are aimed at evaluating the ensemble weights among states within a small region of the phase space, in which the states are mutually accessible to each other. Work-based methods are based on a free energy perturbation that evaluates the work along the path between two states[27]. As described in the sampling section, MD routines need to evaluate the potential energy at every time step to be able to derive the forces. The potential energy function (often referred to as "force field" or "potential") relies on parametrisable models of inter-atomic interactions, which can be of a pair-wise nature (i.e. two-body) or higher order. For example, the CHARMM force field derives its parameters from quantum mechanical (QM) calculations that are often validated against experimental data. According to CHARMM36[28], the

potential energy V_{tot} is as follows:

$$\begin{aligned} V_{tot} = & \sum_{bonds} k_{bond}(r - r_o)^2 \\ & + \sum_{angles} k_{angle}(\theta - \theta_o)^2 \\ & + \sum_{Urey-Bradley} k_{UB}(u - u_o)^2 \\ & + \sum_{impropers} k_{improper}(\omega - \omega_o)^2 \\ & + \sum_{dihedrals} k_{dihedral}[1 + cos(n\phi + \delta)] \\ & + \sum_{CMAPs} \sum_{i=1}^{4} \sum_{j=1}^{4} c_{ij} \left(\frac{\phi - \phi_L}{\Delta_\phi} \right)^{i-1} \left(\frac{\psi - \psi_L}{\Delta_\psi} \right)^{j-1} \\ & + \sum_{LJ} \gamma_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] \\ & + \sum_{electrostatic} \frac{q_i q_j}{4\pi\epsilon_o r_{ij}} \end{aligned} \tag{1.11}$$

Where the first four terms are single-welled harmonic potentials with force constants k_{term}, and equilibrium values: r_o for bond lengths, θ_o for three bonded atoms angles, u_o for 1-3 bonded atom distance (for UB cross-term that corrects for angle bending), and ω_o for improper dihedral (out-of-plane) angle. The fifth term describes a 4-body torsion angle potential with force constant $k_{dihedral}$, periodicity n, phase shift δ, calculated at torsion ϕ. The CMAP term provides a corrective cross-term for the backbone (ϕ, ψ) dihedrals, where the input is a relative free energy grid derived from QM-level potential of mean force (PMF) calculations, resolved at $(\Delta_\phi, \Delta_\psi)$ steps. The term is a bicubic interpolation function of the input grid (where c_{ij} are the precomputable interpolation coefficients for a given map), which is a smooth differentiable function with continuous second derivatives across the boundaries. The Lennard-Jones term approximates the London dispersion forces where γ_{ij} is the well depth for atom pair (i, j), r_{ij} is the atom pair distance, and $R_{min,ij}$ is the distance between the same atom pair corresponding to the minimum of the well. Finally, the electrostatic interaction can be described by a Coulombic term between a pair of atomic charges (q_i, q_j),

separated by distance r_{ij}, at a dielectric constant ϵ_o. Practically though, this last term is poorly scalable and very expensive to compute directly, as the potential decays slowly with distance (i.e. $1/r$ for Coloumb's *vs.* $1/r^6$ for LJ), so its long-range component is computed through more efficient methods (e.g. Particle Mesh Ewald), which reduce its complexity from $O(N^2)$ to $O(NlogN)$.

Previous work has investigated the utility of temporal, density-of-states approach in the context of protein design (the VALOCIDY method)[29]. The VALOCIDY approach is aimed at the estimation of absolute free energies from unperturbed ensembles, and only attempted a temperature elevation scheme to traverse energy barriers. Although the approach shows good convergence behaviour, it only reached that after microsecond-scale simulations and only for seven-residue-long peptides. Given the intractability of global density-of-states methods in the context of design problems (where the protein size is much larger than a few amino acids), decomposing the problem and evaluating relative free energies across predefined state transitions would provide an appealing proposition. Throughout the work I present here, I attempt to test the utility of work-based methods in protein design for estimating relative free energies *in silico*, and thus address the scoring problem in a tractable and convergent manner. To implement this successfully, the search space must be effectively simplified and reduced, which emphasises the necessity for reducing the sampling spaces and simplifying our objective optimisation functions (as described above). Free energy perturbation methods and perturb-probe schemes in particular, were tested for the ability to capture the intermediate state energetics, and hopefully, achieve more accurate free energy estimations. Such intermediate states may be easily dismissed by density-of-states methods (due to the obligate under-sampling at average-sized proteins) or by end-state methods (due to the assumption of a linear transition between initial and final states).

Protein dynamics

Early studies on enzymatic catalysis[30] hypothesised the necessity of *template flexibility* to account for the the underlying reaction specificity. This started the historical debate between the *conformational selection*[31] and the *induced fit*[32] models of dynamic recognition

mechanisms, and motivated a broad range of dynamics studies. Various classes of techniques have been used to query dynamic properties of proteins, like deuterium exchange, IR spectroscopy, fluorescence spectroscopy and Raman spectroscopy[33]. However, none of these techniques could specifically map the probed dynamical properties at atomic resolution. The exclusive capacity of nuclear magnetic resonance spectroscopy (NMR) to acquire atomically-mapped, dynamics-related observables for proteins in solution has placed it front and centre in the protein dynamics field.

The simplest 1D NMR spectrum presents readily obtained experimental observables that can directly report on protein dynamics. The three fundamental quantities obtainable from such spectra are peak intensity I (i.e. peak integral), absolute resonance frequency v (which can also be represented as a relative chemical shift δ; $\delta = \frac{v}{\gamma B_o}$), and linewidth λ (i.e. peak width at half hight). In modern FT-NMR, data is collected within mid-millisecond time frames during the *detection time* of an experiment. This data comes in the form of a time function of the detected signal intensity $I(t)$, and is Fourier-transformed into the frequency domain (that can be represented in absolute Hz units or relative ppm units), in which the spectrum appears as a function of frequency $I(v)$ (or $I(\delta)$).

Figure 1.1 illustrates the possible spectral scenarios for a given magnetically active nucleus undergoing *chemical exchange*; i.e. dynamically transitioning between two states with distinct chemical environments, and thus appearing at two different chemical shifts *chemical shifts* (i.e. frequencies). States A and B are connected through rate constants k_A and k_B, an exchange rate constant $k_{ex} = k_A + k_B$, and possess relative probabilities of P_A and P_B. Here, v (or δ) reports on the chemical environment change between the exchanging species, λ reports on their interconversion rates (i.e. kinetics), and (v) reports on their relative probabilities (i.e. thermodynamics) in the frequency domain.

A range of other dynamical quantities can be derived from simple NMR experiments at different time scales. For example, the magnitudes of the principal magnetisation vectors along the z-axis (i.e. *polarisation* along the constant field B_o direction) and along the rotating frame of reference $x'y'$ plane (i.e. *coherence*). The decay of these two vectors to their equilibrium values (i.e. the recovery of net polarisation along z and the loss of coherence

around $x'y'$, yield The corresponding T_1 and T_2 relaxation time constants. These are readily obtainable through standard relaxation experiments, and these time constants are used to derive a generalised order parameter S^2 for individual bond vectors according to:

$$S^2 = \frac{T_2^{-1} - T_1^{-1}}{T_{2R}^{-1} - T_{1R}^{-1}} \tag{1.12}$$

Where T_{iR} is the respective time constant denoting the rigid tumbling limit value[34]. S^2 yields information on the local disorder and can be directly converted into conformational entropy values through a Boltzmann factor upon a perturbation. It was also shown that for proteins where $\Lambda > 1$ an experimental PMF can be established as function of temperature (Eq. 1.13)[35]; i.e. localised energy change can be mapped.

$$\Lambda = \frac{\Delta ln(1 - S^2)}{\Delta lnT} \tag{1.13}$$

Advanced pulse sequences can further serve to extract information from otherwise obfuscated observables as with the cases of intermediate and fast exchange regimes in Figure 1.1. Specialised pulse sequences like CPMG and rotating frame relaxation dispersion[36] have succeeded in decomposing the apparent exchange rates in the intermediate and fast exchange regimes through refocusing the exchange-based broadening and thus allowing for evaluating the underlying exchange rates.

Although the above are just a few examples of many acquirable NMR observables on protein dynamics, the information acquired through NMR does not explicitly involve structural coordinates. In combination with the fairly long evolution, detection, and relaxation delays, and the obligate ensemble averaging, there are still fundamental limitations to the utility of this information. That necessitates the close association of NMR and molecular dynamics studies to attempt to solve for causality of events, internal correlations, and mechanistic interpretation of the dynamic events. My conformational mapping research presented below capitalises on major advances recently achieved in both fields (NMR and MD) to extract

new information and enable more accurate solution structure determination.

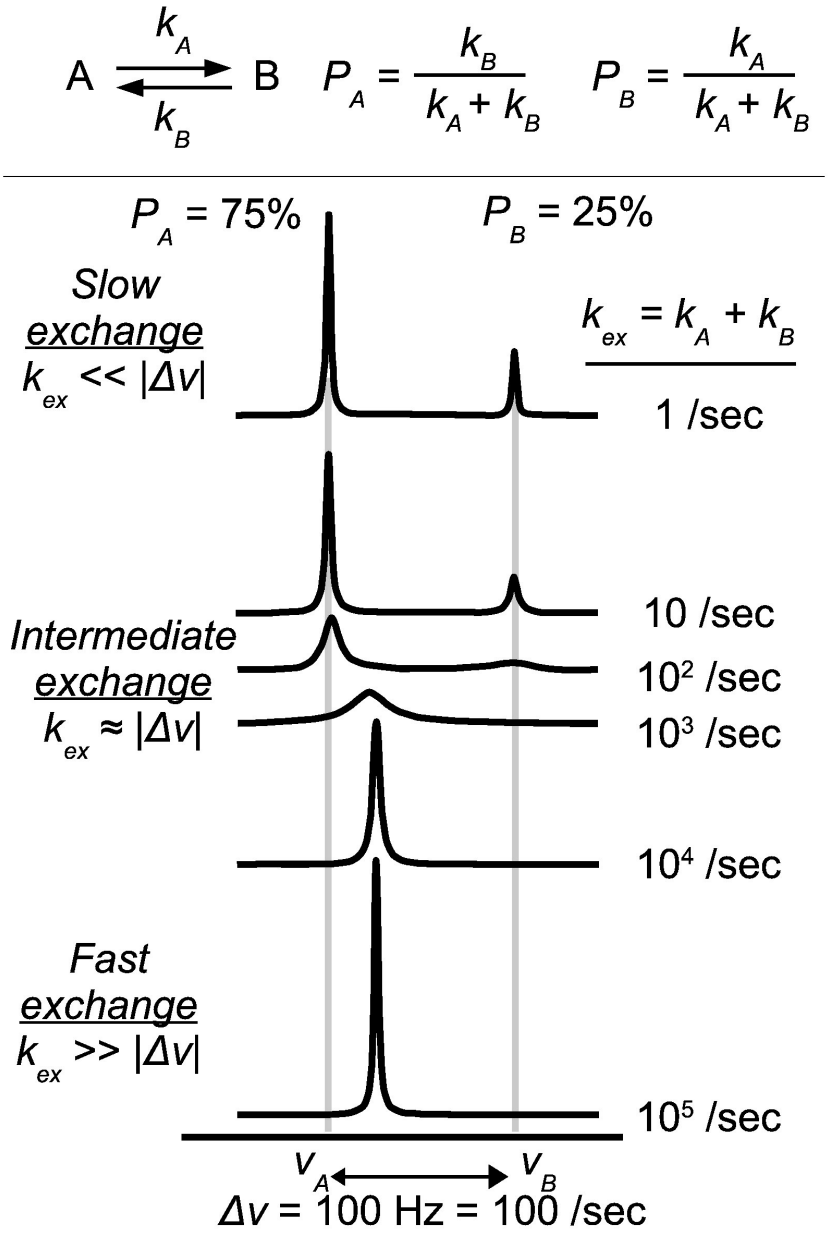


Figure 1.1: The three limits of chemical exchange regimes and the associated effects on three main NMR observables. Intensity $I(v)$, resonant frequency v and linewidth *lambda* report on the kinetics and thermodynamics of chemical exchange between two species. The slow, intermediate and fast exchange rates represent the three limiting cases for the exchange rate constant k_{ex}. Reproduced with permission from Fig. 3 in Ref.[36]. Source: Elsevier

Research focus

The fundamental and conceptual overlaps between the design and folding problems have started to trickle down to technical and methodical levels during the past two decades. Thanks primarily to the huge leaps in accessible computing power and advancements in sampling and scoring methods. This progress has effectively blurred the lines between the topics of protein design and structure prediction problems[37]. This has also driven the recruitment and repurposing of robust protein design algorithms towards solving structure determination problems, spanning the fields of NMR[38], X-ray crystallography[39] and cryo-electron microscopy[40].

Throughout my research I aim to present novel design and structure determination strategies that emphasise the roles of protein dynamics on both fronts. I provide a strong case for exhaustive, localised sampling schemes in combination with temporal, dynamics-based evaluation schemes as common bases for both high-precision protein design and structure determination.

Asymmetric globular domain design

Many natural protein structures show tandem repetition of supersecondary structural units. That detectable homology is often preserved among their constituent fragments and has suggested amplification as a powerful, parsimonious means for fold evolution[41]. Such repeat proteins have a special status as protein folding model systems, owing to their low-dimensional folding landscapes[42]. The simplicity and supersecondary structural modularity of repeat proteins have inspired numerous successful protein design studies that recombine conserved supersecondary structures[43]. To this end, sequence profile-based consensus design combined with computational design has proven a robust means for generating thermodynamically superior proteins with novel sequences[44].

In contrast to previous studies that deployed motif-based design to construct repeat proteins, this design study is aimed at designing a non-repetitive globular fold from two con-

served supersecondary structures. Since in repeat proteins the folding enthalpy is provided by an interface that is amplified multiple times, their folding stability is directly proportional to the number of interface repetitions. Their folding free energies start to become negative mostly after 2 or more interfaces are formed[45]. Therefore the challenge for an asymmetric, non-repetitive fragment recombination lies in that the intra-molecular interface between these two fragments is almost the sole source of folding enthalpy. Here, I tried to overcome this by building on two principles. The first is that consensus and consequently conserved sequences tend to average out kinetic traps and improve folding stability[46] (Fig. 1.2). The second is that even if the supersecondary structural building blocks used cannot fold on their on, they still possess residual folding information. This should allow me to focus computational sampling on only the inter-fragment interface, perform such sampling at high granularity, and deploy more expensive scoring routines, as the available computing power is redirected to a small portion of the protein.

My goal here was to design a novel dRP lyase domain from two conserved supersecondary structures, namely, an $\alpha\alpha$-hairpin and a helix-hairpin-helix motif (HhH-motif). This was motivated by the observation that two or more evolutionarily conserved supersecondary structures cannot be detected to coexist in a single domain[47], this in spite of the ubiquitous natural repetition of these individual supersecondary structures in many folds. The TPR-like $\alpha\alpha$-hairpin and HhH-motif combination thus provides an attractive goal, given their widespread occurrence in numerous natural folds. The design objectives were to achieve a compact, single-domain protein that possesses no full-length homology to any existing protein, yet optimally, more stable than its closest structurally similar counterpart, the human dRP lyase domain.

It is worth emphasising here that the reliance on ever smaller fragments in a fragment-based design framework allows stricter control of the target topology. It is also a very effective strategy in bringing together fragments with structural oddities that are otherwise difficult to sample *de novo*[48]. Here I describe my interface-driven computational design strategy and the experimental evaluation (biophysical characterisation and structure determination) of the designs resulting from the scheme described in Figure 1.3.

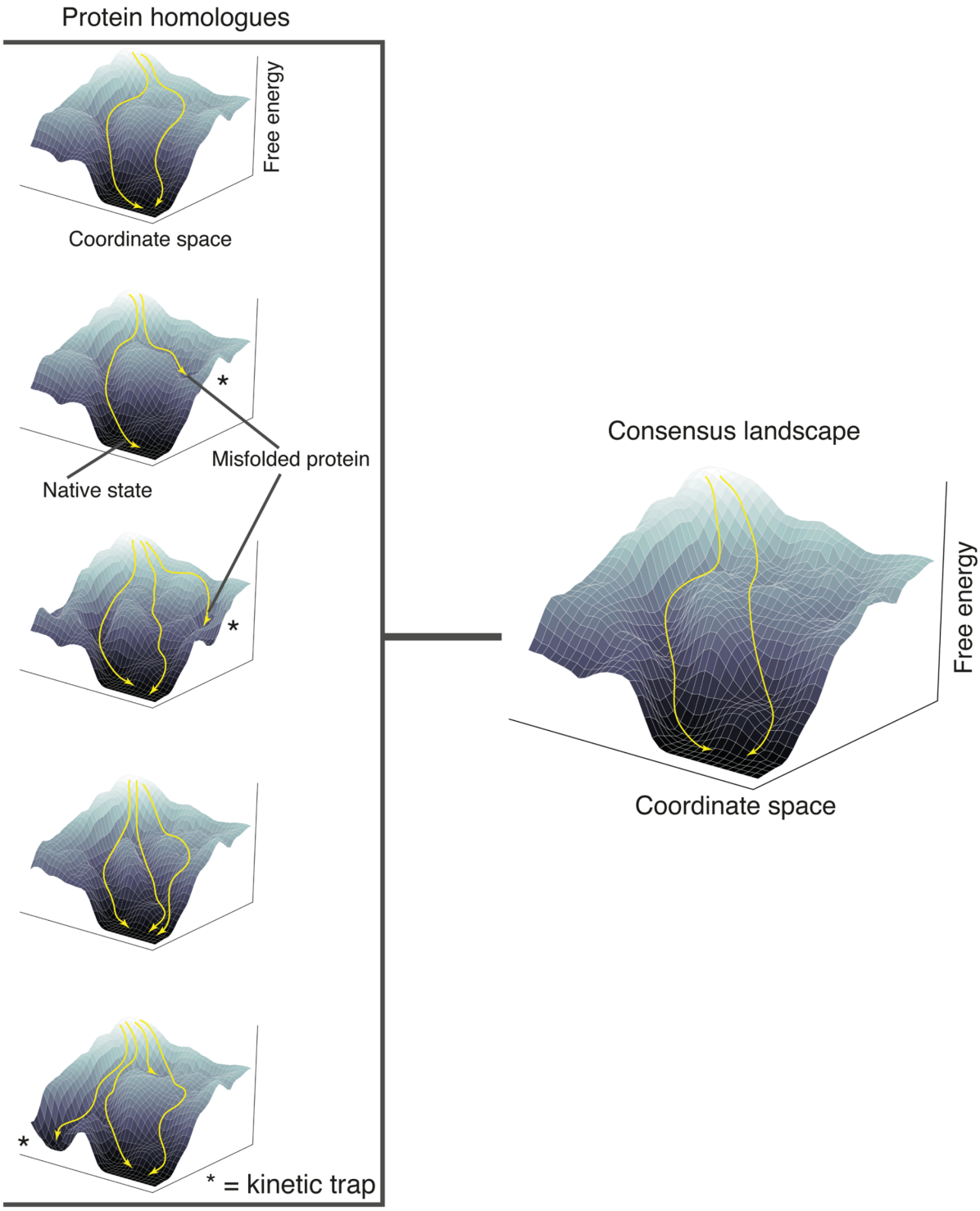


Figure 1.2: Sequence averaging is associated with energy landscape averaging. The average of five hypothetical folding landscapes for five homologous sequences levels out kinetic traps and relatively broadens the minimum basin. Reproduced with permission from Fig. 2 in Ref.[46]. Source: Oxford University Press

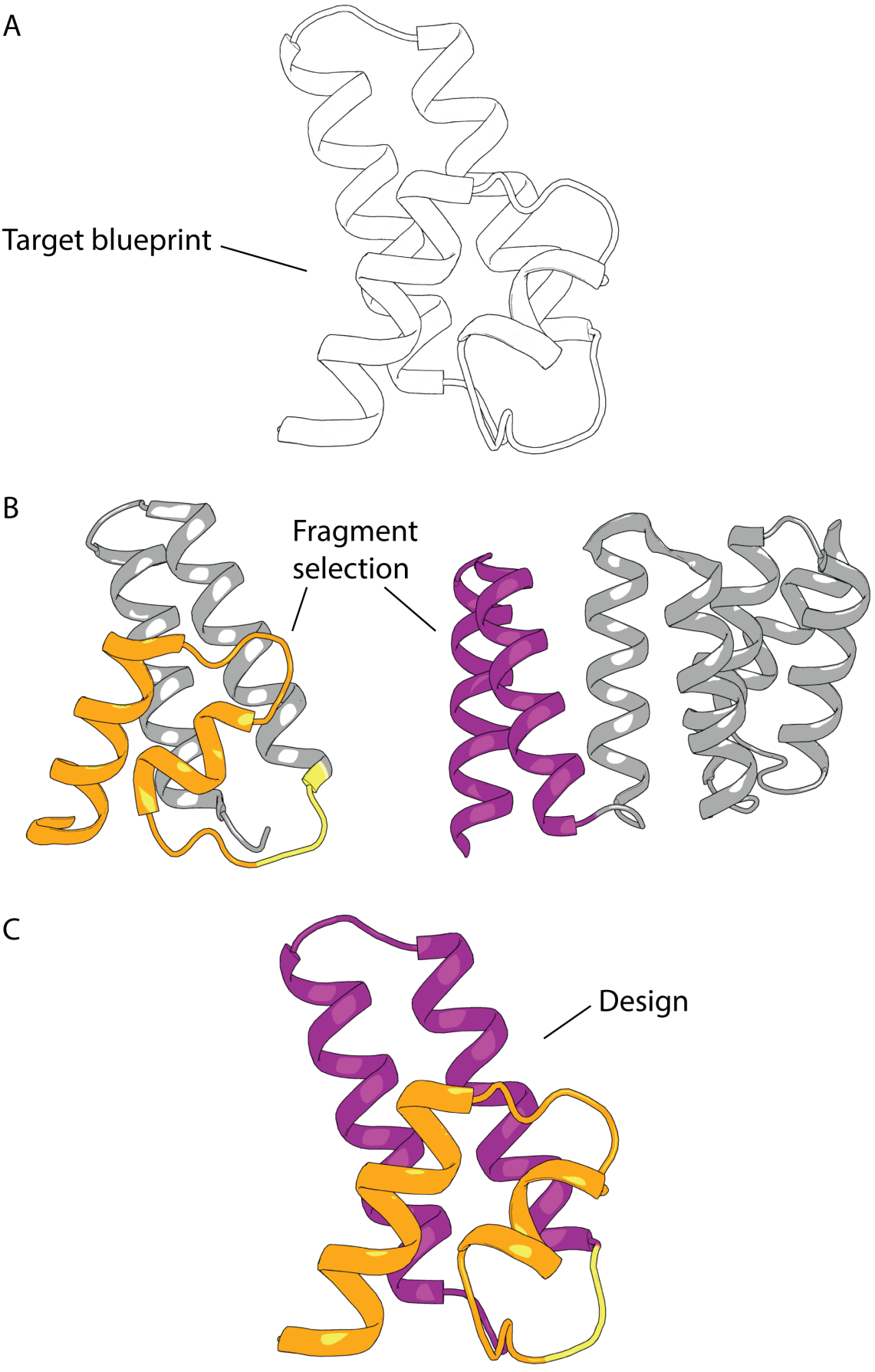

A
Target blueprint
B
Fragment
selection
C
Design

Figure 1.3: The design of a globular asymmetric domain from heterologous fragment recombination. A) The definition of a target backbone blueprint based on the fold of the human dRP lyase. B) Structural fragments are chosen from a pool of conserved sequences on a geometric basis, and deliberately adopted from unrelated sources. C) Design process aimed at constructing a novel hydrophobic core at the interface of arbitrary juxtaposed fragments.

In the first chapter, I describe my design pipeline in detail, and proceed to report my experimental characterisation results for four different designs.

Design of a novel protein architecture

Designing proteins with tailored structural features like internal orientations, pockets, curved surfaces, or grooves is still a challenging task. The most demanding form of such manipulations is the design of proteins with novel folds, ultimately at atomic precision. This is particularly difficult as the biophysical properties of the target fold are not known *a priori* and no sequence profile exists to describe its features. Therefore, most computational design efforts so far have been directed towards creating novel proteins recapitulating existing folds. Moreover, the difficulty of this challenge (if difficulty is defined by the computational power required per successful design) is highly non-uniform. This is due to several complicating factors such as the protein size, the size-scalability of the design algorithms, the involved structural motifs, and even the optimal folding landscape of the target blueprint itself (e.g. excessive internal symmetries can cause obligate kinetic frustration[49]).

To date, two successful attempts that used different computational strategies to generate novel folds have been reported. The first used iterative rounds of global Monte Carlo sampling and *ab initio* structure prediction to optimise the sequence and rotamers starting from an initial backbone blueprint[50]. The second utilised overlapping helical stretches as junction points to transition between natural fragments[48]. While the first strategy offers strict control over target topology (assuming no topological drift across iterations), the *en masse* MC

sampling strategy is not size-scalable and does not guarantee convergence, even for small-sized proteins. The second strategy, on the other hand, while deploying a size-scalable sampling with tractable convergence, does not possess control over the resulting topology, as the latter is strictly contingent on the chance of finding viable interfaces across the available fragments.

Here I aim to lay out a strategy centred upon the design of novel intra-molecular interfaces that enables the construction of a target fold from a set of starting fragments with an arbitrary orientation. In addition to its generality, this strategy effectively reduces the amount of computational sampling necessary to achieve an optimal sequence, without compromising the level of topological control. The fragment-based approach applied here preserves block-wise, linear size-scalability and focused sampling, which improves convergence properties. On the other hand, the *de novo* design of loops and arbitrarily oriented interfaces guarantees the enforced topology (Fig 1.4). This strategy should be applicable to starting fragments of any size (i.e. motifs or domains), or composition (i.e. designed or natural), which should be particularly useful for incorporation of functional but energetically perturbed motifs or structural oddities that are otherwise difficult to sample[51].

I provide an example by aiming at designing a novel *corrugated* protein architecture that does not exist in nature. The solenoid architecture is defined by a uniform connectivity across its repeat units and thus winds into a continuous superhelix. The implied single loop and interface types and obligate, uniform handedness per repeat is commensurate with a sawtooth wave (Fig. 1.5A). This inspired me to try to double this level of topological complexity, through doubling the waveform phase span; reversing the polarity periodically lead to a triangle waveform. Figure 1.5B shows the conceived *corrugated* architecture, that implies two interface types and two loop types, resulting in a bi-handed repeat. This novel fold would be satisfied upon the adjoining of simpler building blocks (up-down four-helix-bundles) into a new single domain; achieving the target fold.

I propose the following scheme:

1. Fragment picking, arbitrary docking and conformational refinement.

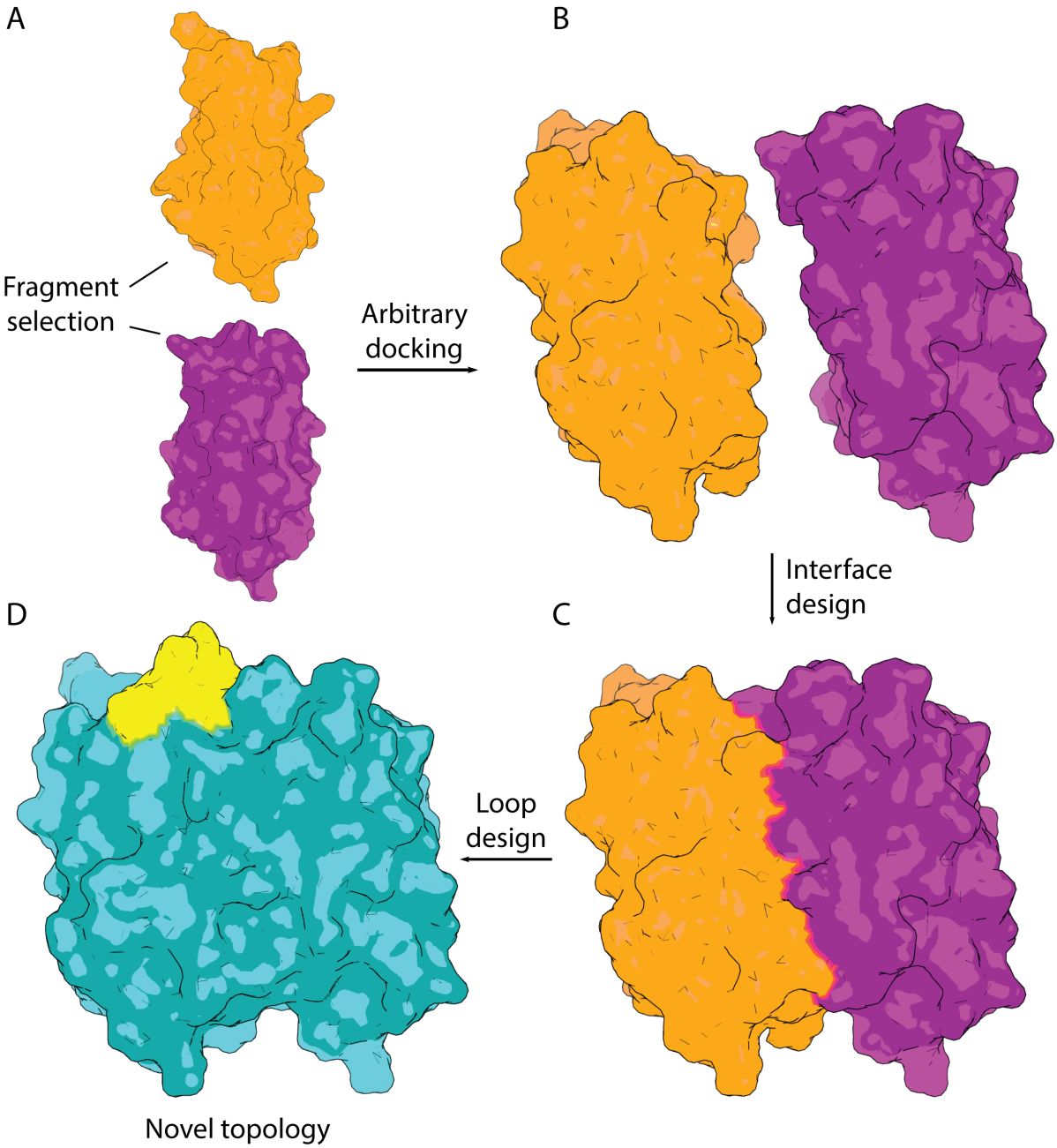


Figure 1.4: The interface-driven design strategy can efficiently yield proteins with novel topologies. The three-step design process follows the main stages of: A) Selection of the participating fragments according to the geometric criteria of choice. B) Arbitrary docking of the two fragments and constrained refinement of the defined poses. C) Design of an inter-fragment *de novo* interface. D) Design of a loop that bridges the fragments across the interface.

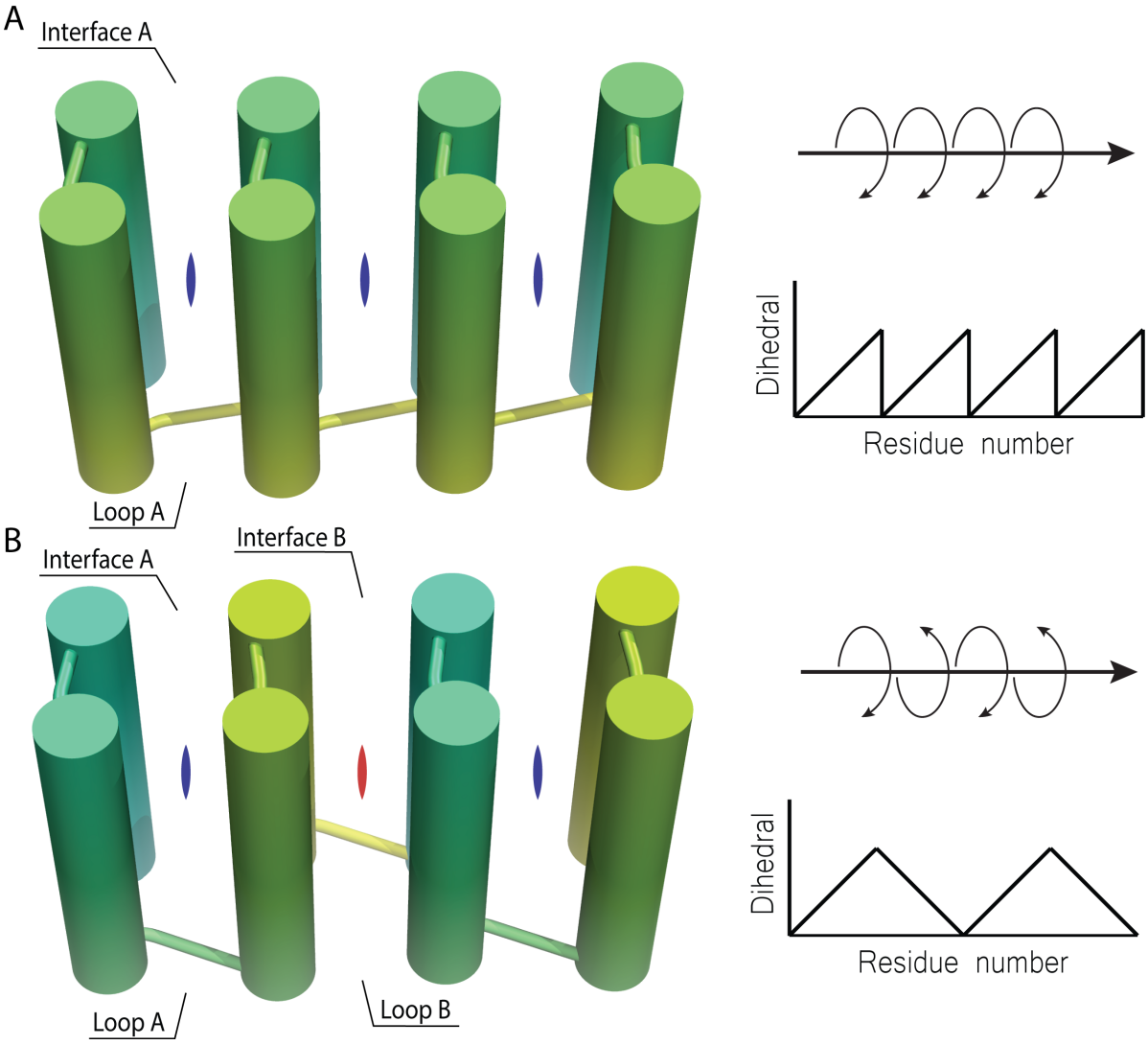


Figure 1.5: A corrugated repeat entails double the topological complexity of a solenoid repeat. A) Natural solenoids encompass single interface and loop types, uniform handedness, and net writhe that is analogous to a sawtooth waveform. B) In contrast, a triangle waveform is twice as complex (with double the phase span). The latter waveform models into a corrugated repeat that encompasses two loop and two interface types, no net writhe, and a bi-handed form.

- The three designs were attempted based on diverse building blocks, where two were adopted from natural helical bundles, and the third from a previously computationally designed bundle. This was done to demonstrate the method's robustness, regardless of the source of the blocks.
- Only geometric terms were used to select bundle regularity and self-compatibility in the context of the complex repeat. The fragments then underwent constrained flexible docking with their translated images so as to minimise bend, twist and curvature at the designed interface.

2. Combined design and geometric filtering.

   - I describe a design protocol that comprises cycles of softened-repulsion sequence sampling, side-chain sampling, backbone optimisation, docking and global conformational refinement. This was interlaced with geometric filtering using an analytically derived metric that can detect packing irregularities.
   - This metric has was intended to detect over- or under-packing in the averaged atomic environment. These irregularities in the atomic pair correlation function have been successful in discriminating high-resolution structures from poorly refined ones, and robust from poor designs.

3. Unidimensional PMF-based interface ranking.

   - I then deploy an adaptive potential of mean force (PMF) scheme with the aim of calculating a variable-velocity, variable-force dissociation work function. Such a PMF scheme, upon sufficient sampling, should account for the system's intermediate states and transient configurational changes along the reaction coordinate of protomers dissociation. Therefore the simulation was calibrated for the best correlation of affinities with experimental data at the fastest possible pulling velocity against a protein-protein affinity benchmark. This was used to rank interfaces based upon a more expensive and accurate estimation of their formation free energies.

4. Compatible loop retrieval and design.

- I then developed a fast, geometric search routine that applies a vectorial description for protein fragments across a gapped sequence sliding window. The aim of this was to seek initial loop structures in the PDB with similar landing sites to the termini of the designed fragments.
- These starting loop configurations are grafted across the corresponding designed interfaces and put through a design protocol for extensive sequence and rotamer sampling. This search-graft-design scheme greatly simplifies the backbone configurational sampling required during the design scheme, as it provides potentially viable initial loop configurations that can be optimised into their structural context.

5. Rotational Force Dissipation (RFD).

    - I here again conceived and implemented an accelerated sampling scheme where I titrate a ramp of crankshaft torque along the centre-most peptide bond in the designed loop, and evaluate the resulting kinetic perturbation from that force. This scheme was also validated against an internally compiled dataset of structured *vs.* random-sequence loops.
    - The least kinetically responsive loops in this perturb-probe scheme were chosen for experimental evaluation.

In the second chapter, I describe my design strategy in detail, and report my biophysical and structural results for three different design that have been experimentally evaluated.

Elucidating protein conformational landscapes in solution

For proteins to act as active effectors, they must possess complex dynamic landscapes that are at least bistable, as functions like catalysis, signal transduction or transport necessitate the protein molecule to partition between at least two different configurations. This motivates a structural description commensurate with the underlying complex dynamical properties; a description that does not just aim at generating an ensemble representation, but also

reconciles the macroscopic experimental observables with the statistical mechanical features of the underlying system.

Among the main drawbacks of current protein structure determination methods is that they either yield unrealistically static or unrealistically averaged models of the protein molecule. In case of protein crystallography, although sub-atomic resolutions are practically attainable, cryo-crystallography acquires data in an environment of heavily depressed atomic displacements[52]. While NMR offers the advantage of acquiring data from a thermalised, solution-state environment, the method's sensitivity and acquisition *shutter speed* only result in data that describe time and ensemble averages of the underlying protein dynamics[53]. Moreover, the current paradigm of solving protein NMR structures primarily relies on querying the binary information of whether an inter-atomic contact exists at a particular chemical shift from NOESY spectra, which largely dismisses the spectral complexity and information content, and entirely ignores the negative information conveyed by peak absence. These extracted contacts are then deployed as restraints in simulations with the goal of minimising restraint violations simultaneously across all restraints; achieving coordinates ensembles closest to the average[54]. Such an over-restrained description effectively veils the underlying microstate distributions and their associated covariances.

My goal through this work was to accurately and comprehensively query the configurational distributions underlying the acquired average quantities. Although the problem of structural determination in general is under-determined, the reliance on constraints can render it tractable. Unlike other structure determination approaches that routinely deploy bioinformatic biases and heuristic methods, this work was aimed at solely deploying unbiased mathematical and physical constraints to render the structure determination problem solvable. The solution sought in this work aimed at solving protein solution structures while unravelling the averaged NOESY spectral data to solve the microstate probability distribution function. The reliance on the nature and merits of the CNH-NOESY experiment allows for generating spectra clear of many experimental artefacts and features that would otherwise complicate the accuracy of back-calculating ^{1}H-edited NOESY spectra. Once the theoretical back-calculation of a spectroscopic quantity can describe the experimentally acquired data down to the noise level, a direct mapping can be established between the spectral space and

the conformational space.

This direct mapping between the spectral and their associated conformational spaces can pave way for quantitative approaches to solution structure determination, validation and refinement. Here I aimed to capitalise on this mapping capacity to develop a new approach to solve protein solution structure and describe its dynamics through a scheme we now call CoMAND (for Conformational Mapping via Analytical NOESY Decomposition). Figure 1.6 describes the main structure and dynamics elucidation scheme of the CoMAND approach.

I have defined my objectives and their assumptions sequentially as follows:

1. Spectral decomposition using an accurate theoretical feature set.

    - I commence by the assumption that every amino acid in the protein assumes one or more conformations in solution, and hence, a NOESY experimental spectrum of that residue is a linear combination of one or more spectral components. Each of these spectral components corresponds to a unique conformation.
    - I seek the back-calculation of theoretical spectra that describe every conformer individually. In this step, it is affordable to perform systematic localised sampling of tripeptide or even pentapeptides along a sliding sequence window of the protein. At this stage the angular step size can be varied for the $(\phi_i, \psi_{i-1}, \chi_{i,1}, \chi_{i,2})$ conformational degrees of freedom, as shown in Figure 1.7A (ψ_{i-1} is referred to here as υ_i). This *shifted-Ramachandran* space of (ϕ_i, υ_i) captures more of the variation in peak intensities observed for the amide proton of residue i than the conventional Ramachandran space.
    - The back-calculation is performed using a existing software, SHINE (Simulation of Hetero-Indirect NOESY Experiments), which is a heteronuclear adaptation of the original SPIRIT software[55]. The back-calculated spectra of the systematically sampled conformers are used to construct a matrix W. Figure 1.7B shows an overlapping plot of all spectra (vectors) within the matrix of L67 of human ubiquitin (hUb). These represent the NOE contacts between the residue's amide

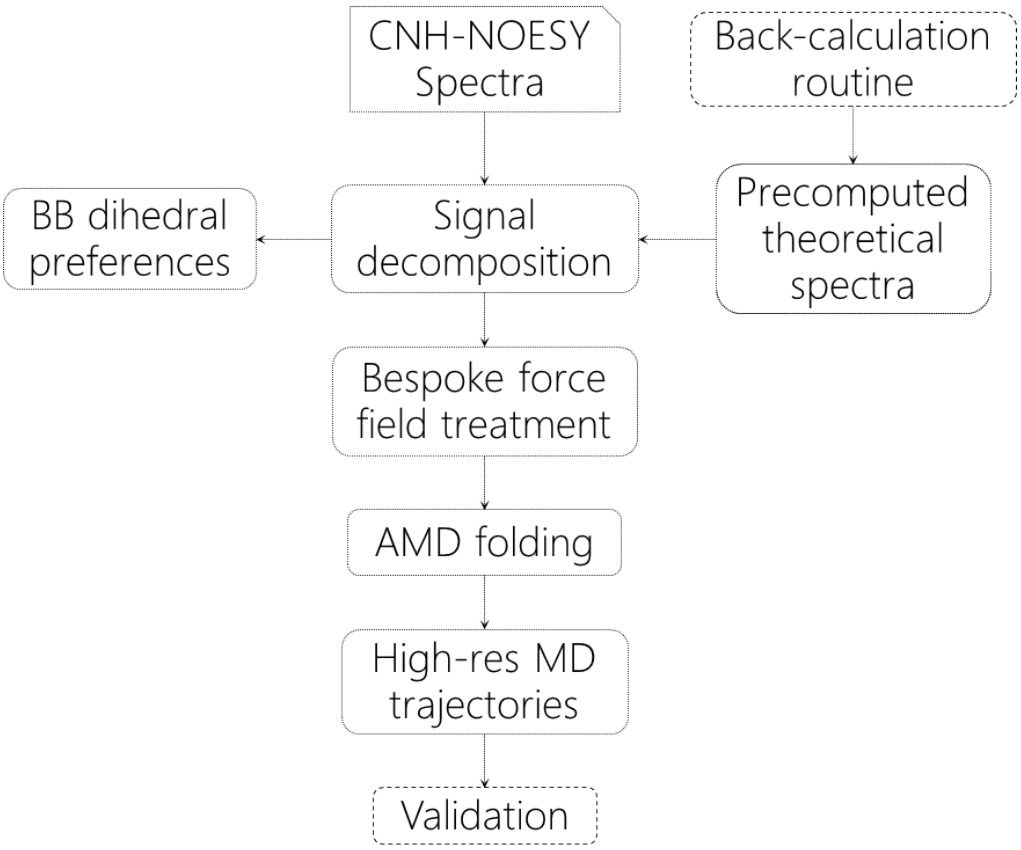


Figure 1.6: The general scheme of the CoMAND method. The protocol starts by the acquisition of a single 3D ^{13}C-HSQC-NOESY-^{13}N-HSQC experiment, from which spectral strips are extracted, each containing contacts to one backbone amide proton. The second step is to conduct the signal decomposition calculations using theoretical spectra that represent systematically sampled local conformers around the target proton. These matrices can be either computed on-the-fly or recovered from a precomputed database (with constant experimental parameters). Dotted boxes represent routine tasks carried out on every structure determination project. The third step entails the generation of local backbone potential energy surface grids that describe the relative free energy change associated with every (ϕ_i, ψ_{i-1}) combination at every residue. The latter energy maps are then embedded into a classical force field (CHARMM36[28] was the force field of choice here), where there smooth splines will serve as fold-guiding potentials. The fourth step is running accelerated molecular dynamics simulations (AMD) to build initial models that are purely based on physical restraints and experimentally-derived conformational distributions. The penultimate stage aims at generating highly refined ensembles that best explain the experimental spectra through performing equilibrium molecular dynamics (MD) simulations and subsequent frame selection. Finally, I attempt to validate the high-resolution ensembles by demonstrating their accuracy in explaining unrelated experimental data to that used in structure elucidation. Dashed boxes represent tasks that do not need to be carried out routinely. For example, theoretical spectra for fixed length peptide stretches of all possible sequences can be precomputed and stored as feature set matrices for reuse in every new project. Orthogonal validation of the method presented here on the other hand, once established for a set of benchmark test-cases, does not have to be carried out for every new project.

proton and all of the locally proximal carbon-bound protons within the conformationally sampled peptide.

- As it can be shown that matrix W, even if square, can never be non-singular, I formulate a solution that aims to maximise uniqueness. Eventually, I tackle special test-cases to robustly demonstrate the solution uniqueness.
- I then perform positive matrix factorisation of the experimental spectrum by the theoretical spectra, in order to solve for the combination of theoretical spectra and their relative weights vector (h) that best explains the mixture underlying the experimental spectrum. This should give accurate starting information on the local conformational preferences across the protein. For example, Figure 1.7C shows the corresponding factorisation result for hUb L67, where the recovered theoretical spectrum can explain the experimental spectrum down to a level close to the noise level.

2. Generate an experimentally derived, residue-wise, potential energy surface.

   - If this factorisation is conducted on a relative basis, it can offer the relative probability of every conformer to contribute to the observed ensemble. Therefore a multidimensional probability distribution the full conformational space can be derived.
   - I then convert this probability landscape, through a Boltzmann factor, into a residue-wise relative free energy landscape. Figure 1.7D shows the corresponding hUb L67 conformational energy map across the (ϕ_i, υ_i) space, compared against a high-resolution crystallographic value. In addition to yielding the right conformer as the global minimum, the full map should supply an amount of information that cannot be generated by any other structure determination method.

3. Construct bespoke force fields.

   - Since these energy maps can form the basis for initial model building, I sought to encode them into a classical mechanics force field that can later guide folding simulations for initial model building. The resulting supplemented force field would be bespoke on a residue-by-residue and protein-by-protein basis.

- The CHARMM36 force field is particularly suited for this since it already provides corrective backbone cross terms (CMAPs), that are derived from QM-level PMF simulations. This CMAP term was replaced by a bespoke residue CMAP, wherever an experimental NOESY spectrum was available for that residue.

4. Develop accelerated simulations for initial model building.

    - I then describe a guided-swarm accelerated scheme that combines the concepts of replica exchange and simulated annealing, with the goal of overcoming the extreme landscape ruggedness faced upon folding an entire protein. I call this purpose-built scheme SARS for Simulated Annealing Replica Seilschaft, owing to its rope team-like behaviour in exploring the potential energy landscape.
    - I demonstrate that this routine, with the help of the guiding energy maps, can successfully fold a linear peptide chain into a structural model that is a good starting point for generating refined ensembles.

5. Build highly refined ensemble.

    - Once an initial model is available, I equilibrate it, and run long, unbiased, canonical ensemble simulations. These would routinely generate tens of thousands of frames that represent microstates in thermal equilibrium. These trajectories would constitute the pool of microstates from which the final ensemble would be compiled by picking only a subset of few frames.
    - To avoid overfitting, a small number of frames corresponding to the experimental data signal-to-noise ratio should be finally selected to represent the solution ensemble. Assuming the canonical ensemble pool is k-frame-large, selecting the lowest R-factor subset of n members implies an $\binom{n}{k}$ computational complexity. I attempt to evaluate the convergence and performance of several approximations, that compile the lowest R-factor ensemble in an acceptable execution time. This is a form of collective optimisation, where all of the residue R-factors are considered in average, with the aim of minimising across the whole protein. Figure 1.7E shows the corresponding selected ensemble at L67 of hUb.

6. Validate the refined ensembles.

    - After an ensemble is compiled on the basis of minimising the disagreement between its average back-calculated spectra and the experimental NOESY spectra, its accuracy can be judged through using it to back-calculate other experimental quantities that are not derived from NOESY data; i.e. to carry out true prospective validation. Given that hUb has long served as a model system for NMR studies, extensive solution NMR datasets are available for this purpose.
    - I evaluate how can the compiled ensemble compares, in terms of correlation coefficients with experimental values, to the most accurate previously published ensembles with respect to reproducing the N-H bond order parameter $S^2_N H$, the $H_N H_\alpha$ scalar coupling constants ${}^3J_{H_N H_\alpha}$, and residual dipolar coupling constants (RDCs).

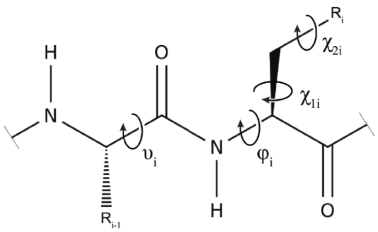
7. Decompose overlapped spectra.

    - Overlapped resonances in the ^{15}N and ^{1}H dimensions result in unsolvable cases of overlapped NOESY spectra (i.e. unassignable cross-peaks). This would constitute a particularly challenging test-case for the factorisation proposition delineated above. Here, I demonstrate that a standard two-component factorisation by a concatenated feature set matrix (i.e. $(W_i|W_j)$, for overlapped residues i and j) can trivially yield the correct conformers solutions.
    - I apply this to nine different residues from two different proteins and even demonstrate that the recovered initial solutions may still be used to generate secondary conformational distribution maps. While this may initially lose the long-range contact information as degenerate *noise*, once an initial model is built, this information can be still recovered within the folded model context.

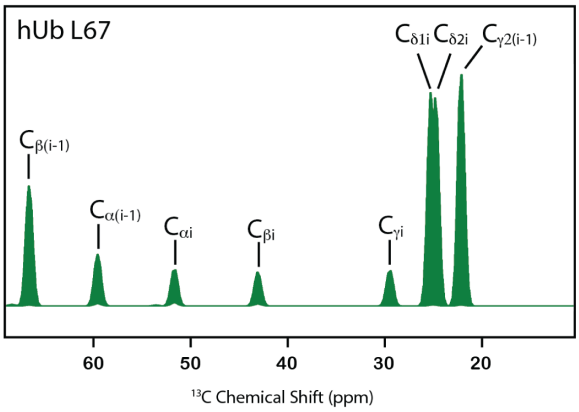
In the third chapter, I describe the method development research in detail, and move on to carry out the structure determination and refinement results for five different proteins of diverse topologies from different structure biology projects in our institute. Special emphasis is paid to the *de novo* human ubiquitin ensemble given this proteins benchmark status in the field.

A

Systematic Conformational Sampling



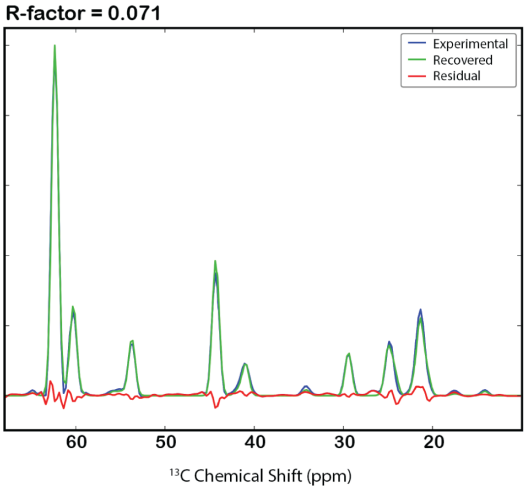
B



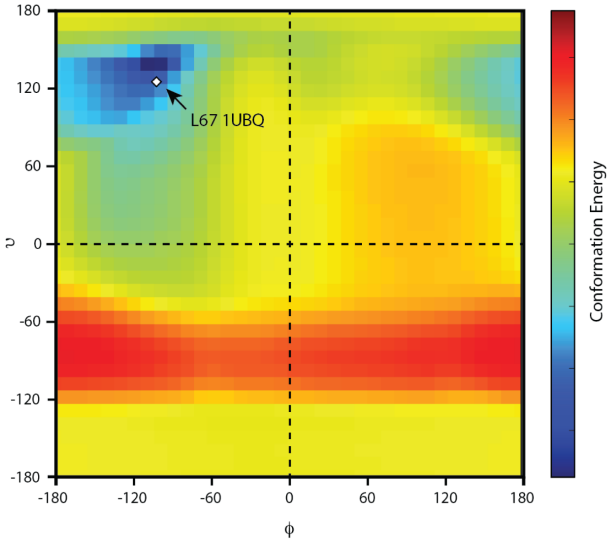
Matrix of Back-calculated Spectra

C

Spectral Decomposition and R-factor Calculation



D



De novo Conformational Mapping

E

Structure Ensemble

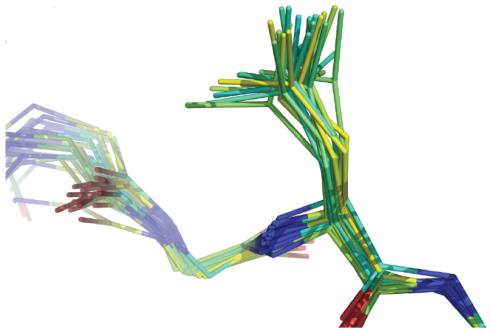


Figure 1.7: Illustration of the key steps along the CoMAND scheme. A) The sampled conformational degrees of freedom ($\phi_i, \upsilon_i, \chi_{i,1}, \chi_{i,2}$) for sequence position i (where υ_i is ψ_{i-1}). These rotamers were systematically sampled within a tripeptide fragment for every residue along the protein sequence. B) An overlay plot of the normalised back-calculated intensity against the ^{13}C chemical shift dimension for all the sampled conformers spectra for hUb L67, with a designation of the source for every cross-peak. C) A plot of the experimental spectrum, the recovered spectrum from the positive matrix factorisation, residual intensity across them, and the R-factor for hUb L67. D) The *de novo* generated conformational relative free energy map to the best two-component solution for the same residue. The white diamond shows the high-residue crystal structure (ϕ_i, υ_i) value (PDB: 1UBQ). E) A stick representation of the conformational preferences of the same residue in the final ensemble.

Chapter 2: Asymmetric Protein Design from Conserved Supersecondary Structures

Status: Under review

Asymmetric protein design from conserved supersecondary structures

Mohammad ElGamacy, Murray Coles, Andrei Lupas

Dept. of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

Corresponding author: Andrei Lupas
Department of Protein Evolution, Max-Planck-Institute for Developmental Biology. Max-Planck-Ring 5, D-72076 Tübingen, Germany
Tel.: +49-7071-601-341
Fax: +49-7071-601-349
E-mail: andrei.lupas@tuebingen.mpg.de

Keywords: Computational protein design, interface-driven strategy, conserved motifs, globular protein design.

Abbreviations: Abbreviations: HhH, Helix-hairpin-Helix; AP, Apurinic/apyrimidinic; dRP, deoxyribophosphate; PSSM, position-specific scoring matix; SMD, steered molecular dynamics; CD, circular dichroism.

Computational design with supersecondary structures as building blocks has proven effective in the construction of new proteins with controlled geometries. So far, this approach has primarily exploited amplification, effectively harnessing the internal folding propensity of self-compatible fragments to achieve sufficient enthalpy for folding. Here we exploit an interface-driven strategy to depart from the repeat design realm, constructing an asymmetric, globular domain from heterologous supersecondary structures. We report the successful design of a dRP lyase domain fold, which agrees with the experimental NMR structure at atomic accuracy (backbone RMSD of 0.94 Å). Our results show that the residual folding information within conserved fragments, combined with efficient interface-directed sampling, can effectively yield globular proteins with novel sequences and biophysical properties.

1. Introduction

Many proteins domains show clear internal symmetries in their structure and sequence, suggesting they have evolved by amplification of subdomain-sized fragments. Such re-use of small, self-compatible structural motifs provides an efficient means of evolving folds via a few tandem repetition steps. Repetition on the sequence level introduces obligate symmetries on the structural level, manifesting as pure rotational symmetries for closed toroids, such as TIM-barrels and β-propellers, or in combination with translational symmetries for open-ended solenoids, such as TPR and armadillo repeats (Söding and Lupas, 2003). This evolutionary mechanism has inspired numerous research efforts aimed at designing proteins from conserved supersecondary structures, thus generating novel sequences with new properties, while maintaining the parent topology (Höcker, 2014).

To date, protein design efforts have been centred on this paradigm of repetition, effectively extracting the conserved sequence determinants that underlie fragment self-compatibility. These minimal patterns of intramolecular interactions can then be further optimised through energy-based scoring functions to reach more stable, idealised forms (Parmeggiani and Huang, 2017; Parmeggiani et al., 2015). While constituting a robust means for consensus design, the power of this approach fundamentally rests on the amplification of the folding enthalpies: by a factor of n in case of closed toroids, and by a factor of $(n-1)$ in open-ended solenoids, where n is the number of repeats. The free energy of interaction between repeats has been well established as the dominant term in folding thermodynamics for a wide range of such folds, and has been accurately recapitulated by a 1D nearest-neighbour model (Kajander et al., 2005; Kloss et al., 2008). However, this dominance and the existence of multiple similar interfaces can lead to design failures via unintended associations, manifesting themselves as oligomeric self-assemblies, structural plasticity and domain swapping (Parmeggiani and Huang, 2017; Voet et al., 2014).

In this work, we aim to depart from the repetition paradigm by designing a globular domain from two unrelated supersecondary structures, which excludes any element of symmetry from the design. This constraint of asymmetry eliminates the reliance on self-compatibility of the motifs, as it implies that a single interface between the two constituent fragments will be the sole enthalpic driving force for folding. Our strategy is based on two assumptions. The first is that the sequences of subdomain-sized fragments still retain a fraction of the information required for spontaneous folding, even though they may not possess sufficient internal interactions to do so. Such information content increases toward the centre of sequence clusters built from homologous proteins, and thus favours the choice of the best-conserved building blocks (Porebski and Buckle, 2016; Wheeler et al., 2016). The second assumption is that the sampling problem can be simplified by restricting sequence optimisation to just the inter-motif interface. This enables the computational sequence and conformer sampling to be conducted at a very fine granularity, but also introduces minimal disruption of the structurally important features of the motifs.

Our target fold is the dRP lyase domain of the human DNA polymerase beta, which consists of a helical hairpin motif (αα-hairpin) and a helix-hairpin-helix (HhH) motif (figure 1). Of the two, only the HhH-motif is widespread across many protein families (Alva et al., 2015). Our goal was therefore to replace the αα-hairpin, which is restricted to the dRP lyase family, with a ubiquitous αα-hairpin conserved across different families. Here we show that this approach can yield a more stable version of the fold, with modest sequence perturbation and a design precision at the atomic level.

2. Materials and Methods

2.1. Computational design

As a starting point, an HhH-motif was obtained from the dRP lyase domain of human polymerase beta (PDB: 4KLI). A diverse set of starting αα-hairpins belonging to the a.118 fold of the SCOPe database was docked against this to best align to the full dRP lyase domain. Since mutagenesis was restricted to the αα-hairpin, the variable sequence positions were restricted to the new interface, and sampled amino acid types were defined by the fragment's sequence profile and the position's solvent exposure. The close-vicinity sequence profile of the αα-hairpins was constructed by running the fragments' sequences through BLAST+ (Camacho et al., 2009) against the nr database, keeping those above 60% identity to the query, and then clustering (Li and Godzik, 2006) them at 95% identity. PSSMs were constructed using PSI-BLAST (Schäffer et al., 2001) and a lower log-odds threshold of 0.0 was used to build the sequence sampling restraints for Rosetta, which were appended with extra polar or apolar amino acid identities, depending on the expected amino acid solvent exposure.

The sequence and conformer sampling for the design of the new interface was performed via RosettaScripts (Fleishman et al., 2011). The protocol comprised a single generic Monte Carlo loop of 6 cycles, optimising for the energy per residue using the talaris2013 scoring function (Leaver-Fay et al., 2013), further filtering of decoys was done using the packstat score (Sheffler and Baker, 2009). Each cycle executed a routine comprising soft-repulsion sequence sampling, backbone optimisation (Smith and Kortemme, 2008), and FastRelax conformational refinement (Tyka et al., 2011). The output was filtered through an accelerated steered molecular dynamics routine that aims at approximately assessing the potential of mean force (PMF) of unbinding across the designed interface. The free energy of unbinding (W) was evaluated as $W_{t_o \to t_e} = \boldsymbol{v} \int_{t_o}^{t_e} \boldsymbol{F}(t)dt$ where $\boldsymbol{F}(t)$ and $\boldsymbol{v}$ are the pulling force and velocity vector at time t and the constant pulling velocity (where $\boldsymbol{v} \approx \boldsymbol{v}_{ref}$), respectively. All decoys were aligned against a reference orientation of the HhH-motif, while the other motif was pulled along a single dimension through a stiff spring to achieve a constant-velocity, variable-force steering setup that yields the free energy profile along the unbinding path. The protein was modelled using the CHARMM36 force field (Best et al., 2012), the simulations were performed in explicit solvent (TIP3P water model) and 0.15 M sodium chloride as NPT ensembles at 310 K and 1 atmosphere using a Langevin thermostat and a Langevin barostat as implemented in the NAMD engine (Phillips et al., 2005). Particle Mesh Ewald electrostatics grid of 1 Å was used with a long-range cutoff set at 12 Å (switching at 10 Å) and a timestep of 2 fs. The reference pulling velocity ($\boldsymbol{v}_{ref}$) was calibrated to 2 Å/ns with a spring constant (k) of 50 kcal·mol^{-1}·Å^{-2} where the applied force ($\boldsymbol{F}(t)$) was computed as $-\nabla[\frac{1}{2}k[t\boldsymbol{v}_{ref} - (\boldsymbol{r_t} - \boldsymbol{r_o}) \cdot \boldsymbol{n}]^2]$ ($\boldsymbol{r_t}$ being the position vector of the steered atom group and $\boldsymbol{n}$ being the pulling direction vector). The systems underwent 2000 steps of conjugate gradient minimisation before random initialisation of atom velocities and force application on the backbone carbonyl carbon atoms within 10 angstroms the motif centre-of-mass. The calculated work was used to rank designs for the next stage.

The final stage was to conduct the Rosetta *ab initio* structure prediction calculations at full-atom detail (Raman et al., 2009). This was performed for 30 different designs, from which the top four were chosen as the most well funnelled folding trajectories (Table 1, Figure 2).

2.2. Expression and purification

The genes were acquired from Eurofin Genomics, cloned into pETHIS-a using NcoI and XhoI cloning sites and in-frame with an N-terminal hexaHis-tag and a TEV cleavage site, while harbouring a kanamycin resistance gene as a selection marker. The plasmids were used to transform chemically competent *E. coli* BL21(DE3) by means of heat-shock. The expression procedure entailed growing of the cells in LB medium and inducing with IPTG at OD_{600} of 0.5~1 with overnight expression at 25 °C. For expression of labelled protein, a preculture in LB medium was grown, cells collected, washed twice in PBS buffer, and resuspended in M9 minimal medium (240 mM Na_2HPO_4, 110 mM KH_2PO_4, 43 mM NaCl), supplemented with 10 µM $FeSO_4$, 0.4 µM H_3BO_3, 10 nM $CuSO_4$, 10 nM $ZnSO_4$, 80 nM $MnCl_2$, 30 nM $CoCl_2$ and 38 µM kanamycin sulfate, to an OD_{600} of 0.5~1. After 40 minutes of incubation at 25 °C, 2.0 gm ^{15}N-labelled ammonium chloride (Sigma-Aldrich cat.nr. 299251) and 6.25 gm ^{13}C D-glucose (Cambridge Isotope Laboratories, Inc. cat.nr. CLM-1396) were added in a 2.5 L culture. After another 40 minutes IPTG was added to 1 mM final concentration for overnight expression. Cells were collected by centrifugation at 5,000 g for 15 minutes, lysed by a Branson Sonifier S-250 (Fisher Scientific) in hypotonic 50 mM Tris-HCl buffer supplemented with one tablet of the cOmplete protease cocktail (Sigma-Aldrich cat.nr. 4693159001) and 3 mg of lyophilised DNase I (5200 U/mg; Applichem cat.nr. A3778). The insoluble fraction was pelleted by 25,000 *g* centrifugation for 50 minutes, and the soluble fraction was filtered (0.45 µm filter pore size) and directly applied to a Ni-NTA column. A 5 mL HisTrapFF immobilised nickel column (GE Healthcare Life Sciences cat.nr. 17-5255-01) was used for this purpose, washed consecutively by 30 mL 150 mM NaCl, 30 mM Tris buffer (pH 8.5) at 0, 30 and 60 mM imidazole. Fractions were collected by a gradient elution at > 60 mM imidazole. The eluate was concentrated using 3 kDa MWCO centrifugal filters (Merck Millipore cat.nr. UFC901024) and loaded onto an equilibrated Superdex 75 gel filtration column (GE Healthcare Life Sciences cat.nr. 17517401). The gel filtration buffer used was always 100 mM sodium phosphate buffer (for NMR and CD transparency) composed to a pH of 8.0. An ÄktaFPLC system (GE Healthcare Life Sciences) was used for all chromatography runs.

2.3. Biophysical characterisation

The analytical gel filtration experiments were all done on a Superdex 75 10/300 GL (GE Healthcare Life Sciences cat.nr 17517501), and the collected fractions from the eluate were used directly for CD or NMR measurements. ^{1}H NMR spectra were collected on Bruker AVIII-800 or Bruker AVIII-600 spectrometers. CD spectra were recorded on a Jasco J-810 spectrometer, with a spectral scan window of 200-240 nm, with a sweep delta of 0.1 nm while averaging over 5 scans. Melting curves were measured from 20 to 100 °C, recording the ellipticity at 222 nm every 0.5 °C, while heating at a 1 °C/min rate.

2.4. NMR structure determination

All spectra were recorded at 313 K on Bruker AVIII-600 and AVIII-800 spectrometers. Backbone sequential and aliphatic sidechain assignments were completed using standard triple resonance experiments, while aromatic assignments were made by linking aromatic spin systems to the respective $C^{\beta}H_2$ protons in a 2D-NOESY spectrum. Distance data were derived from a set of five 3D-NOESY spectra, including the heteronuclear edited NNH-, CCH-, and CNH-NOESY spectra (Dierks et al., 1999) in addition to conventional ^{15}N- and ^{13}C-HSQC-NOESY spectra. A ^{12}C-filtered 2D-NOESY spectrum was recorded for the observation of contacts to aromatic groups. Backbone dihedral angle restraints were derived

using the TALOS-N server (Cornilescu et al., 1999). Sidechain rotamers were assessed using an HNHB experiment. Hydrogen bond restraints were applied for amide protons that were protected from solvent exchange and where acceptors were consistently identified in preliminary calculations. These were applied in the simulated annealing calculations as pseudo-covalent bonds.

Refinement was carried out by comparing experimental and back-calculated NOESY spectra using in-house software. 1D strips were back calculated for the amide protons of all ordered atoms, plus selected sidechain groups. These were compared to the experimental spectra, yielding backbone and sidechain dihedral angles. As the intensities of intra-residue and sequential NOESY cross peaks are evaluated in this process, only medium and long-range contacts were further considered as distance restraints. Structures were calculated with XPLOR (NIH version 2.9.4) using standard protocols with modifications for the inclusion of hydrogen bonds as pseudo-covalent bonds. Non-bonded parameters were updated to match those used in the MOLPROBITY structure quality evaluation suite (Chen et al., 2010; Davis et al., 2007). For the final ensemble, 50 structures were calculated and 21 chosen on the basis of lowest restraint violations. An average structure was calculated and regularized to give a structure representative of the ensemble. Details of the input data and the final ensemble are given in Table 2. The coordinates for the structure were deposited into the Protein Data Bank with accession: 6H5H.

3. Results

3.1. Novel dRP lyase-like domain assembly

Apurinic/apyrimidinic (AP) sites, in addition to being a form of DNA damage, constitute a major intermediate product along base excision repair pathways. The repair of these sites is performed via 5' incision, deoxyribophosphate (dRP) excision, DNA extension, and ligation. Polymerase beta participates in one such pathway by conducting the DNA extension, through a typical arrangement of finger, palm and thumb domains that are collectively responsible for the DNA polymerase activity. Additionally, the polymerase possesses an 8 kDa N-terminal domain capable of carrying out the dRP excision step; the dRP lyase domain (Matsumoto and Kim, 1995). This domain is comprised of an αα-motif and an HhH-motif, and mutational analysis has identified two essential residues for the dRP lyase acitivity: K72 and Y39. The actual catalytic role as the Schiff base mediating the β-elimination reaction was shown to be played by K72 (on the HhH-motif), while Y39 (on the αα-hairpin) structurally stacks against the AP site (Matsumoto et al., 1998). Here we sought to reconstruct the dRP domain by combining this ancestral HhH-motif with a heterologous ancestral αα-hairpin obtained from a TPR-like fold. The main goal of the design process was to obtain a new hydrophobic core at the interface between these motifs (Figure 1C). This was performed with the additional constraint of keeping the HhH-motif, which harbours the catalytic centre, compositionally fixed, in order to enable future mechanistic and functional work.

3.2. Computational cross-fragment interface design

To keep the HhH-motif constant, we performed one-against-many docking; a single HhH-motif against a set of ancestral TPR-like αα-hairpins. The docking stage aims at conformationally refining the initial models under an RMSD constraint to the target topology, with the reference being the human dRP lyase domain. In order to avoid sequence

deviations far from the starting fragments, mutational sampling bias was imposed towards the existing sequence diversity within the close-vicinity profiles of the αα-hairpin. This was done by building and extracting the information from position-specific scoring matrices (PSSMs). While sampling for the HhH-motif was restricted to conformational refinement, sequence sampling and conformational sampling were conducted on the αα-hairpin to minimise the inter-motif interaction energy. The main sampling routine thus comprised a combination of sequence sampling, sidechain rotamer sampling, backbone refinement, and rigid-body docking, which were performed with an initially softened steric repulsion term. The generated decoys were filtered progressively, with each stage involving more computationally intensive criteria. At first, these iterations were filtered by measures of atomic packing quality and total energy. Secondly - with the goal of more accurately estimating the interaction free energy of the designed interfaces - potential-of-mean-force calculations were conducted through constant-velocity, variable-force steered molecular dynamics (SMD) simulations (Materials and Methods). The interaction free energy was calculated from the velocity-scaled integral of the force as a function of time, and was used to rank the candidates accepted for the next stage. The final filtering stage was conducted through template-free structure prediction of the designed sequences using Rosetta *ab initio* folding simulations. The aim was to select the designs for which the predicted decoys with the lowest energy score have the lowest backbone RMSD from the design coordinates. Thirty candidates have undergone such sampling, and the top four designs with the most funnel-shaped folding landscapes were finally chosen for experimental evaluation (Figure 2, sequences in Table 1).

3.3. Biophysical properties

All of the four selected designs were expressed and soluble in *Escherichia coli* to varying extents. Two were successfully purified in large quantities and appeared to be monomeric. The fourth design (polb4) appeared to be stably folded and purely monomeric, as evidenced by analytical size-exclusion chromatography (Figure 3A). Circular Dichroism (CD) experiments and 1D NMR spectra showed polb4 to be of a strongly alpha helical character (Figure 3B) and to possess good NMR dispersion. Thermal unfolding experiments of polb4 resulted in a single-phase equilibrium unfolding transition at 72 °C (Figure 3C), which is conducive of a stable single domain behaviour. The van't Hoff fit for polb4 (Figure 3D) indicated an unfolding free energy of 34.8 kJ/mol (at 20 °C), with an underlying enthalpy change of 231.1 kJ/mol and an entropy change of 0.7 kJ/mol·K, assuming a two-state model. We have also compared the folding stability of the designed domain against the wild type domain with the added C-terminal sequence: KLRKLEKIRQDDTSSSINFLT, which is the extra C-terminal sequence included in the PMF simulations. The results showed a melting transition for the wild type at 46 °C compared to 77.8 °C for the design. Moreover, at the same concentration, the design shows more than twice the ellipticity of the wild type (Figure 4). The low entropy value, combined with the high free energy of unfolding for polb4, is indicative of very favourable thermodynamics for polb4. Given its high expression level, 1D ^{1}H spectral dispersion and its apparent monomeric status, we sought to solve the structure of polb4 using NMR spectroscopy.

3.4. Solution structure of polb4

The compiled NMR ensemble consisted of 20 frames and was very focused, with a backbone RMSD from the average structure stood at 0.28 ±0.08 Å. This conformational homogeneity was matched to atomic agreement between the experimental structure ensemble and the designed coordinates, where the average backbone RMSD between the

ensemble and the design was 0.94 Å (Figure 5A). Most of that deviation stems from the slight departure of the experimental structure's C-terminal helix in comparison to the design. The design appears to be more densely packed, which may be expected, as we compare the thermalised solution structure to the energy minimised design coordinates. Inspecting the side chain rotameric accuracy at the interface shows that 14 out of 16 side chain rotamers were correctly predicted by the design model, with an average all-atom RMSD distance of 1.4 Å between the structure and the design (Figure 5B).

4. Discussion

A significant level of internal redundancy has been shown to exist across natural proteins, where similar subdomain-sized fragments are scattered across different folds (Alva et al., 2015; Söding and Lupas, 2003). The detectable homology among some of these fragments has led to the proposal that illegitimate recombination of DNA fragments encoding them represents a parsimonious mechanism of fold evolution. The natural re-use of such conserved motifs within different structural contexts suggests their potential robustness as building blocks that are optimized for folding. Several previous protein design studies have successfully utilised such supersecondary structures to build scaffolds with diversified biophysical properties (Höcker, 2014), however, they were all exploiting repetition (on the structure or sequence level), yielding pseudo-symmetric repeat proteins with sizes that are multiples of their unique constituting fragments. This repetition easily amplifies the folding enthalpies arising from inter-fragment contacts, rendering these designs more attainable.

Here we progress to construct an asymmetric fold with a single inter-fragment interface, consisting of two heterologous supersecondary structures. Whereas the HhH-motif has homologues across a wide range of protein families, the αα-hairpin of dRP lyase is limited to this family. This prompted us to study whether the latter could be replaced by an αα-hairpin that is equally wide-spread across numerous protein families, based on the premise that highly conserved sequences have more residual folding information, and thus are more robust to sequence optimisation in different contexts (Porebski and Buckle, 2016; Wheeler et al., 2016). Since this was an effort to replace one fragment by another, we limited the sequence design to the inbound fragment, while keeping the rest of the domain constant.

Such non-repetitive configurations pose a significant challenge, as the sampled folding enthalpies stem only from a single inter-fragment interface. This represents a shallow potential energy basin compared to repetitive architectures. It has been previously shown that computationally designing large interface surface areas and incorporation of large, flexible side chain residues across protein-protein interfaces becomes prohibitive due to the associated breadth of sampling (Stranges and Kuhlman, 2013). We therefore propose that, despite the shallowness of the basins involved, small interfaces nevertheless provide a narrow search space that allows for extensive rotameric sampling at a fine granularity, and that more accurate estimations of their binding free energies can improve the design success rate. The advantage brought by such rigorous sampling not only averts the poor packing obtained from coarse-grained rotameric sampling, but also greatly enhances the design precision, as emphasized by the highly accurate prediction of rotameric states achieved in our design.

Acknowledgements

The authors thank Eva Hertle and Jörg Martin for assistance with the biochemical experiments, and Vincent Truffault for support of the NMR measurements. This work was supported by institutional funds from the Max Planck Society.

References

Alva, V., Söding, J., and Lupas, A.N., 2015. A vocabulary of ancient peptides at the origin of folded proteins. eLife *4*, e09410.

Best, R.B., Zhu, X., Shim, J., Lopes, P.E.M., Mittal, J., Feig, M., and MacKerell, A.D., 2012. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. J. Chem. Theory Comput. *8*, 3257-3273.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. *10*, 421-421.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D: Biol. Crystallogr. *66*, 12-21.

Cornilescu, G., Delaglio, F., and Bax, A., 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR *13*, 289-302.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., Richardson, D.C., 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. *35*, W375-W383.

Dierks, T., Coles, M., and Kessler, H., 1999. An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. J. Biomol. NMR *15*, 177-180.

Fleishman, S.J., Leaver-Fay, A., Corn, J.E., Strauch, E.-M., Khare, S.D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., Baker, D., 2011. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLoS ONE *6*, e20161.

Höcker, B., 2014. Design of proteins from smaller fragments—learning from evolution. Curr. Opin. Struct. Biol. *27*, 56-62.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L., 2005. A New Folding Paradigm for Repeat Proteins. J. Am. Chem. Soc. *127*, 10188-10190.

Kloss, E., Courtemanche, N., and Barrick, D., 2008. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. Arch. Biochem Biophys. *469*, 83-99.

Leaver-Fay, A., O'Meara, M.J., Tyka, M., Jacak, R., Song, Y., Kellogg, E.H., Thompson, J., Davis, I.W., Pache, R.A., Lyskov, S., Grey, J.G., Kortemme, T., Richardson, J.S., Havranek, J.J., Snoeyink, J., Baker, D., Kuhlman, B., 2013. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. Methods Enzymol. *523*, 109-143.

Li, W., and Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Matsumoto, Y., and Kim, K.,1995. Excision of deoxyribose phosphate residues by DNA polymerase beta during DNA repair. Science *269*, 699.

Matsumoto, Y., Kim, K., Katz, D.S., and Feng, J.-a.,1998. Catalytic Center of DNA Polymerase β for Excision of Deoxyribose Phosphate Groups. Biochemistry *37*, 6456-6464.

Parmeggiani, F., and Huang, P.-S., 2017. Designing repeat proteins: a modular approach to protein design. Curr. Opin. Struct. Biol. *45*, 116-123.

Parmeggiani, F., Huang, P.-S., Vorobiev, S., Xiao, R., Park, K., Caprari, S., Su, M., Jayaraman, S., Mao, L., Janjua, H., Montelione, G.T., Hunt, J., Baker, D., 2015. A General Computational Approach for Repeat Protein Design. J. Mol. Biol. *427*, 563-575.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K., 2005. Scalable Molecular Dynamics with NAMD. J. Comput. Chem. *26*, 1781-1802.

Porebski, B.T., and Buckle, A.M., 2016. Consensus protein design. Protein Eng, Des. & Sel. *29*, 245-251.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B.H., Das, R., Grishin, N.V., Baker, D., 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins *77*, 89-99.

Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F., 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. *29*, 2994-3005.

Sheffler, W., and Baker, D., 2009. RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. *18*, 229-239.

Smith, C.A., and Kortemme, T., 2008. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol Biol. *380*, 742-756.

Söding, J., and Lupas, A.N., 2003. More than the sum of their parts: On the evolution of proteins from peptides. BioEssays *25*, 837-846.

Stranges, P.B., and Kuhlman, B., 2013. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Sci. *22*, 74-82.

Tyka, M.D., Keedy, D.A., André, I., DiMaio, F., Song, Y., Richardson, D.C., Richardsonb, J.S., and Baker, D., 2011. Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607-618.

Voet, A.R.D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S.-Y., Zhang, K.Y.J., and Tame, J.R.H., 2014. Computational design of a self-assembling symmetrical β-propeller protein. PNAS. *111*, 15102-15107.

Wheeler, L.C., Lim, S.A., Marqusee, S., and Harms, M.J., 2016. The thermostability and specificity of ancient proteins. Curr. Opin. Struct. Biol. *38*, 37-43.

Tables

Table 1. Experimentally tested design sequences and their starting fragments

```
polb1          TLNGVIKNMLVEEAKLAEQAERYDDMAVAMKAVTVIAKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATG
frag1(2o02)    ----MDKNELVQKAKLAEQAERYDDMAACMKSVTEQG--------------------------------
                   : ** **::***************..**:**

polb2          TLNGAIATMLAELARYAFNNQWWDLSVQEAIAAKVLAKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATG
frag2(1QSA)    ---SKSKTEQAQLARYAFNNQWWDLSVQATIAGKLWD---------------------------------
                  .   *  *:***************** :**.*:

polb3          TLNAFVASMLVEIANALRRVGDERDATTYLIAACKVGKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATG
frag3(1wy6)    ---EVSASILVAIANALRRVGDERDATTLLIEACKKG---------------------------------
                     **:** ***************** ** *** *

polb4          TLNGALVNMLKEEGNKALSVGNIDDALQYYAAAITLDKYPHKIKSGAEAKKLPGVGTKIAEKIDEFLATG
frag4(1elw)    ----EQVNELKEKGNKALSVGNIDDALQCYSEAIKLD---------------------------------
                     ** ***:***************** *: **.**
```

Table 2. Solution structure statistics

	SA	$\langle SA \rangle_r$
Restraint Violations[1]		
Distance restraints (Å)		
All (139)	0.008 ± 0.001	0.006
Medium range (9)	0.016 ± 0.001	0.015
Long range (79)	0.009 ± 0.001	0.006
H-bond (51)	0.000 ± 0.000	0.000
Persistent viol. thres.[2]	0.038	-
Dihedral restraints (°)		
All (181)	0.14 ± 0.006	0.14
Persistent viol. thres.[2]	0.65	-
H-bond restraints[3]		
Distance (Å) (51)	2.07 ± 0.11	2.02 ± 0.03
Antecedent angle (°)	15.2 ± 5.4	16.2 ± 4.3
Covalent Geometry		
Bonds ($Å \times 10^{-3}$)	2.05 ± 0.03	1.90
Angles (°)	0.56 ± 0.01	0.56
Impropers (°)	1.53 ± 0.03	1.52
Structure Quality Indicators[4]		
Ramachandran Map (%)	100.0 / 0.0 / 0.0	100.0 / 0.0 / 0.0
Atomic R.M.S.D (Å)[5]		
	Backbone Heavy Atom	All Heavy Atom
SA vs $\langle SA \rangle$	0.28 ± 0.08	0.80 ± 0.09
SA vs $\langle SA \rangle_r$	0.42 ± 0.14	1.06 ± 0.09
$\langle SA \rangle$ vs $\langle SA \rangle_r$	0.32	0.82

[1] Violations are expressed as RMSD ± SD unless otherwise stated. Numbers in brackets indicate the number of restraints of each type.

[2] Persistent violations are defined as those occurring in at least 75% of all structures. The thresholds at which no persistent violations occur are tabulated.

[3] Hydrogen bonds were treated as pseudo-covalent bonds. Deviations are expressed as the average distance/average deviation from linearity for restrained hydrogen bonds.

[4] Defined as the percentage of residues in the favored/allowed/outlier regions of the Ramachandran map as determined by MOLPROBITY (Chen et al., 2010; Davis et al., 2007).

[5] Structures are labeled as follows: SA, the final set of 21 simulated annealing structures; <SA>, the mean structure calculated by averaging the coordinates of SA structures after fitting over secondary structure elements; $<SA>_r$, the structure obtained by regularizing the mean structure under experimental restraints. RMSD values were obtained based on superimpositions over ordered residue (defined as E1-G71).

Figures

Figure 1. Stages of asymmetric globular protein design from conserved supersecondary structures through an interface-driven strategy. The top-left pane shows our initial target fold, which is of a human DNA polymerase beta dRP lyase domain. The top-right pane shows the first stage of diverse fragment selection, where we select motifs from heterologous and unrelated domains. This is followed by an arbitrary rigid-body optimisation and conformational refinement that precedes the total interface design procedure (bottom-left pane). The bottom-right pane shows the final output design coordinates (polb4).

Figure 2. Experimentally tested design coordinates and their Rosetta *ab initio* folding funnels. Top pane shows a cartoon representation of the designs of polb1, polb2, polb3 and polb4. Bottom pane shows the corresponding Rosetta *ab initio* folding predictions scores (in Rosetta energy units) against the root mean square deviation from the design backbone coordinates.

Figure 3. The hydrodynamic and thermodynamic properties of polb4. (A) Size exclusion chromatography shows polb4 to be almost exclusively monomeric. Gray line shows hydrodynamic markers designating (1) Elution void volume, and globular proteins of molecular weights: (2) 75 kDa, (3) 29 kDa, and (4) 13.7 kDa. Solid line shows polb4 elution peak, which corresponds to its expected monomeric state. (B) Circular dichroism spectrum polb4 showing an ellipticity pattern of a majorly helical protein. (C) Equilibrium melting curve as a function of ellipiticity at 222 nm, with melting transition inflection at 72 °C. (D) The van't Hoff fit explains the unfolding transition well and estimates a unfolding free energy change of 34.8 kJ/mol at 20 °C (unfolding enthalpy and entropy changes are shown for the linear fit).

Figure 4. Polb4 is significantly more stable than the wild type. (A) Melting curve for the wildtype dRP lyase domain with the C-terminal extension KLRKLEKIRQDDTSSSINFLT. The inset CD spectrum shows a weak helical signal, with melting transition inflection at 46 °C. (B) Melting curve for the corresponding polb4 construct, with melting inflection at 77.8 °C, and much stronger helical character (inset).

Figure 5. NMR structure of polb4 and the conformational precision of its design coordinates. (A) The experimental NMR structure ensemble (purple ribbons; Protein Data Bank accession: 6H5H) and the design coordinates (orange cartoon). The overlay shows the highly focused structure of NMR ensemble with a backbone RMSD of 0.28 Å from the average structure, while the average RMSD distance of ensemble from the design was 0.94 Å. (B) The designed side chain rotamers at the interface match the experimental structure very well; 14 out of 16 rotamers were correctly predicted by the design model.

Figure1

Click here to download high resolution image

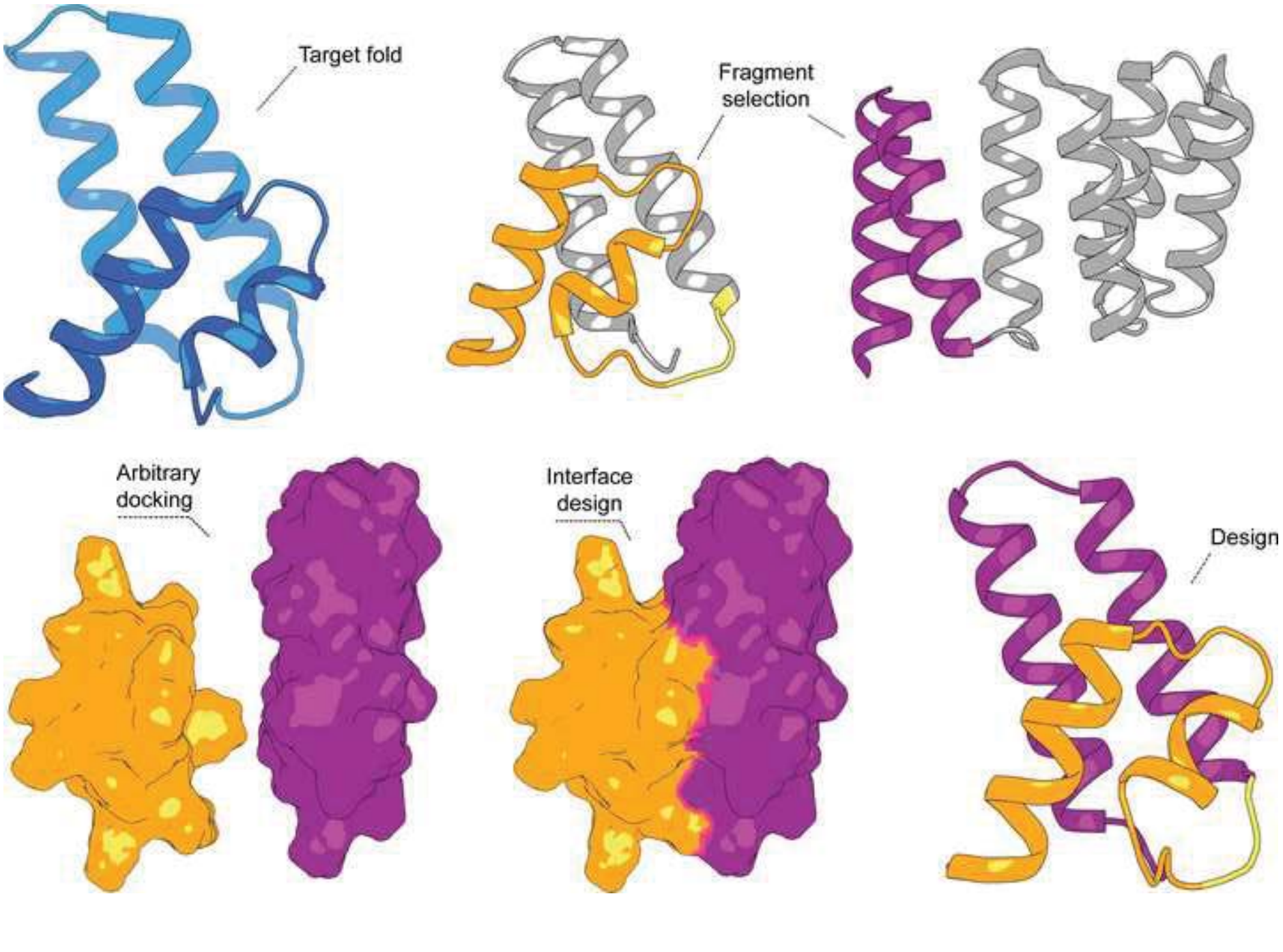


Figure2

Click here to download high resolution image

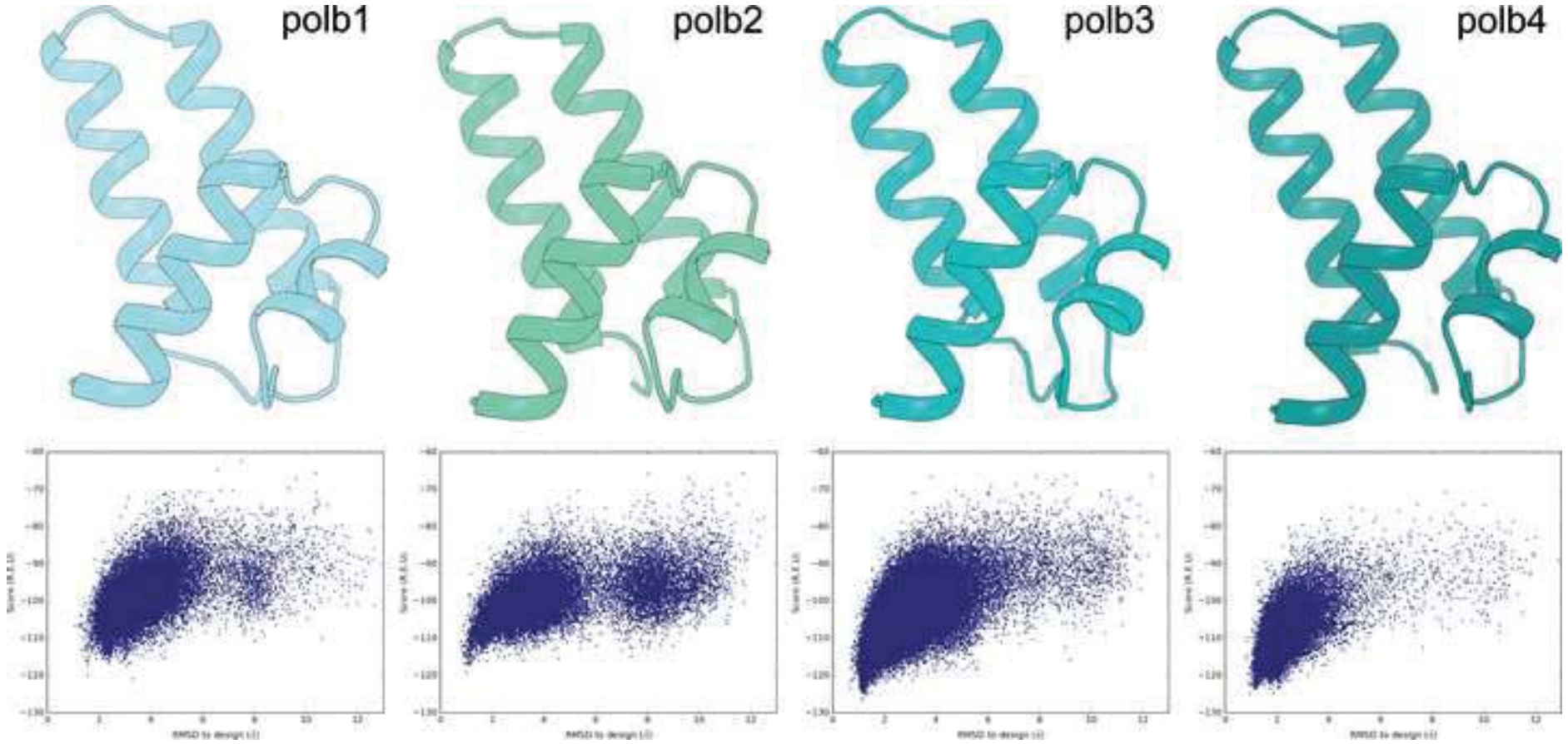


Figure3

Click here to download high resolution image

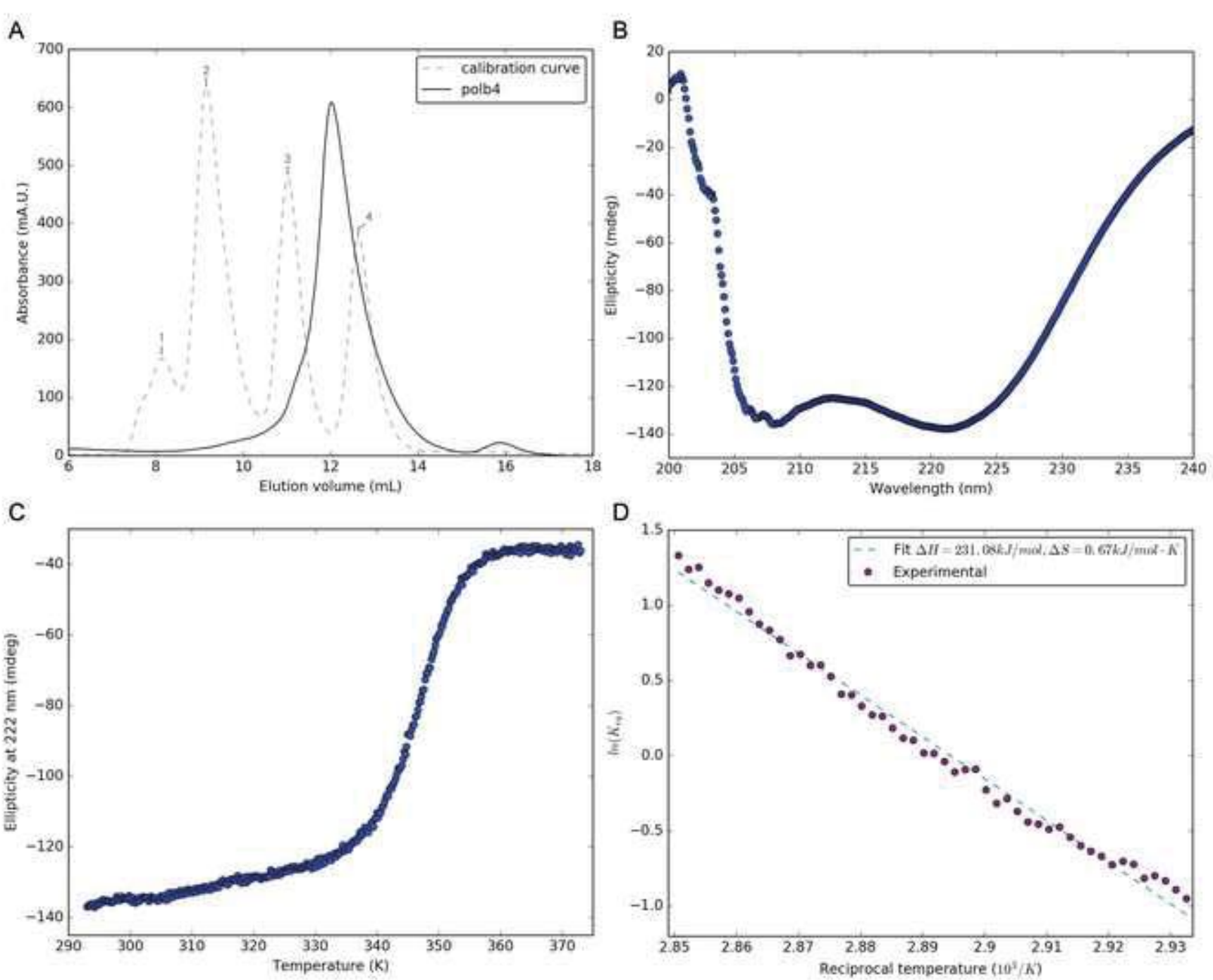


Figure4

Click here to download high resolution image

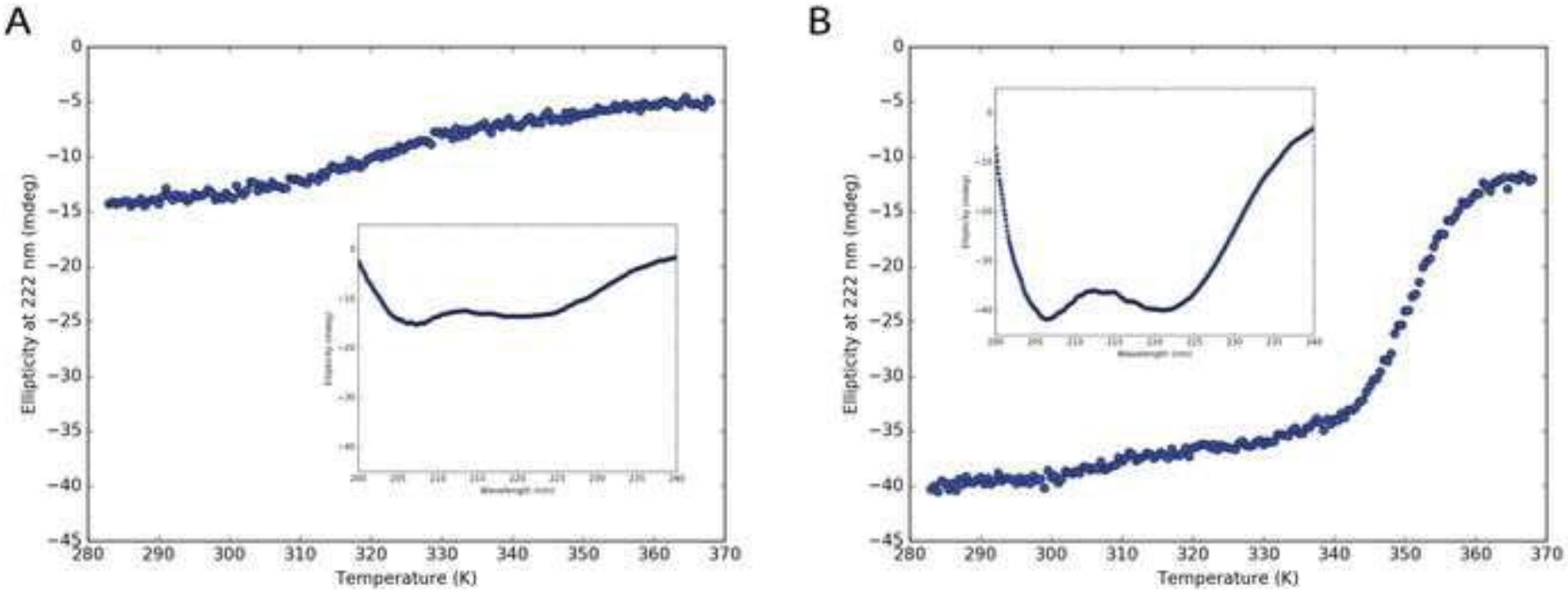
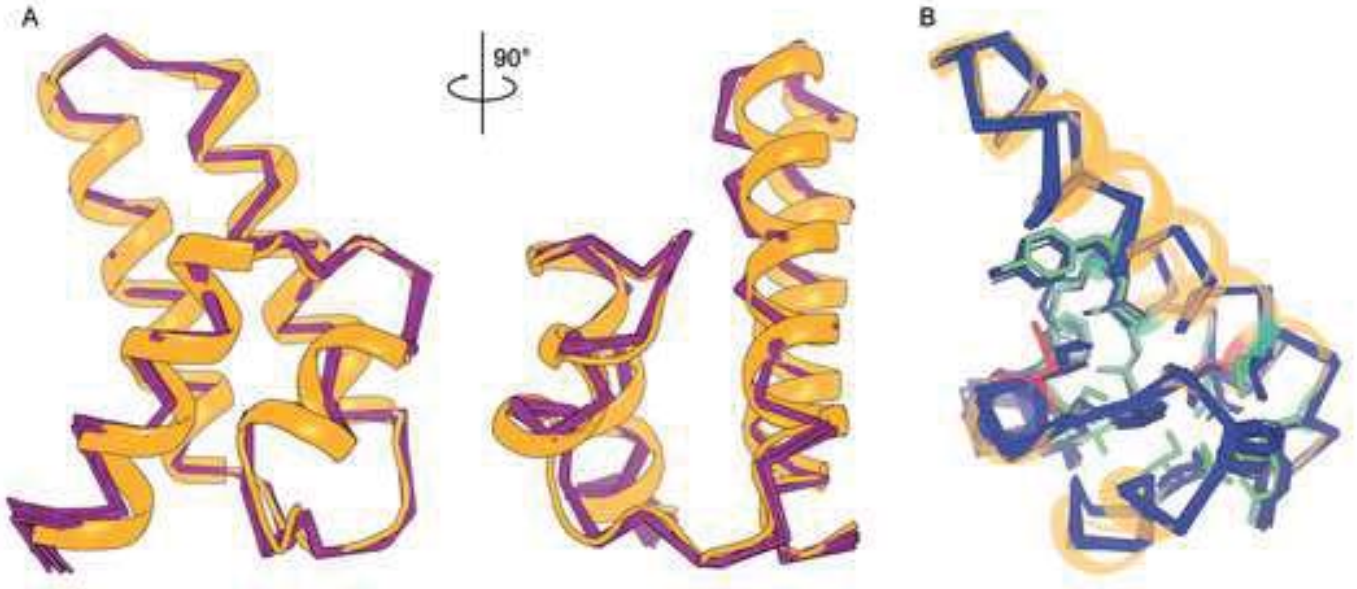


Figure5

Click here to download high resolution image



Chapter 3: An Interface-Driven Design Strategy Yields a Novel Corrugated Protein Architecture

Status: Published

ACS Synthetic Biology
Cite This: ACS Synth. Biol. XXXX, XXX, XXX−XXX
Research Article
pubs.acs.org/synthbio

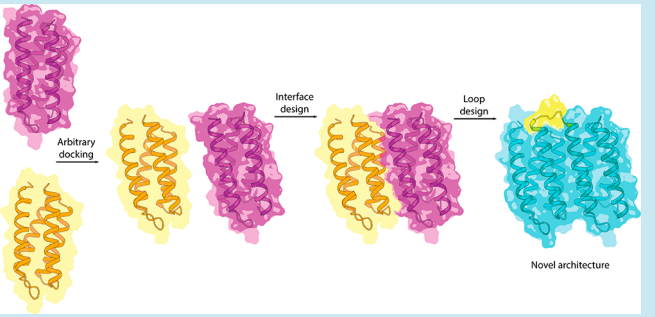
An Interface-Driven Design Strategy Yields a Novel, Corrugated Protein Architecture

Mohammad ElGamacy,[†] Murray Coles,[†] Patrick Ernst,[‡] Hongbo Zhu,[†] Marcus D. Hartmann,[†] Andreas Plückthun,[‡] and Andrei N. Lupas*[,†]

[†]Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany
[‡]Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

Supporting Information

ABSTRACT: Designing proteins with novel folds remains a major challenge, as the biophysical properties of the target fold are not known *a priori* and no sequence profile exists to describe its features. Therefore, most computational design efforts so far have been directed toward creating proteins that recapitulate existing folds. Here we present a strategy centered upon the design of novel intramolecular interfaces that enables the construction of a target fold from a set of starting fragments. This strategy effectively reduces the amount of computational sampling necessary to achieve an optimal sequence, without compromising the level of topological control. The solenoid architecture has been a target of extensive protein design efforts, as it provides a highly modular platform of low topological complexity. However, none of the previous efforts have attempted to depart from the natural form, which is characterized by a uniformly handed superhelical architecture. Here we aimed to design a more complex platform, abolishing the superhelicity by introducing internally alternating handedness, resulting in a novel, corrugated architecture. We employed our interface-driven strategy, designing three proteins and confirming the design by solving the structure of two examples.



KEYWORDS: *alternating handedness, novel fold, protein design, protein structure, repeat protein*

Computational design has thus far been very successful in diversifying the geometries and sequences of existing folds. This has been largely assisted by the presence of one or more starting structures for redesigning a particular fold, and the associated data that underpin its sequence determinants. In contrast, designing novel folds with a predetermined backbone blueprint, while offering a vast new range of designable folds, renders the gross sequence and rotamer sampling problem intractable. A fragment-based approach can greatly reduce this search space, as starting building blocks already carry intrinsic folding information. In addition to maintaining control over the level of adherence to a target fold, this approach also offers the possibility of coarse-graining the assembly problem by choice of the building block sizes. The latter may range from secondary structural elements to large subdomain or domain-sized fragments.[1] This effectively decomposes the problem into searching for optimal interfragment interfaces and loops. This promises to focus the available computing resources on accurately and exhaustively exploring restricted spaces, instead of sparsely exploring much larger ones. Here we demonstrate the capacity of this interface-driven approach as an efficient means for novel fold design.

For many years repeat proteins—in particular solenoids—have been a central topic of protein design. Unlike globular proteins, their low contact order and compositional uniformity have made them excellent platforms for investigating sequence-structure relationships and dissecting the energetics of protein folding.[2] They have also served a wide range of applications as antibody-like tailored synthetic binding proteins selected from libraries, and some have even progressed to late-stage clinical trials.[3] Because of their favorable biophysical properties, they have also been developed into crystallization chaperones.[4] Initially, design efforts on solenoids were aimed at generating more robust variants through sequence idealization.[5] More recently, the vast potential of solenoid proteins as tunable scaffolds has motivated computational design aimed at expanding the available repertoire of solenoid configurations with atomic accuracy.[6] These controlled geometries have included previously unobserved forms. However, despite this considerable success, to date the general solenoid architecture has not been altered. Here we aim to move beyond the solenoid, exploiting an incremental increase in the topological complexity to create a corrugated arrangement so far not observed in nature.

Solenoid proteins are characterized by a uniform connectivity between repeat units and thus wind into a continuous

Received: May 28, 2018
Published: August 27, 2018

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

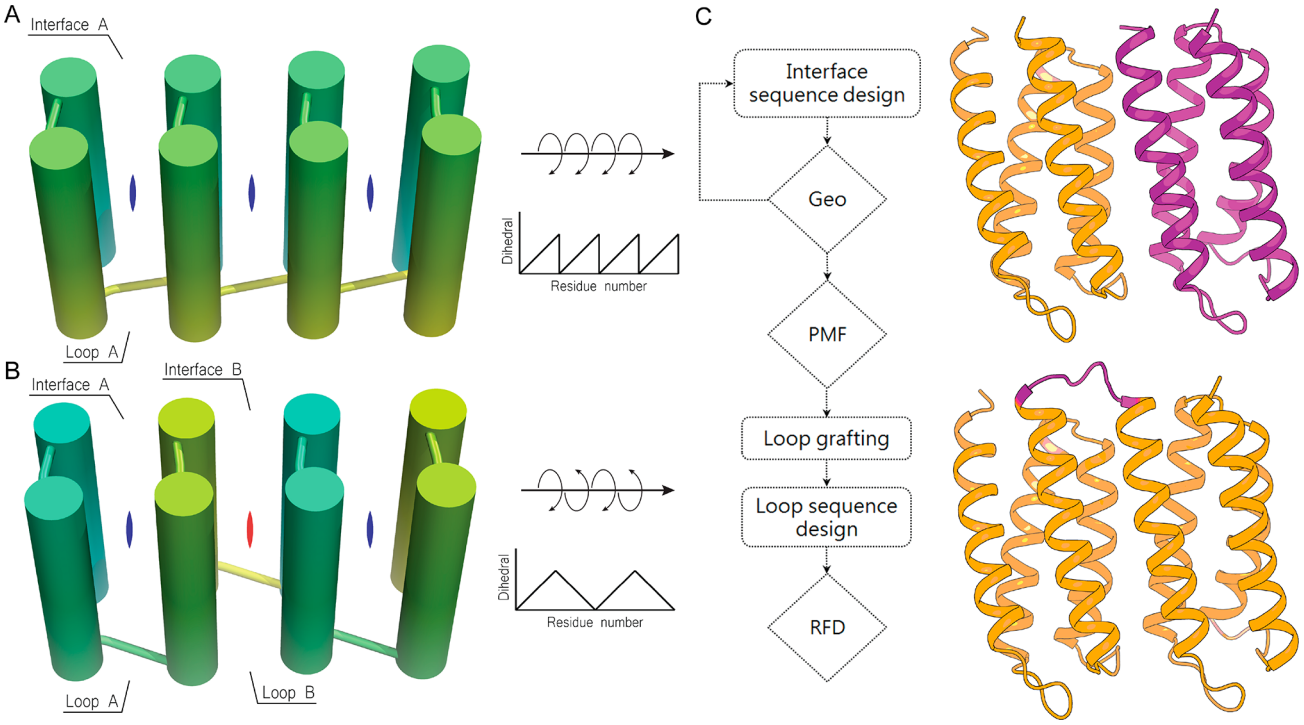


Figure 1. Solenoid repeat, corrugated repeat and design strategy. (A) Natural solenoids have repeats with a single junction type (blue circle) and a uniform handedness; they can thus be described by a sawtooth wave. The wave represents the torsion angle along the superhelical axis between the first and the *n*th residue. (B) A corrugated fold, represented by a triangle wave, would entail bihanded repeats, and thus require two junction types; Figure S9 shows actual values for idealized templates. (C) A two-stage strategy of interface design and loop construction; Geo: geometric filtering calculations, PMF: Potential of mean force energy calculations, RFD: Rotational force dissipation simulations. The target fold is built from a four-helix-bundle (orange) and its translational symmetry image (purple); top panel. The interface was then spanned by a grafted loop (purple); bottom panel.

superhelix.[7] This implies a single interunit junction type (defined as an interface and a connecting loop), where the units are bound to possess the same handedness. A waveform description of this periodic fold can be made by plotting the change in dihedral angle around the superhelical axis *versus* sequence position, which takes the form of a sawtooth wave (Figure 1A and Figure S9). The next step in complexity involves the alternating use of two junction types with opposite handedness; under the waveform description this topology adopts the form of a triangle wave (Figure 1B and Figure S9). Such a bihanded topology can be obtained by doubling the size of the building block and introducing a new junction of the opposite handedness to that of the starting block. In contrast to solenoids, the alternating handedness eliminates supercoiling. Here we have taken this approach to construct the corrugated target architecture, using an interface-driven design strategy that builds upon existing, simpler structural blocks and minimizes the amount of sampling required to achieve a target fold.

RESULTS AND DISCUSSION

Design Strategy. To construct the target topology, two unique helical hairpins, two unique interfaces, and two unique loops are required. The use of an up−down four-helix-bundle as a starting point provides two hairpins, a single interface and a single loop. The design of a second interface with the translated image of the bundle and of a second connecting loop is then sufficient to complete the target fold (Figure 1C). For this purpose, we employ a two-stage strategy: The first aimed at designing an intramolecular interface between two arbitrarily posed building blocks. This arbitrary docking step is only constrained by N- to C-terminal distance between the two blocks, a distance that can be defined by the allowed loop length. The second stage is aimed at contructing a loop across this interface.

We began by compiling a set of four-helix-bundles from the Protein Data Bank (PDB) that satisfied a set of geometric criteria defining regularity, bundle height range, and internal hairpin similarity. Initial poses were built between the bundle backbones and their translation images. The relative orientations were made to minimize the twist and curvature at the connecting interface along the central axis. This step was followed by the main sampling routine, where a combination of sequence sampling, side chain rotamer sampling, backbone refinement, and rigid-body docking were performed with an initially softened steric repulsion term. For efficient sampling, different iterations of the main Monte Carlo sampling loops were interlaced with a geometric filter. The latter being aimed at eliminating solutions with poor residue packing quality, before further rounds of design are resumed. The sequence sampling was restricted to the interface positions, while conformational refinement was performed globally (Materials and Methods, Figure S1). With the goal of further filtering the generated decoys by estimating the interaction free energy of the designed interfaces, more expensive potential-of-mean-force calculations were conducted through variable-velocity, variable-force steered molecular dynamics (SMD) simulations. The interaction free energy between the building blocks was calculated from the convolution of the velocity and force functions, and was used to rank the candidates accepted for the

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

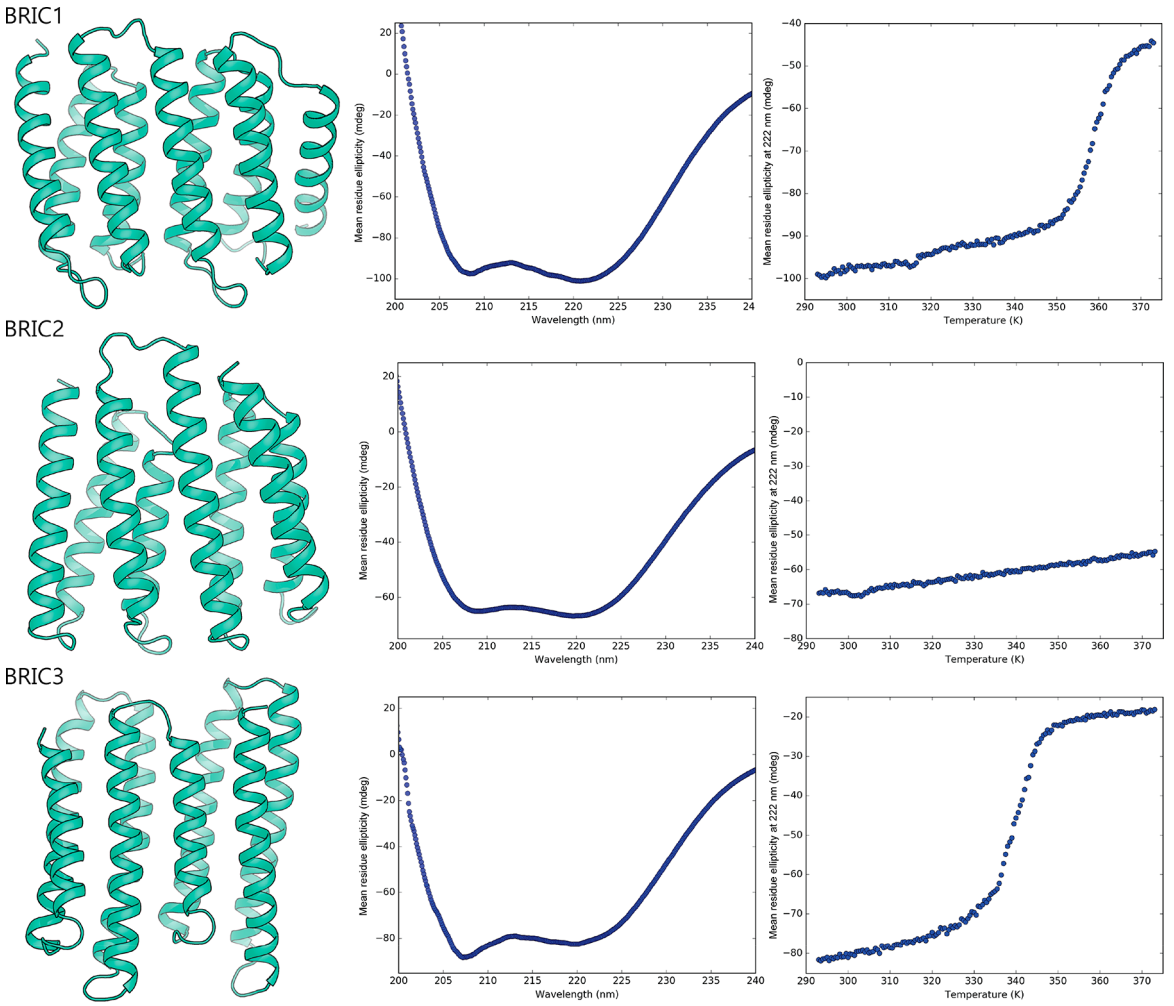


Figure 2. All three designs were folded. The first column shows the designed models as cartoon representation. The second column shows the respective CD spectra of the designs. The third column shows the melting curves of the designs, where BRIC1 and BRIC3 exhibit monophasic unfolding, while BRIC2 does not thermally unfold below 100 °C.

loop design stage. The described simulation setup applies a more adaptive pulling scheme of a previous constant-velocity setup that we have previously benchmarked against a subset of a protein−protein affinity data set[8] (Materials and Methods). This accelerated form of free energy estimation method has been shown to be particularly suited for protomers that do not undergo major conformational changes upon unbinding,[9] which was assumed here given the nature of our building blocks.

The next stage was to construct the loop that connects the newly designed interface. For this we searched the PDB for loop configurations that could serve as initial templates. The search routine scanned structures with a gapped sliding window, based on a generic description of the geometry defined by the ending and starting segments of adjacent repeat units (Figure S2). This description was obtained using the dihedrals profile, the axial vectors of the relevant segments and their orientations (Materials and Methods). The grafted loops were then subjected to combined sequence and conformer sampling, and all resulting loop compositions were evaluated using an accelerated molecular dynamics scheme. The routine applies a linear ramp of rotational force across the peptide bond at the center of the loop in a crankshaft fashion. This linear force titration, results in a nonlinear rotational response. the resulting nonlinear rotational kinetic energy is evaluated across the simulation time. The loop compositions that required the highest force magnitudes to induce rotational motion were selected for experimental evaluation (Materials and Methods, Figure S3).

Choice of Building Blocks. Three starting template bundles were adopted from three different natural proteins, to evaluate the generality of the approach and the choice of purely geometric criteria for template inclusion. The first design, BRIC1 (for Bihanded Repeat with Internal Corrugation), was constructed from a template bundle from the CheA histidine phosphotransfer domain (PDB: 1I5N);[10] the second, BRIC2, from the DRNN four-helix-bundle, which had previously undergone a total computational redesign of its hydrophobic core (PDB: 2LCH);[11] and the third, BRIC3, from a focal adhesion targeting domain (PDB: 3B71).[12] While one 2LCH is a monomeric solution structure, 1I5N and 3B71 do not possess any crystallographic arrangement similar to that proposed in Figure 1C. In the design process, BRIC1 underwent 12 mutations on the N-terminal face of the designed interface and 13 on the C-terminal face. BRIC2 underwent 12 mutations on the N-terminal face and 12 mutations on the C-terminal face; in addition to 2 mutations in the core of each bundle. BRIC3 underwent 20 mutations on

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

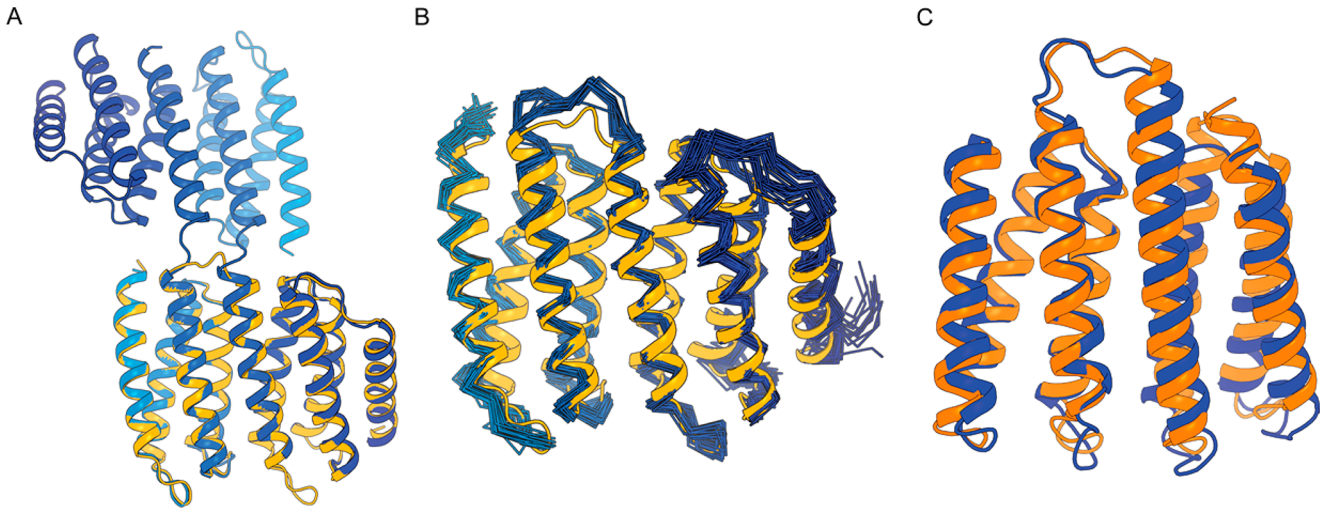


Figure 3. Experimental structures of BRIC1 and BRIC2 confirm the design. (A) The crystal structure of the 3D domain-swapped dimer, with individual protomers (colored by sequence position from cyan to blue) superimposed on the design (yellow). (B) The low-resolution NMR model of the BRIC1 monomer superimposed on the design (colors as in panel A). (C) The crystal structure of BRIC2 (blue) superimposed on the design (orange).

the N-terminal face and 17 mutations on the C-terminal face. Structure-based sequence alignments of the designs to their respective starting templates are shown in Table S1. The interface for BRIC1 was bridged by a 5-residue-loop, for BRIC2 by a 7-residue-loop, and for BRIC3 by a 6-residue-loop containing a disulfide bridge. For each of the designs, we explored experimentally the minimal form consisting of two repeat units. For BRIC1, we retained the native C-terminal helix of the phosphotransfer domain as a C-terminal capping helix.

Biophysical Characterization. We expressed the proteins in *Escherichia coli* and purified them using immobilized metal ion affinity chromatography (Materials and Methods). All three were primarily monomeric by analytical size-exclusion chromatography, although they all showed pH-dependent oligomerization. Their well-dispersed 1D NMR spectra were consistent with folded proteins (Figures S6 and S7) and their circular dichroism (CD) spectra with predominantly helical secondary structure (Figure 2). In thermal unfolding experiments, BRIC1 and BRIC3 showed single-phase equilibrium unfolding at 86 and 67 °C, respectively, while BRIC2 did not exhibit any melting transition. The monophasic melting transition of BRIC1 corresponds to that of a single, compact domain. This emphasizes the success of our interface design, as it implies that the enthalpy of the designed junction matches that of the native one. The three constructs underwent crystallization screening and only BRIC1 readily yielded diffracting crystals, BRIC2 was fused to a crystallization chaperone, while BRIC3 did not express in sufficient yield in M9 minimal medium or in fusion with the crystallization chaperone

Crystal Structure of BRIC1. Crystallization screens yielded well-diffracting BRIC1 crystals, for which we obtained data to 2.5 Å resolution in space group C2. The crystals contained one BRIC1 monomer in the asymmetric unit, which was unambiguously located in a molecular replacement trial searching with the full design model, in the first attempt and with high contrast. However, after initial refinement, it became apparent that the connectivity between the two four-helical halves of the protein differed from the design. Clear electron density showed the linker in an extended conformation, resulting from a domain-swapped dimeric assembly. In this assembly, two elongated BRIC1 protomers, related by the crystallographic 2-fold symmetry, are associated in an antiparallel fashion, such that the N-terminal four-helix bundle of one protomer interfaces with the C-terminal bundle of the other protomer, and vice versa (Figure 3 and Table S2). Given that BRIC1 also shows a minor dimeric form in solution (Figure S4), it appears that this form was selectively crystallized. As a result, the inter-repeat interface has entirely retained the designed interface features. This had a swapped backbone RMSD to the design of 1.82 Å (all-atom RMSD was 2.1 Å) across the entire structure, excluding the loop.

NMR Structure of BRIC1. To address the nature of the monomeric form of BRIC1, we prepared isotope labeled samples for solution NMR. Diffusion coefficients measured on freshly prepared samples were consistent with the designed monomer (Figure S5). However, dimeric and higher oligomeric forms accumulated over time, impacting on the quality of spectra. This feature, combined with the ambiguity intrinsic to repeat sequences, precluded full resonance assignment and thus high-resolution structure determination. We therefore adopted a strategy aimed at creating a low-resolution model, using a sample selectively ^{13}C-labeled on methionine methyl groups to define interhelical contacts (Materials and Methods). An initial observation was the similarity of chemical shifts between the repeats, indicating that both adopt very similar structures. Interhelical contacts then defined intra- and inter-repeat junctions very similar to those observed in the crystal, with the C-terminal repeat identified by contacts to the unambiguously assigned C-terminal capping helix. The compiled data were sufficient to define the monomer structure, using the domain-swapped crystallographic protomer as a starting point (Figure 3B and Table S3). The calculated monomer ensemble agrees well with the design, with an average backbone RMSD of 1.8 Å (all-atom RMSD ranged from 2.5 to 2.9 Å, excluding the capping helix).

Crystal Structure of BRIC2. In contrast to BRIC1, BRIC2 did not yield well-diffracting crystals in the first attempt. For this reason, a rigid shared helix fusion to DARPin D12

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

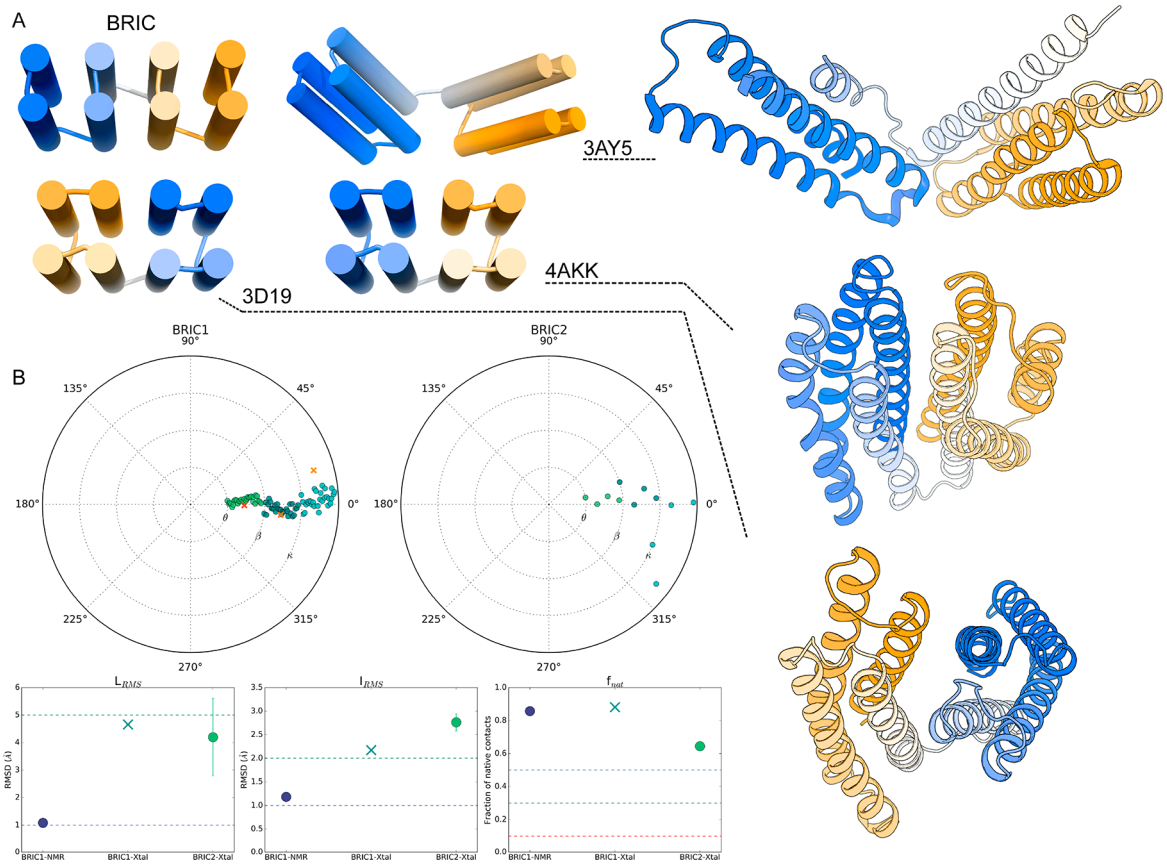


Figure 4. Architectural uniqueness and interface accuracy of the BRIC designs. (A) A comparison between the idealized BRIC architecture and the closest architectures through structural similarity searches. The structures are colored by chain path from blue to yellow. (B) Polar and Cartesian disparity between the designs and experimental structures of BRIC1 and BRIC2. The top panel shows the angular deviation from design values for the tilt (θ), bend (β) and curvature (κ) across the designed interface in green, teal and cyan, respectively. Each dot represents either an NMR model or one of the asymmetric unit chains (the single chain of the BRIC1 crystal structure is represented by orange crosses). The bottom panel shows the CAPRI evaluation criteria (L_{rms}, I_{rms} and f_{nat}) for the designed interfaces (defined in the Materials and Methods). The red, green and blue dashed horizontal lines mark the high, medium and acceptable ranks, respectively. Error bars represent the standard error across asymmetric unit chains of NMR models (some error bars are within the dot diameter).

(designed ankyrin repeat protein D12) was constructed. DARPin D12 had been previously identified as well-crystallizing under many different conditions and thus serving as a crystallization chaperone when rigidly fused to other repeat proteins.[4a] An N-terminal fusion of the DARPin was built *in silico* and both the shared helix and residues within 5 Å proximity were sequence-optimized using *Rosetta fixed backbone design*, as previously described.[4a] Crystals appeared after 25 days, diffracted to 3.0 Å resolution and the data were integrated in space group P1. For the molecular replacement, a model of the DARPin was used as a search model and the design models of BRIC2 were manually fitted into the density (refinement statistics are provided in Table S4). Out of the four molecules of the asymmetric unit, chain B and D looked as designed and a slight bend of the shared helix was observed for chains A and C, due to crystal forces (Figure S8). Clear electron density was visible for the whole BRIC2 domain and the designed loop, connecting the two repeats, could be built. In comparison to BRIC1, BRIC2 was monomeric and no domain swap was observed in all of the four chains, proving the successful interface design between the two helical bundles (Figure 3C). The overall backbone RMSD ranged from 2.27 to 3.0 Å (all-atom RMSD ranged from 2.8 to 3.4 Å), and confirmed both the design and the potential of DARPin D12 as a crystallization chaperone.

Architectural Uniqueness and Interface Design Precision. To contrast the BRIC architecture to the nearest existing folds, we conducted structure searches against the entire PDB using PDBeFOLD[13] and DALI,[14] and the ECOD database[15] using TM-align.[16] No folds were found that structurally align along the full length of our designed structures. Any similarity detected was largely localized to a four-helix-bundle substructure. PDBeFOLD did not recover any significantly related hits, while the best TM-align hits had TM-scores <0.55 and did not share significant similarity with our BRICs. For the DALI searches we selected three structures based on their alignment lengths and secondary structures arrangements. Figure 4A shows the structures and idealized topologies of these hits contrasted against the BRIC topology. Two of these hits (3D19 and 4AKK) were topologically similar to each other, but with opposite chain paths. These were composed of two uniformly handed four-helix-bundles with N- and C-termini abutting each other at the connecting interface; a close-ended configuration that results from the parallel orientation of the helical hairpins to the main axis. The third hit (3AY5) consisted of two dissociated antiparallel helical domains, with one being a right-handed, side-connected

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

bundle, and the other a left-handed, diagonally connected bundle.

To evaluate the interface design precision, we measured the polar and Cartesian error across the interface (Figure 4B). The polar deviations of the interface between designs and experimental structures were defined in terms of tilt (θ), bend (β) and curvature (κ). Polar deviations in BRIC1 were lower in the solution structure than in the crystal structure, with the highest deviation in the κ dimension. BRIC2, however, exhibited a larger range of deviations between the asymmetric unit protomers, particularly, along the β and κ dimensions. As these large interprotomer deviations would potentially build-up in a multirepeat scenario, we carried out molecular dynamics simulations for a five-bundle-repeat model of BRIC2. The deviations in the solvated simulations were of smaller magnitude, averaging below 8.0°, 6.6° and 3.0° for $|\theta|$, $|\beta|$ and $|\kappa|$, respectively, at the four designed interfaces. We therefore expect most of these deviations to stem from the crystal packing. For estimating the Cartesian precision at the interface, we calculated the evaluation criteria used in the CAPRI interaction prediction competition: L_{rms}, I_{rms} and f_{nat}.[17] The three structures ranked *medium* on the L_{rms} score. The BRIC1 solution structure ranked *medium* on the I_{rms} score, while the two crystal structures ranked *acceptable*. All three structures ranked *high* on the f_{nat} score. In spite of the asymmetric nature of the two-sided interface design, the intramolecular four-helix-bundle backbone RMSD was minor; 0.8 Å within design and 1.1 Å within structure for BRIC1, and 0.9 Å within design and 1.3 Å within structure for BRIC2. The design *vs* structure values were 0.6 Å for both respective bundles of BRIC1, and 1.2 and 1.3 Å for the first and second bundles of BRIC2, which affirms the rigid incorporation of the building blocks.

CONCLUSIONS

At the frontier of protein design is the aim to provide new scaffolds for functionalization, this potential has made repeat architectures attractive design targets. Internal cross-alignments in our experimental structures show that minimal structural perturbation has been introduced to the starting building blocks, leading to the possibility of constructing longer repeats. With this architecture, the large-sized building blocks can harbor functional features from selected parent blocks, or afford more extensive engineering owing to their expanded substructural diversity, as compared to repeats with smaller building blocks.

Protein design efforts have so far been biased toward assembling idealized secondary structure elements. As such they do not reflect natural proteins, where structural deformities are common and often associated with functional motifs. The difficulty of sampling such deformities is an inherent barrier to the designability of these motifs.[18] A successful design strategy that sidesteps this problem, and even creates new topologies, has been to combine natural substructures by directly using structurally similar fragments as overlapping connectors.[19] This, however, does not allow strict control of target topologies, since it is contingent upon the existence of overlaps that yield viable intramolecular interfaces. In contrast to this connectivity-driven approach, we introduce an interface-driven approach that is capable of delivering novel topologies from an arbitrary arrangement of building blocks. Moreover, this strategy employs sequence and conformational sampling focused only on the junctions between building blocks, and separates interface optimization from loop design, thus adding to the overall efficiency.

MATERIALS AND METHODS

Computational Design. The interface sequence design stage was performed in multiple consecutive rounds filtering the top 10−20 candidates from each round and feeding them as input to the next. Each round was performed using a RosettaScripts[20] protocol comprising two generic Monte Carlo loops separated by packstat[21] and total energy (talaris2013 scoring function[22]) filters. Each loop executed a protocol comprising soft-repulsion sequence sampling, backbone optimization,[23] docking and conformational refinement. Between the consecutive rounds, under- or overpacking was evaluated by calculating the average deviation from high-resolution structures packing density probability. The last round's output was filtered through an accelerated SMD routine that aims at approximately assessing the potential of mean force of unbinding across the designed interface. The free energy of unbinding (W) was evaluated as $W_{t_o \to t_e} = \int_{t_o}^{t_e} \mathbf{v}(t)\,\mathbf{F}(t)\mathrm{d}t$ where $\mathbf{F}(t)$ and $\mathbf{v}(t)$ are the pulling force and velocity vectors at time t, respectively. One partner was fixed and aligned against a reference orientation while the other was pulled along a single dimension through a loose spring to achieve a variable-velocity, variable-force SMD setup that yields the free energy profile along the unbinding path. The protein was modeled using the CHARMM36 force field,[24] where the simulations were performed in explicit solvent (TIP3P water model) and 0.15 M sodium chloride as NPT ensembles at 310 K and 1 atm using a Langevin thermostat and a Langevin barostat as implemented in the NAMD engine.[25] Particle Mesh Ewald electrostatics grid of 1 Å resolution was used with a long-range cutoff set at 12 Å (switching at 10 Å) and a time step of 2 fs. The reference pulling velocity ($\mathbf{v}_{ref}$) was calibrated to 2.5 Å/ns with a spring constant (k) of 20 kcal·mol^{-1}·Å^{-2} where the applied force ($\mathbf{F}(t)$) was computed as $-\nabla\left[\frac{1}{2}k[t\mathbf{v}_{ref} - (\mathbf{r}_t - \mathbf{r}_o)\cdot\mathbf{n}]^2\right]$ ($\mathbf{r}_t$ being the position vector of the steered atom group and $\mathbf{n}$ being the pulling direction vector). The systems underwent 2000 steps of conjugate gradient minimization before random atom velocities initialization and force application on the backbone carbonyl carbon atoms. The calculated work was used to rank designs for the next stage.

The loop design stage begins with a structural search using a gapped sliding window across the whole PDB, where the landing sites are defined by two N-to-C vectors and a single (φ, ψ) array. Given the latter representation, every subject landing site was compared to the query geometry by means of dihedral profiles similarity, landing sites lengths similarity and landing sites relative orientation similarity. Loop lengths of 4 and up to 8 were searched for, with landing sites of lengths ranging from 4 to 8 residues. The best matches according to the previous metrics were then grafted onto the top ranking interface designs and subjected to loop mutagenesis using a Rosetta script that performs sequence sampling, backrub refinement, and side chain refinement in a Monte Carlo looper. The designed loops were evaluated by applying reciprocating crankshaft force across the peptide bond at the center of the loop with a reciprocation frequency of 20 fs^{-1}. A 60 ps span of equilibration was followed by equal torques applied to the peptide bond hydrogen and oxygen atoms around the peptide bond axis, starting by an angular

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

acceleration of 2 rad·ps^{-2}. The latter rotational acceleration was incrementally ramped up every 40 fs by a value of 2 rad·ps^{-2} using the updated atomic positions every 20 fs so as not to apply any forces against the peptide axis itself. The simulation was performed in triplicates in durations of 300 ps with similar system setup parameters to the SMD described above. The distributions of the loop atoms root-mean-squared-fluctuation and rotational kinetic energy were assessed to choose the designs of the lowest mean and standard deviation of these variables. The top designs at this point were directly taken to the laboratory.

Expression and Purification. The genes were acquired from Synbio Technologies, already cloned into pET-28a(+) using NcoI and NdeI cloning sites and in-frame with an N-terminal hexaHis-tag and a thrombin cleavage site, while harboring a kanamycin resistance gene as a selection marker. The plasmids were used to transform chemically competent *E. coli* BL21(DE3) by means of heat-shock. The expression procedure entailed growing of the cells in LB medium and inducing with IPTG at OD_{600} of 0.5−1 with overnight expression at 25 °C. For expression of labeled protein, a preculture in LB medium was grown, cells collected, and resuspended in M9 minimal medium (240 mM Na_2HPO_4, 110 mM KH_2PO_4, 43 mM NaCl), supplemented with 10 μM $FeSO_4$, 0.4 μM H_3BO_3, 10 nM $CuSO_4$, 10 nM $ZnSO_4$, 80 nM $MnCl_2$, 30 nM $CoCl_2$ and 38 μM kanamycin sulfate, to an OD_{600} of 0.5−1. After 40 min of incubation at 25 °C, 2.0 g ^{15}N-labeled ammonium chloride (Sigma-Aldrich cat. no. 299251) and 6.25 g ^{13}C D-glucose (Cambridge Isotope Laboratories, Inc. cat. no. CLM-1396) were added, or 100 mg methyl-^{13}C L-methionine (Sigma-Aldrich cat. no. 299146) in case of selective labeling in a 2.5 L culture. After another 40 min IPTG was added to 1 mM final concentration for overnight expression. Cells were collected by centrifugation at 5000*g* for 15 min, lysed by a Branson Sonifier S-250 (Fisher Scientific) in hypotonic 50 mM Tris-HCl buffer supplemented with one tablet of the complete protease cocktail (Sigma-Aldrich cat. no. 4693159001) and 3 mg of lyophilized DNase I (5200 U/mg; Applichem cat. no. A3778). The insoluble fraction was pelleted by 25 000*g* centrifugation for 50 min, and the soluble fraction was filtered (0.45 μm filter pore size) and directly applied to a Ni-NTA column. A 5 mL HisTrapFF immobilized nickel column (GE Healthcare Life Sciences cat. no. 17−5255−01) was used for this purpose, washed consecutively by 30 mL 150 mM NaCl, 30 mM Tris buffer (pH 8.5) at 0, 30, and 60 mM imidazole. Fractions were collected by a gradient elution at >60 mM imidazole. The eluate was concentrated using 10 kDa MWCO centrifugal filters (Merck Millipore cat. no. UFC901024) and loaded onto an equilibrated Superdex 75 gel filtration column (GE Healthcare Life Sciences cat. no. 17517401). The gel filtration buffer used was always 100 mM sodium phosphate buffer (for NMR and CD transparency) composed to the target pH, where BRIC1 was eluted in pH 8.5, while BRIC2 and BRIC3 at pH 5.5. An ÄktaFPLC system (GE Healthcare Life Sciences) was used for all chromatography runs.

For the D12-BRIC2 chimera, the shared helix between D12 and BRIC2 was introduced by assembly PCR and the resulting fragment was cloned into a pQE30LIC_3C (Qiagen) based plasmid *via* BamHI and HindIII restriction sites. Chemo-competent BL21DE3 cells were transformed with the plasmid and the protein was expressed in autoinduction medium at 25 °C for 16 h.[26] Cells were resuspended in 50 mM Tris/HCl pH 8, 500 mM NaCl, 20 mM imidazole and lysed *via* sonication. Insoluble material was spun down by centrifugation for 30 min at 30 000*g* and the supernatant was loaded on 5 mL NiNTA resin, equilibrated with resuspension buffer. The column was washed with 25 mL resuspension buffer and protein was eluted with 15 mL resuspension buffer containing 250 mM imidazole. The elution fraction was dialyzed overnight against 50 mM Tris/HCl pH8, 300 mM NaCl and the N-terminal 10xHis-tag was removed by cleavage with 3C-protease (2% w/w). Following a second NiNTA step to remove the protease and the His-tag, the protein solution was concentrated to 5 mL and further purified by gel filtration on an S200 16/600 column (GE healthcare) equilibrated with 10 mM Tris/HCl pH 8, 100 mM NaCl.

Biophysical Characterization. The analytical gel filtration experiments were all done on a Superdex 200 10/300 GL (GE Healthcare Life Sciences cat. no. 17517501), and the collected fractions from the eluate were used for CD or NMR measurements directly after. ^{1}H NMR spectra were collected on a Bruker AVIII-800. NMR diffusion ordered spectroscopy experiments were performed on a Bruker AVIII-600 using the relevant functionality in the TopSpin software, running the analysis over multiple aliphatic proton peaks. The structure-based prediction of the diffusion coefficient was done using the HYDROpro software,[27] setting the corresponding temperature to 310 K and viscosity to 0.007 P. CD spectra were recorded on a Jasco J-810 spectrometer, with a spectral scan window of 200−240 nm, with a sweep delta of 0.1 nm while averaging over 5 scans. Melting curves were measured from 20 to 100 °C, recording the ellipticity at 222 nm every 0.5 °C, while heating at a 1 °C/min rate.

X-ray Crystallography. For BRIC1 crystallization, the protein was concentrated to 13 mg/mL in 25 mM Tris buffer, pH 8.5, 150 mM NaCl. The D12-BRIC2 fusion was concentrated to 40 mg/mL in 10 mM Tris buffer, pH 8, 100 mM NaCl. Sitting-drop vapor diffusion crystallization trials were performed in 96-well format, equilibrating drops containing 300 nL of protein solution and 300 nL of reservoir solution against 50 μL of reservoir solution. For D12-BRIC2, the drop size was 150 nL + 150 nL and the reservoir contained 75 μL of mother liquor. Best diffracting crystals were obtained with a reservoir solution containing 20% v/v PEG 500 MME, 10% w/v PEG 20 000, 30 mM $MgCl_2$, 30 mM $CaCl_2$ and 100 mM Tris-BICINE pH 8.5, loop-mounted, and flash-frozen in liquid nitrogen. For D12-BRIC2, an initial hit was found in 0.2 M $(NH_4)_2SO_4$, 25% w/v PEG 3350 and 100 mM Bis-Tris pH 5.5. A fine screen with two perpendicular gradients of the PEG concentration and the pH was set up to yield diffracting crystals, which were flash-frozen in mother liquor containing 20% v/v ethylene glycol. Data were collected at beamline X10SA at the Swiss Light Source, at 100 K with an X-ray wavelength of 1 Å and a PILATUS 6M-F detector (Dectris) for BRIC1 or an EIGER 16 M X detector (Dectris) for D12-BRIC2. Data for BRIC1 were indexed, integrated and scaled to a resolution of 2.5 Å in space group C2, using XDS.[28] For D12-BRIC2, two crystals were indexed and integrated in space group P1. After merging the two data sets, the data were scaled to 3 Å. According to the unit cell dimensions, one BRIC1 monomer was expected in the asymmetric unit with a solvent content of 50%. Molecular replacement was carried out using MOLREP,[29] using the designed coordinates as a search model. A unique solution was found in the first attempt with high contrast. After rigid-body refinement with Refmac5,[30] a

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

different conformation of the designed connecting loop became apparent and was manually rebuilt in Coot.[31] The structure was completed and finalized in cycles of manual modeling in Coot and refinement with BUSTER or PHENIX-refine.[32] Data processing and refinement statistics are summarized in Tables S2 and S4.

NMR Structure Determination. Spectra were recorded at 310 K on Bruker AVIII-600 and AVIII-800 spectrometers. Backbone sequential assignments were made using standard triple-resonance experiments and by tracing strong NOESY contacts between sequential amide protons in helical segments. Aliphatic side chain assignments were completed with TOCSY-based experiments, while partial aromatic assignments were made by linking aromatic spin systems to unambiguously assigned aliphatic groups in NOESY spectra. The oligomeric purity of samples was checked with diffusion-ordered (DOSY) spectra. These confirmed that fresh samples used in diffusion experiments were predominantly monomeric.

To identify interhelical contacts, we exploited the uneven distribution of methionine residues observed in the dimeric crystal structure. The 16 methionine residues in this structure fall into three broad clusters, one within each repeat and a third at the inter-repeat interface. To assign these we produced a sample selectively ^{13}C-labeled on methionine methyl groups on a ^{12}C, ^{15}N-labeled background. Members of each cluster could be identified by contacts between the labeled methyl groups in a 3D CCH-NOESY experiment.[33] Contacts to unambiguously assigned protons in a ^{13}C-HSQC-NOESY spectrum then allowed the assignment of all members within each cluster. Thus, assigned, these methyl groups were effective probes of the interhelical interfaces providing 34 long-range distance restraints. These were applied, in simulated annealing calculations, together with other unambiguously assigned contacts and TALOS-based dihedral restraints. A summary of the input data and final structure statistics is given in Supplementary Table S3.

Structures were calculated with XPLOR-NIH (version 2.9.4) using a monomer extracted from the domain-swapped dimer as a starting structure; *i.e.*, an open structure with no interunit interface. Simulated annealing runs were first aimed at closing this interface by treating the four-helix bundles as pseudorigid bodies. The resulting set of 50 structures defined an interface very similar to that observed in the crystal structure. Refinement was performed using atomistic molecular dynamics computations in isothermal−isobaric ensembles to accommodate large conformational changes, where the overall explicit solvent simulations setup was similar to that described above. A total of 135 ns were collected while deploying the NMR-derived dihedral and distance restraints using the harmonic restraint terms $k_{\text{torsion}}(\theta_t - \theta_{\text{ref}})^2$ and $k_{\text{distance}}(x_t - x_{\text{ref}})^2$, respectively. Here k_{torsion} and k_{distance} are the dihedral and distance spring constants (set at 1 and 0.1, respectively), θ_t is the φ or ψ angle at time t, x_t is the atom pair distance at time t, while θ_{ref} and x_{ref} are the NMR-derived values. Fifty frames from these runs were picked on the basis of agreement with distance restraints and minimized under restraints in XPLOR-NIH to regularize covalent geometry. The final ensemble consisted of 26 structures chosen on the basis of lowest restraint violations.

Structural Analysis. Searching among existing structures for similar folds was performed using three different methods. The PDBeFOLD[13] and DALI[14] servers were used to search against the entire PDB for similar existing folds to the experimental structures of BRIC1 and BRIC2. The resulting hits were sorted by their alignment lengths, and manually inspected the top 100 hits for similar topologies. Additionally, the ECOD database[15] (ECOD40 subset) was searched using TM-align[16] for the same purpose. Only hits with TM-score equal or above 0.5 were manually inspected for potential similarity.

Polar precision at the designed interface was assessed by calculating the deviation between the designs and experimental structures for three quantities; the tilt (θ), bend (β) and curvature (κ) across the designed interface. The three quantities are supposed to represent the plane-projected angular change between the two helical hairpins across the designed interface, along the three mutually orthogonal planes. The assessment of the designed interface accuracy in Cartesian and qualitative terms was done using the CAPRI interface criteria: L_{rms}, I_{rms} and f_{nat}.[17] The L_{rms} represents the backbone RMSD of the protein unit downstream of the designed interface, after structurally aligning the pair by their upstream units. The I_{rms} was calculated as the backbone RMSD between the residues at the designed interface (defined by a distance cutoff of 10 Å). The f_{nat} represents the number of contacts common across the designed interface between the design and experimental structure, divided by the total number of contacts in the experimental structure. A contact is defined by the existence of any interatomic distance within 5 Å between two residues across either side of the interface (the designed loop residues were not considered).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.8b00224.

Figures S1−S9; Tables S1−S3 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: andrei.lupas@tuebingen.mpg.de.

ORCID

Murray Coles: 0000-0001-6716-6150
Marcus D. Hartmann: 0000-0001-6937-5677
Andreas Plückthun: 0000-0003-4191-5306
Andrei N. Lupas: 0000-0002-1959-4836

Notes

The authors declare no competing financial interest.
The crystal structure of the dimeric BRIC1 has been deposited in the RCSB Protein Data Bank under the code 6FF6. The crystal structure of the D12-BRIC2 fusion has been deposited in the RCSB Protein Data Bank under the code 6FES.

■ ACKNOWLEDGMENTS

We thank Reinhard Albrecht and Vincent Truffault from the Max Planck Institute for Developmental Biology for technical assistance in structure determination experiments. We also thank Beat Blattmann from the UZH Protein Crystallization Center for setting up the crystallization experiments for the D12-BRIC2 chimera. This work was supported by institutional funds of the Max Planck Society and by grant 310030B_166676 from the Swiss National Science Foundation to AP. The funders had no role in study design, data collection

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

(1) Woolfson, D. N., Bartlett, G. J., Burton, A. J., Heal, J. W., Niitsu, A., Thomson, A. R., and Wood, C. W. (2015) De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol. 33*, 16−26.

(2) (a) Kloss, E., Courtemanche, N., and Barrick, D. (2008) Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys. 469* (1), 83−99. (b) Rowling, P. J. E., Sivertsson, E. M., Perez-Riba, A., Main, E. R. G., and Itzhaki, L. S. (2015) Dissecting and reprogramming the folding and assembly of tandem-repeat proteins. *Biochem. Soc. Trans. 43* (5), 881−888. (c) Kajander, T., Cortajarena, A. L., Main, E. R. G., Mochrie, S. G. J., and Regan, L. (2005) A New Folding Paradigm for Repeat Proteins. *J. Am. Chem. Soc. 127* (29), 10188−10190. (d) Mello, C. C., and Barrick, D. (2004) An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. U. S. A. 101* (39), 14102−14107. (e) Lowe, A. R., and Itzhaki, L. S. (2007) Rational redesign of the folding pathway of a modular protein. *Proc. Natl. Acad. Sci. U. S. A. 104* (8), 2679−2684. (f) Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K., and Plückthun, A. (2008) Folding and Unfolding Mechanism of Highly Stable Full-Consensus Ankyrin Repeat Proteins. *J. Mol. Biol. 376* (1), 241−257. (g) Wetzel, S. K., Ewald, C., Settanni, G., Jurt, S., Plückthun, A., and Zerbe, O. (2010) Residue-Resolved Stability of Full-Consensus Ankyrin Repeat Proteins Probed by NMR. *J. Mol. Biol. 402* (1), 241−258.

(3) Plückthun, A. (2015) Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy. *Annu. Rev. Pharmacol. Toxicol. 55* (1), 489−511.

(4) (a) Wu, Y., Batyuk, A., Honegger, A., Brandl, F., Mittl, P. R. E., and Plückthun, A. (2017) Rigidly connected multispecific artificial binders with adjustable geometries. *Sci. Rep. 7* (1), 11217. (b) Batyuk, A., Wu, Y., Honegger, A., Heberling, M. M., and Plückthun, A. (2016) DARPin-Based Crystallization Chaperones Exploit Molecular Geometry as a Screening Dimension in Protein Crystallography. *J. Mol. Biol. 428* (8), 1574−1588.

(5) (a) Main, E. R. G., Jackson, S. E., and Regan, L. (2003) The folding and design of repeat proteins: reaching a consensus. *Curr. Opin. Struct. Biol. 13* (4), 482−489. (b) Parmeggiani, F., Huang, P.-S., Vorobiev, S., Xiao, R., Park, K., Caprari, S., Su, M., Seetharaman, J., Mao, L., Janjua, H., Montelione, G. T., Hunt, J., and Baker, D. (2015) A General Computational Approach for Repeat Protein Design. *J. Mol. Biol. 427* (2), 563−575. (c) Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P., and Plückthun, A. (2003) Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. *J. Mol. Biol. 332* (2), 489−503. (d) Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A., and Plückthun, A. (2008) Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. *J. Mol. Biol. 376* (5), 1282−1304.

(6) (a) Brunette, T. J., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. (2015) Exploring the repeat protein universe through computational protein design. *Nature 528*, 580. (b) Parmeggiani, F., and Huang, P.-S. (2017) Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol. 45*, 116−123. (c) Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L., and Bradley, P. (2015) Rational design of α-helical tandem repeat proteins with closed architectures. *Nature 528*, 585. (d) Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. (2015) Control of repeat protein curvature by computational protein design. *Nat. Struct. Mol. Biol. 22* (2), 167−174.

(7) Kobe, B., and Kajava, A. V. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci. 25* (10), 509−515.

(8) Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. J. J., and Janin, J. (2011) A structure-based benchmark for protein−protein binding affinity. *Protein Sci. 20* (3), 482−491.

(9) Chen, P.-C., and Kuyucak, S. (2011) Accurate Determination of the Binding Free Energy for KcsA-Charybdotoxin Complex from the Potential of Mean Force Calculations with Restraints. *Biophys. J. 100* (10), 2466−2474.

(10) Mourey, L., Da Re, S., Pédelacq, J.-D., Tolstykh, T., Faurie, C., Guillet, V., Stock, J. B., and Samama, J.-P. (2001) Crystal Structure of the CheA Histidine Phosphotransfer Domain that Mediates Response Regulator Phosphorylation in Bacterial Chemotaxis. *J. Biol. Chem. 276* (33), 31074−31082.

(11) Murphy, G. S., Mills, J. L., Miley, M. J., Machius, M., Szyperski, T., and Kuhlman, B. (2012) Increasing Sequence Diversity with Flexible Backbone Protein Design: The Complete Redesign of a Protein Hydrophobic Core. *Structure (Oxford, U. K.) 20* (6), 1086−1096.

(12) Garron, M.-L., Arthos, J., Guichou, J.-F., McNally, J., Cicala, C., and Arold, S. T. (2008) Structural Basis for the Interaction between Focal Adhesion Kinase and CD4. *J. Mol. Biol. 375* (5), 1320−1328.

(13) Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 60*, 2256−2268.

(14) Holm, L., and Laakso, L. M. (2016) Dali server update. *Nucleic Acids Res. 44*, W351−W355.

(15) Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H., and Grishin, N. V. (2014) ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol. 10* (12), e1003926.

(16) Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res. 33* (7), 2302−2309.

(17) Mendez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003) Assessment of blind predictions of protein−protein interactions: Current status of docking methods. *Proteins: Struct., Funct., Genet. 52* (1), 51−67.

(18) (a) Kim, D. E., Blum, B., Bradley, P., and Baker, D. (2009) Sampling bottlenecks in de novo protein structure prediction. *J. Mol. Biol. 393* (1), 249−260. (b) Huang, P.-S., Boyken, S. E., and Baker, D. (2016) The coming of age of de novo protein design. *Nature 537*, 320.

(19) Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., and Kuhlman, B. (2016) Design of structurally distinct proteins using strategies inspired by evolution. *Science 352* (6286), 687.

(20) Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS One 6* (6), e20161.

(21) Sheffler, W., and Baker, D. (2009) RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci. 18* (1), 229−239.

(22) Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D., and Kuhlman, B. (2013) Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods Enzymol. 523*, 109−143.

(23) Smith, C. A., and Kortemme, T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol. 380* (4), 742−756.

(24) Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell, A. D. (2012) Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput. 8* (9), 3257−3273.

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(25) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable Molecular Dynamics with NAMD. *J. Comput. Chem. 26* (16), 1781−1802.
(26) Studier, F. W. (2005) Protein production by auto-induction in high-density shaking cultures. *Protein Expression Purif. 41* (1), 207−234.
(27) Ortega, A., Amorós, D., and García de la Torre, J. (2011) Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models. *Biophys. J. 101* (4), 892−898.
(28) Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr. 26* (6), 795−800.
(29) Vagin, A., and Teplyakov, A. (2000) An approach to multi-copy search in molecular replacement. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 56* (12), 1622−1624.
(30) Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S., and Dodson, E. J. (1999) Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 55* (1), 247−255.
(31) Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 60*, 2126−2132.
(32) (a) Bricogne, G. B. E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P, S, A., Smart, O. S., Vonrhein, C., and Womack, T. O. (2017) *BUSTER*, version 2.10.3, Global Phasing Ltd., Cambridge, UK. (b) Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 66* (2), 213−221.
(33) Diercks, T., Coles, M., and Kessler, H. (1999) An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. *J. Biomol. NMR 15* (2), 177−180.

DOI: 10.1021/acssynbio.8b00224
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Supporting Information

An interface-driven design strategy yields a novel corrugated protein architecture

Mohammad ElGamacy, Murray Coles, Patrick Ernst[†], Hongbo Zhu, Marcus D. Hartmann, Andreas Plückthun[†] and Andrei N. Lupas[*]

Dept. of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

[†]Dept. of Biochemistry, University of Zurich, 8057 Zurich, Switzerland

.

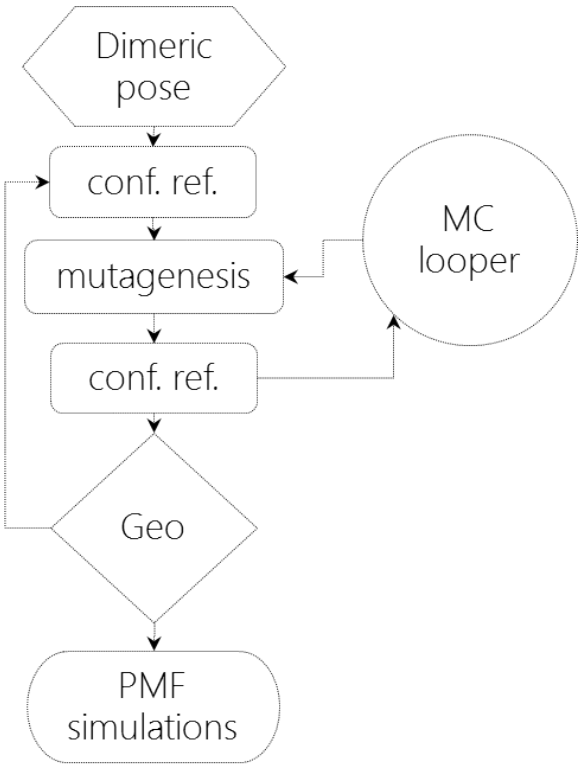


Figure S1. Interface design scheme. The conformational refinement steps consist of a FastRelax mover (repacking only in 2 rounds), a Backrub mover, and a DockingProtocol mover (repacking only with local refinement using soft_rep scoring, with a maximum rigid body perturbation of 3 Å and 2° per round). The mutagenesis step comprised three consecutive RepackMinimize movers (initially with soft_rep scoring and no backbone refinement, and later using talaris13 scoring and performing backbone refinement) and a BackrubDD mover. The GenericMonteCarlo mover was run for 5 loops and all of the output decoys were filtered geometrically and the top few were reused as interface design input. After 4 rounds for BRIC1, 5 rounds for BRIC2 and 7 rounds for BRIC3, the decoys generated were evaluated by the PMF simulations, where the ones exhibiting the highest dissociation work were chosen for loop design.

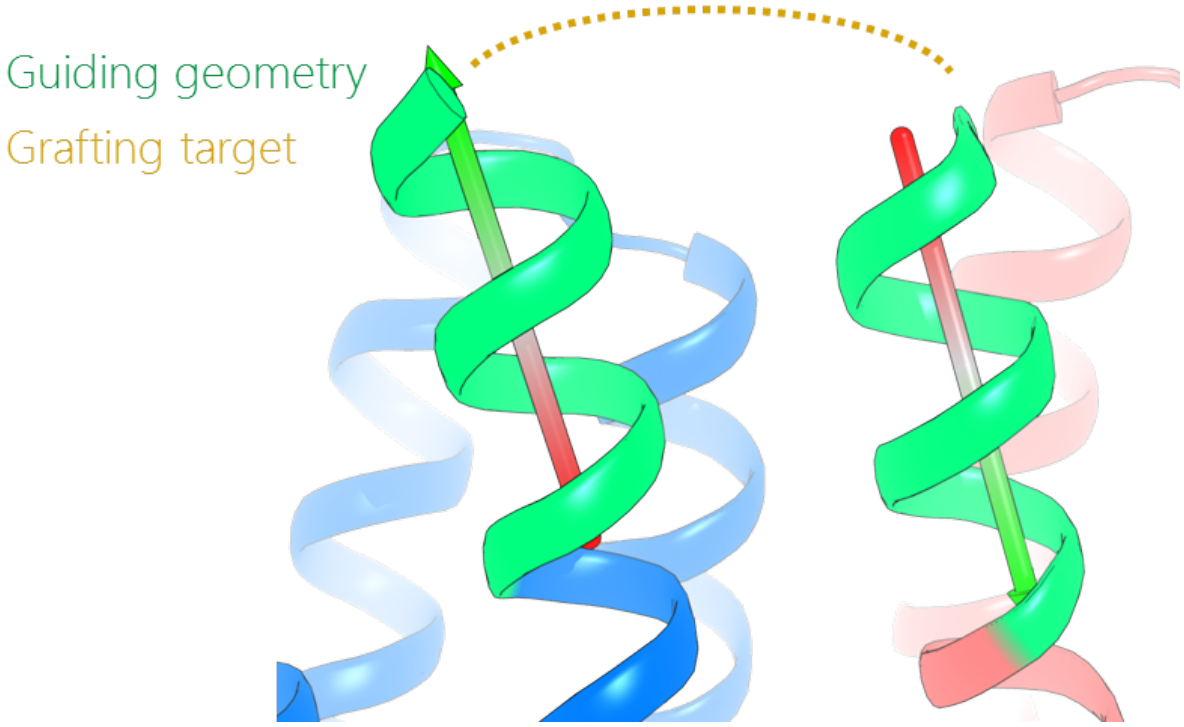


Figure S2. Loop grafting step. Initial loop compositions were fetched from natural structures using an alignment-free geometric matching method that compares the landing site vectors and their associated backbone dihedrals profile (The green stretches of the design interface represent the landing sites).

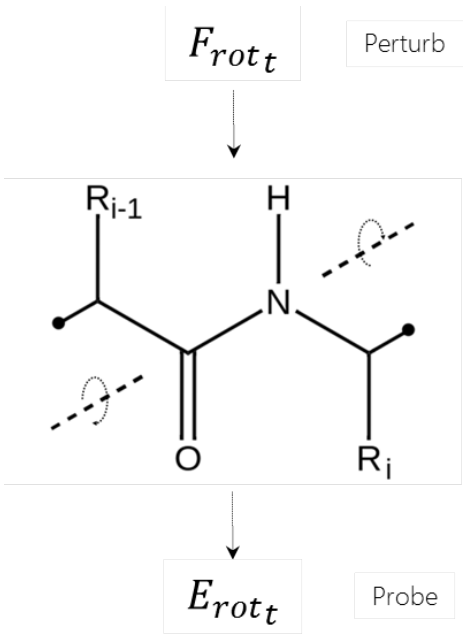


Figure S3. The rotational perturbation axis was taken to be the peptide bond axis, which would force dihedral sampling by diagonal traversal of the shifted-Ramachandran space (i.e. the $(\boldsymbol{\psi_{i-1}}, \boldsymbol{\phi_i})$-space), which was performed while monitoring the instantaneous kinetic energy response**.**

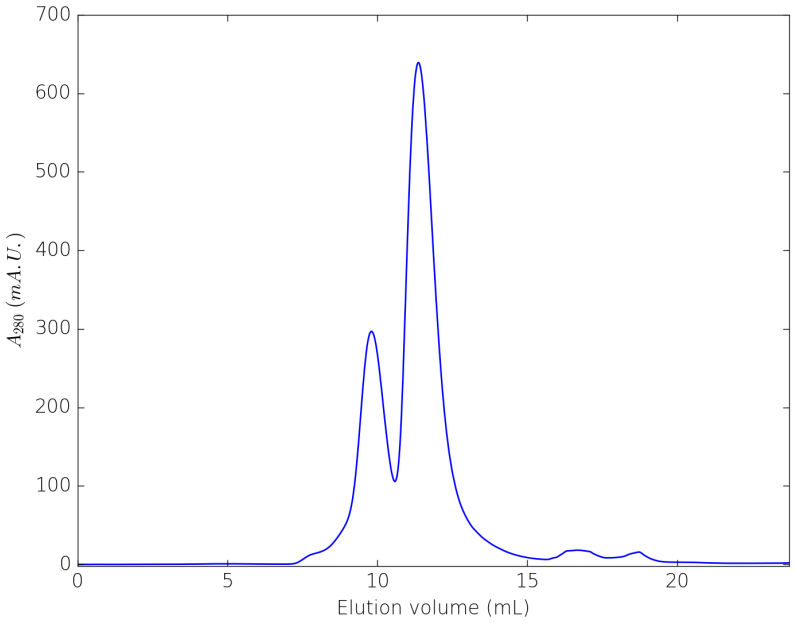


Figure S4. Analytical gel filtration of BRIC1 showing a major monomeric species and a minor dimeric species, using GE Superdex™ 75 10/300.

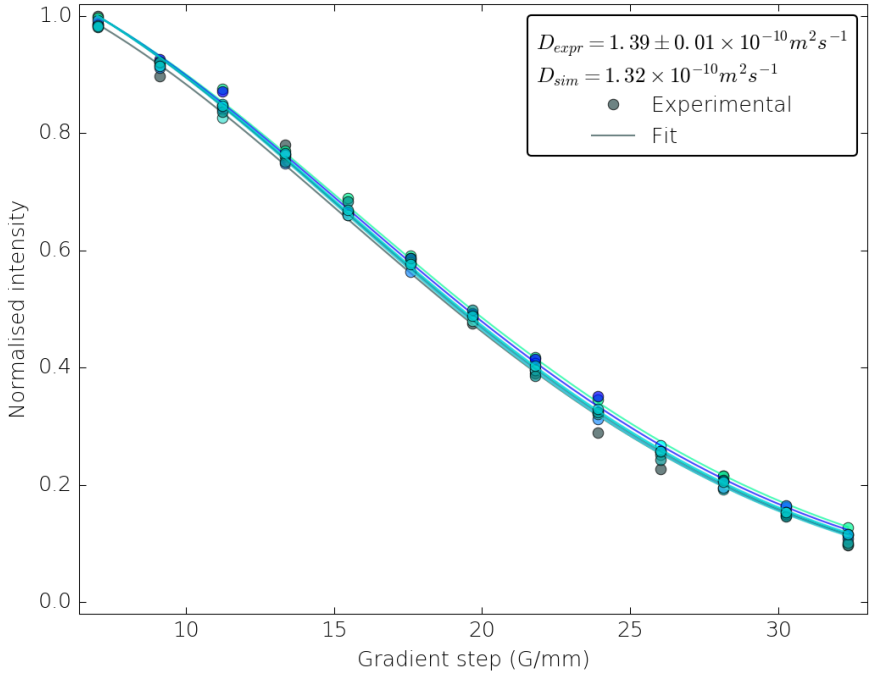


Figure S5. Designed coordinates best explain the diffusion-ordered spectroscopy experiments on BRIC1. Different colours designate data collected for eight different aliphatic proton peaks, where the legend shows the average and standard deviation values of the diffusion coefficient. The predicted translational diffusion coefficient value using the designed coordinates at the same temperature was 1.32×10^{-10}, while that of the swapped dimeric and the swapped monomeric forms were 6.81×10^{-11} and 1.18×10^{-10}, respectively.

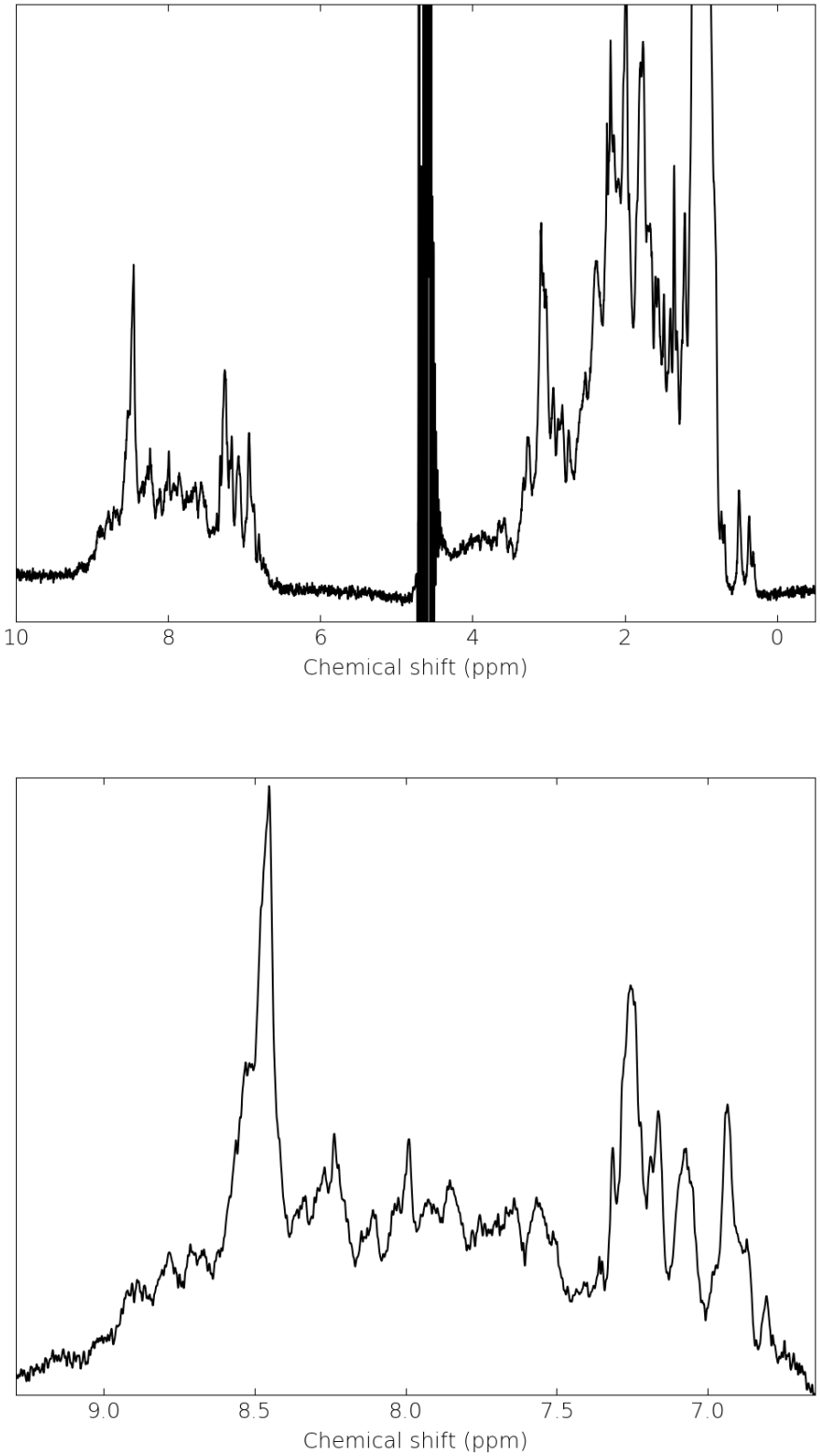


Figure S6. 1D ^{1}H NMR spectrum of BRIC2. The panes show dispersions along the full and the amide spectral ranges, respectively.

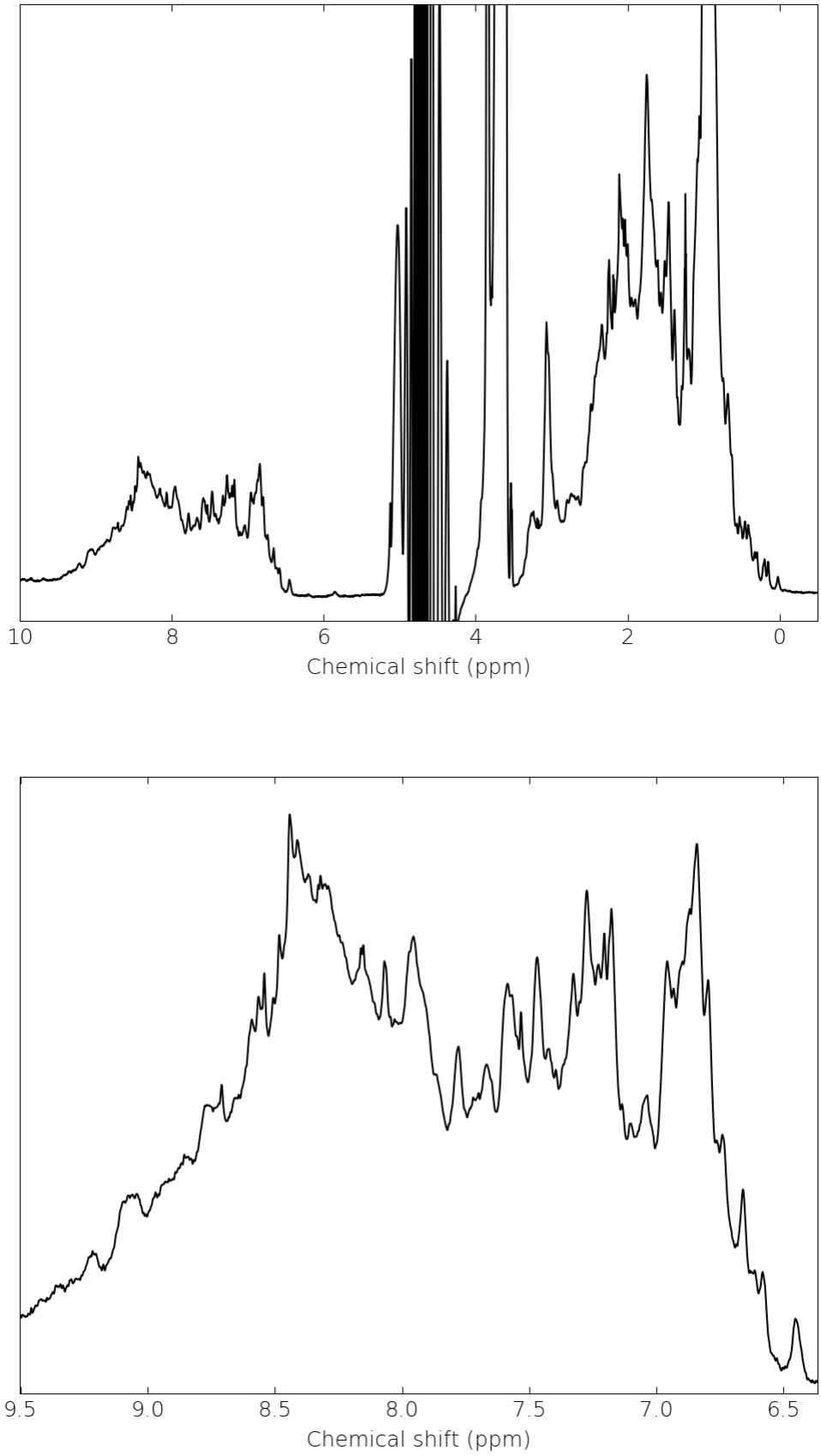


Figure S7. 1D ^{1}H NMR spectrum of BRIC3. The panes show dispersions along the full and the amide spectral ranges, respectively.

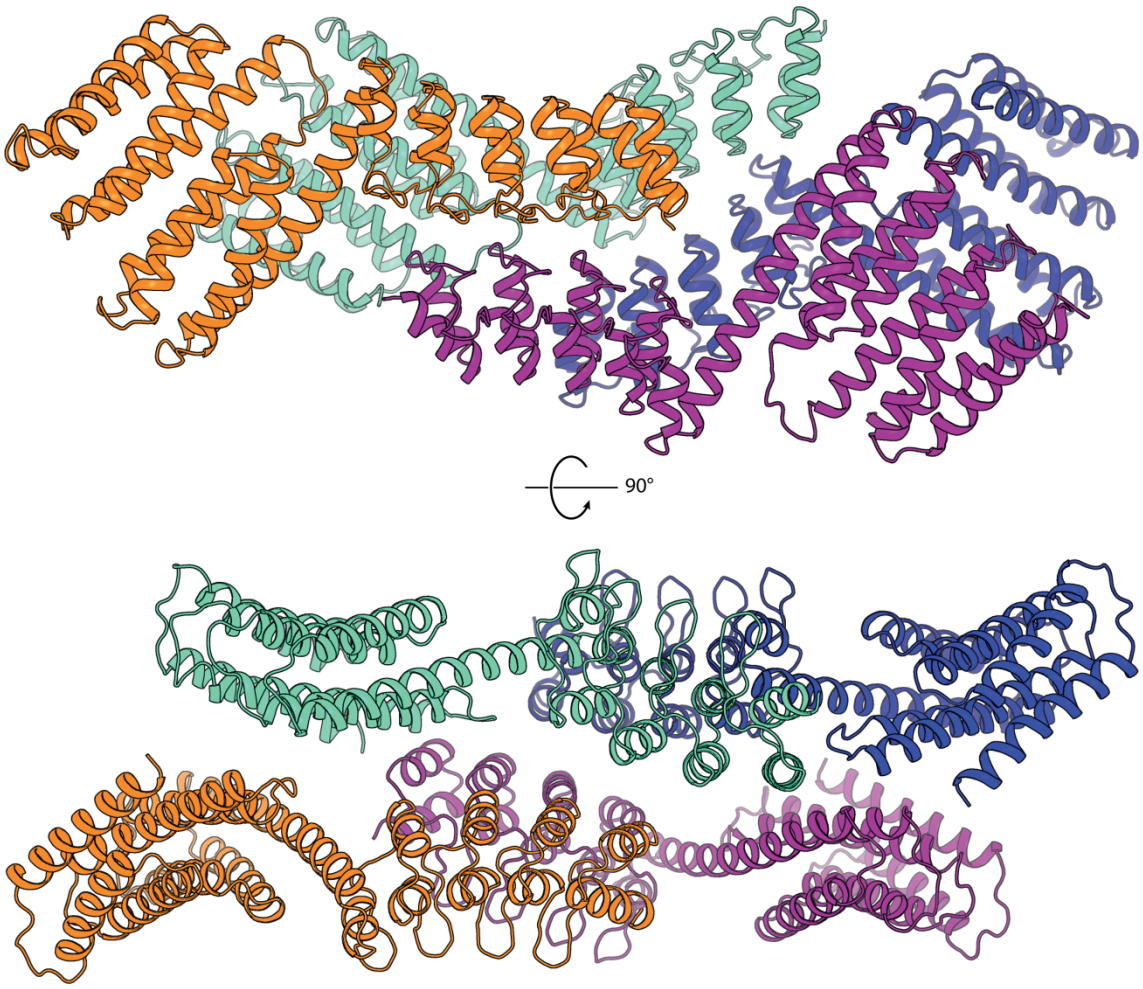


Figure S8. The asymmetric unit of the D12-BRIC2 fusion crystal structure. Different colours show the four chains in the asymmetric unit from two different views.

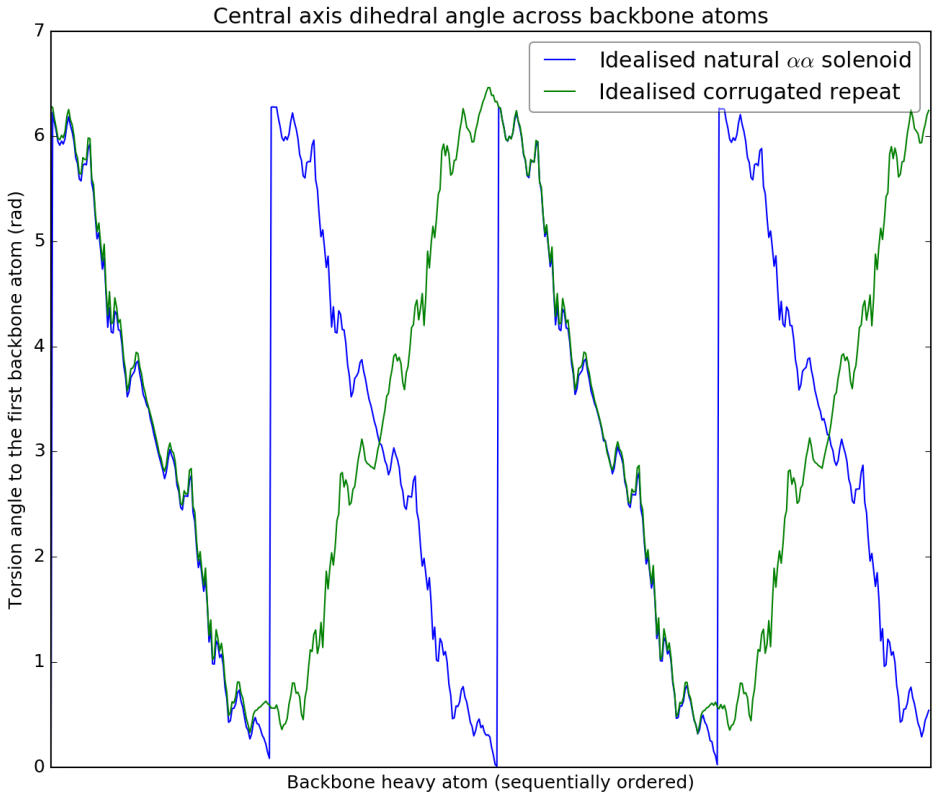


Figure S9. Superhelical axis torsion waveforms of α/α-solenoid *vs.* corrugated fold. The α/α-solenoid shows a clear saw-tooth pattern as compared to the triangle-wave pattern emergent from the more complex corrugated topology. The corrugated topology waveform exhibits double the phase-cycle of the equivalent solenoid illustrating the complexity increment. Both waveforms are distorted by high-frequency interference from the local α-helical pattern.

Table S1. Sequences of the final design constructs aligned structurally against their respective starting templates, with the highlight and underline designating the designed loop and C-terminal cap, respectively.

```
BRIC1           FYQTFFDEADELLADMEQHLLDLVPESPDAEQLNAIFRAAHSIKGGAGTFGFTMLQYAVE
1i5n_tmplt      FYQTFFDEADELLADMEQHLLDLVPESPDAEQLNAIFRAAHSIKGGAGTFGFTILQETTH
                ********************************************************:** :..

BRIC1           LMENMLDFARRGEMQLNTDIINLFLELKDLMQRMLDYYKKPQPCFYQAFFDMADVMLKVM
1i5n_tmplt      LMENLLDEARRGEMQLNTDIINLFLETKDIMQEQLDAYK-----FYQTFFDEADELLADM
                ****:** ****************** **:**. ** **     ***:*** ** :*  *

BRIC1           EQLLKLLVPESPDAAMLNAIFRAAHFIKGAAGTFGFTILQETTHLMENLLDEARRGEMQL
1i5n_tmplt      EQHLLDLVPESPDAEQLNAIFRAAHSIKGGAGTFGFTILQETTHLMENLLDEARRGEMQL
                ** *  ********  ********* ***.****************************

BRIC1           NTDIINLFLETKDIMQEQLDAYKNSEEPDAASFEYICNALRQLALEA
1i5n_tmplt      NTDIINLFLETKDIMQEQLDAYKNSEEPDAASFEYICNALRQLALEA
                *********************************************

BRIC2           YIKKVVDELKELIQNVNDDIKEVEKNPEDMEYWNKIYRLVHTMKEITETMGFSPVALVLE
2lch_tmplt      YIKKVTDELKELIQNVNDDIKEVEKNPEDMEYWNKIYRLVHTMKEITETMGFSSVAKVLH
                *****.*********************************************** ** **.

BRIC2           AIMMLVKLMLNSEIKITSDLIDAVKKMLDMVTRLLDLMVDPNLNEEQYIKMVVDALKILI
2lch_tmplt      TIMNLVDKMLNSEIKITSDLIDKVKKLDMVTRELDKKVS-------YIKKVTDELKELI
                :** **. ************** *** ****** **  *        *** *.* ** **

BRIC2           EAVNVLIKMVEKNPEDMEFWNLIYRLVHVMKEVTETMGFSSVAKVLHTIMNLVDKMLNSE
2lch_tmplt      QNVNDDIKEVEKNPEDMEYWNKIYRLVHTMKEITETMGFSSVAKVLHTIMNLVDKMLNSE
                : **  ** **********:** ******.***:**************************

BRIC2           IKITSDLIDKVKKKLDMVTRELDKMVS
2lch_tmplt      IKITSDLIDKVKKKLDMVTRELDKKVS
                ************************ **

BRIC3           DKVYENVTGLVKAVIEMSSKIQPAPPEEYVPMVKEVGLALRTLLATVDETIPLLPASTHR
3b71_tmplt      DKVYENVTGLVKAVIEMSSKIQPAPPEEYVPMVKEVGLALRTLLATVDETIPLLPASTHR
                ************************************************************

BRIC3           AIELMQELLNIALQLLEIAMKLAQQYVMTSAQQEHKKMMLMAAQVLAEIAKFLLDCITSP
3b71_tmplt      EIEMAQKLLNSDLGELINKMKLAQQYVMTSLQQEYKKQMLTAAHALAVDAKNLLD-----
                 **: *:***  *  *   *********** ***:** ** **:.**  ** ***

BRIC3           CVVYAAVQILVKFVEFMSKFIQPAPPELYVAMVKAVGKALRVLLAIVDMTIPLLPASTHR
3b71_tmplt      -KVYENVTGLVKAVIEMSSKIQPAPPEEYVPMVKEVGLALRTLLATVDETIPLLPASTHR
                  **  *  *** *  **. ******* ** *** ** ***.*** ** ***********

BRIC3           EIEMAQKLLNSDLGELINKMKLAQQYVMTSLQQEYKKQMLTAAHALAVDAKNLLDVIDQ
3b71_tmplt      EIEMAQKLLNSDLGELINKMKLAQQYVMTSLQQEYKKQMLTAAHALAVDAKNLLDVIDQ
                ************************************************************
```

Table S2. Crystallographic structure statistics for BRIC1

Data collection	
Space group	C2
Cell dimensions	
a, b, c (Å)	113.66, 41.95, 58.26
α, β, γ (°)	90.0, 90.46, 90.0
Resolution (Å)	40.84 – 2.50 (2.65 – 2.50)[a]
R_{merge}	0.067 (0.787)
$\langle I \rangle / \langle \sigma(I) \rangle$	12.1 (1.67)
Completeness (%)	99.7 (98.8)
Redundancy	6.55 (6.16)
Refinement	
Resolution (Å)	40.8 – 2.50 (2.79 – 2.50)
No. reflections	9739 (2721)
R_{work} / R_{free} (%)	22.9 / 27.9 (24.9 / 30.2)
No. atoms	1791
Protein	1791
<B-factors> (Å^2)	116.0
Protein	116.0
RMS deviations	
Bond lengths (Å)	0.010
Bond angles (°)	1.03

[a]Values in parentheses are for highest-resolution shell.

Table S3. Solution structure statistics for BRIC1

Restraint Violations[1]	
Inter-Helical Distance restraints (Å)	
All (39)	0.018 ± 0.003
N-terminal interface (11)	
C-terminal interface (15)	
Designed interface (13)	
Dihedral restraints (°)	
All (314)	0.082 ± 0.014
Covalent Geometry	
Bonds (Å × 10^{-3})	1.91 ± 0.04
Angles (°)	0.54 ± 0.01
Impropers (°)	0.88 ± 0.01
Structure Quality Indicators[2]	
Ramachandran Map (%)	98.2 / 1.6 / 0.2
Backbone Heavy Atom R.M.S.D (Å)[3]	
E vs <E>	0.84 ± 0.06
E vs design	1.78 ± 0.11
<E> vs design	1.57

[1] Violations are expressed as RMSD ± SD. The number of each restraint type is shown in brackets.

[2] Defined as the percentage of residues in the favored/allowed/outlier regions of the Ramachandran map as determined by MOLPROBITY (http://molprobity.biochem.duke.edu).

[3] Structures are labeled as follows: E, the final ensemble of 26 structures; <E>, the mean structure calculated by averaging the coordinates of E structures after fitting over ordered residues. RMSD values are based on superimpositions over ordered residues (defined as F1-L225)

Table S4. Crystallographic structure statistics for D12-BRIC2*

Data collection	
Space group	P1
Cell dimensions	
a, b, c (Å)	52.99, 63, 133.54
α, β, γ (°)	97.991, 91.923, 110.815
Resolution (Å)	49.33 – 3 (3.08 - 3.0)
R_{merge}	0.177 (1.308)
<I>/<σ(I)>	7.6 (2.15)
Completeness (%)	99.0 (99.8)
Redundancy	5.6 (5.8)
Refinement	
Resolution (Å)	44.9 – 3 (3.0969 - 3.0)
No. reflections	31544 (2901)
R_{work} / R_{free} (%)	25.09 / 29.19 (35.56 / 35.97)
No. atoms	11544
Protein	11420
Water	51
Ligands	73
<B-factors> (Å2)	97.91
Protein	98.06
RMS deviations	
Bond lengths (Å)	0.003
Bond angles (°)	0.66
Ramachandran plot	
Favoured	95.37
Allowed	4.22
Outliers	0.41

[a]Values in parentheses are for highest-resolution shell.

*Diffraction data from two crystals from the same drop were merged to improve completeness.

Chapter 4: Mapping Local Conformational Landscapes of Proteins in Solution

Status: In revision

Mapping Local Conformational Landscapes of Proteins in Solution

M. ElGamacy[1], M. Riss[2], H. Zhu[1], V. Truffault[1], and M. Coles[1]*

1. Dept. of Protein Evolution, Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

2. Dept. of Informatics, Technical University of Munich, Boltzmannstrasse 3, 85748 Garching, Germany

* Corresponding Author: **murray.coles@tuebingen.mpg.de**

Summary

The ability of proteins to adopt multiple conformational states is essential to their function and elucidating the details of such diversity under physiological conditions has been a major challenge. Here we present a generalized method for mapping protein population landscapes by NMR spectroscopy. Experimental NOESY spectra are directly compared to a set of expectation spectra back-calculated across an arbitrary conformational space. Signal decomposition of the experimental spectrum then directly yields the relative populations of local conformational microstates. In this way, averaged descriptions of conformation can be eliminated. As the method quantitatively compares experimental and expectation spectra, it inherently delivers an R-factor expressing how well structural models explain the input data. We demonstrate that our method extracts sufficient information from a single 3D NOESY experiment to perform initial model building, refinement and validation, thus offering a complete *de novo* structure determination protocol.

Introduction

In order to function, proteins must adopt a distinct three-dimensional fold. However, a vast range of protein functions, including catalysis, molecular recognition and allosteric signalling, also rely on their ability to adopt various local conformations within this structural scaffold. Understanding these processes therefore requires not only accurate descriptions of protein structures, but also their conformational diversity. NMR spectroscopy is uniquely placed to address these issues, offering atomic-resolution data on samples in native-like physical states. Time averaging of NMR parameters has long been exploited to localise and characterize the timescales of internal dynamics [(1)]. However, the data is also ensemble-averaged over all molecules in the sample volume and should thus provide information on the nature and population of underlying conformational microstates. Accessing this data has long been a goal in NMR spectroscopy [(2-5)].

Here we aim to elucidate the propensities of individual microstates by means of spectral decomposition. Systematic back-calculation of expectation spectra across a conformational space allows reconstruction of the experimental spectra. In NMR spectroscopy, the richest source of structural data are NOESY spectra, which report on inter-proton distances within a detection limit of 5 to 6 Å. Due to this short spatial range, a large fraction of NOESY intensity can be explained within short, linear sequence fragments. Each such fragment thus represents a sub-space that could be searched systematically to provide detailed information on local dihedral angles and their distributions. Moreover, comparison of the back-calculated and experimental data would provide a quantitative quality measure: an NMR R-factor.

A difficulty in realising this approach lies in the nature of NOESY data itself. The information content of NOESY spectra is very unevenly distributed across the observed intensities, thus small, informative peaks can be overwhelmed by inaccuracies in back-calculation and spectral artefacts. For this reason, quantitative comparison of back-calculated and experimental spectra has been far less applicable in NMR structure determination than equivalent measures used in crystallography [(6-8)]. Here we show that these obstacles can be largely averted in the 3D CNH-NOESY experiment [(9)]. This is an

implementation of a ^{13}C-HSQC-NOESY-^{15}N-HSQC where the indirect proton dimension has been omitted, thus displaying contacts to backbone amide protons in a well-resolved ^{13}C dimension. It exploits the higher dispersion and more homogeneous effective linewidths of the heteronuclei, while suppressing water-exchange cross-peaks and obviating the need for stereospecific proton assignments. Crucially, it intrinsically lacks large, uninformative diagonal peaks and the associated baseline and truncation artefacts. Combined, these represent decisive advantages in the accuracy of back-calculation.

In this work we demonstrate that a single 3D CNH-NOESY spectrum contains sufficient information to define population maps of local dihedral sub-spaces. Analytical decomposition expresses the experimental spectra as a linear combination of elements of a features set of back-calculated spectra. In this way, both the reliance on a knowledge-base and the interpretation of spectra in terms of peak or assignment lists can be eliminated. This conformational mapping provides highly detailed data for model building and refinement, with progress monitored by a quantitative R-factor. We validate this method against human Ubiquitin (hUb), widely considered the gold standard for NMR-based protein structure determination [5, 10-12]. We further demonstrate the generality of the method by solving the structures of four example proteins.

Results

An R-factor from CNH-NOESY data

We have adapted existing routines to back-calculate CNH-NOESY spectra, obtaining 1D ^{13}C strips for each backbone amide proton. These are compared directly to equivalent strips extracted from the experimental 3D matrix. An R-factor expressing the discrepancy between experimental and expectation spectra is readily calculated as the fractional root mean squared residual (see Experimental Section). This R-factor is analogous to its crystallographic counterpart, except that it is calculated on a per-residue basis. Our back-calculation routines very accurately reproduce the intensities and line-shapes of experimental CNH-NOESY data collected for hUb (Figure 1a). The back-calculated spectra are also highly sensitive to backbone and sidechain dihedral angles (Figure 1b and Supplementary Figure S1), a prerequisite for conformational mapping.

NOESY intensities are time and ensemble averages over the conformational microstates sampled during the measurement. For this reason, R-factors improve with an accurate and comprehensive description of the ensemble. We demonstrate this for hUb using three reference ensembles that have been compiled according to different metrics. Two have been compiled to elucidate internal motions: the 2KOX ensemble using a large set of residual dipolar coupling (RDC) data [11], and the 2NR2 ensemble according to minimum under-restraining, minimum over-restraining criteria and including S^2_{NH} order parameters derived from relaxation data [5]. The third, 2MJB, provides a static control ensemble close to the average structure [12]. High R-factors are obtained where these ensembles either over- or under-estimate the conformational diversity. Moreover, the use of a residue-wise R-factor in evaluating the ensemble can localize such diversity (Figure 1b).

Mapping local conformational spaces

For the CNH-NOESY, the vast majority of cross-peak intensity can be explained by intra-residue contacts and those to the immediately preceding residue. The R-factor for residue *i* is thus strongly dependent on conformation in a shifted Ramachandran space defined by the backbone dihedral angles ψ_{i-1} (here denoted υ_i) and ϕ_i. This is extended by including the relevant sidechain rotamers up to χ^1_i and χ^2_i, representing a periodic space within a dipeptide fragment that can be searched exhaustively (Figure 2a). The back-calculated spectra for these systematically sampled conformers constitute a features set that can be used to decompose the experimental spectra (Figure 2b). Here we characterize the solution ensemble as a linear combination of elements of the features set, weighted by their respective populations. Calculating these weights is analogous to parts-based representation of complex spectral mixtures often encountered in other fields of spectroscopy [13].

Decomposition of the experimental spectrum can be framed as a positive matrix factorisation problem [14]. The features matrix is represented by $\boldsymbol{W}$ comprising back-calculated spectra for *l* conformers, resolved to *m* points along the ^{13}C dimension. A solution can thus be found for a vector of weights $\boldsymbol{H}$ in order to reconstruct the observed experimental spectrum $\boldsymbol{V}$:

$$V \approx WH, \quad V \in \mathbb{R}_{\geq 0}^{m \times n}, W \in \mathbb{R}_{\geq 0}^{m \times l}, H \in \mathbb{R}_{\geq 0}^{l \times n}$$

where n=1 if one spectrum is considered per residue. To map the energetic landscape of the conformational space, the yielded estimates of the population weights can in turn be expressed via a Boltzmann factor relative to a reference conformer (Materials and Methods). Examples of these conformational maps for hUb are shown in Figure 3.

An inherent question in factorization methods is the uniqueness of the solution. Here uniqueness is limited by cross-peaks that cannot be explained within the dipeptide space, but overlap with peak positions in the features set. The larger the fraction of such contamination, the more difficult finding a unique solution becomes. Given the low intensity of potentially contaminating peaks and the high dispersion of the ^{13}C dimension, the extent of contamination is usually minor. The extreme case of contamination is the coincidence of ^{15}N-HSQC positions for two or more residues, resulting in overlap of the experimental strips. Such a situation can be solved by concatenation of the features sets of the overlapped residues and solving in a multiple-dipeptide space. Figure 2c shows a typical example of this situation, where conformational maps have been obtained for two overlapped residues.

Structure determination with conformational maps

The conformational maps obtained from spectral decomposition provide rich information for structure determination. At the simplest level, global minima can provide local torsion angles sufficient for model building. These initial dihedrals constitute an agnostic starting point, as they are derived directly from the data without recourse to heuristics or conformational databases. A unique feature of this method is the deployment of R-factors as an objective convergence test that captures both local and long-range contacts. The latter can be isolated by examining the difference between R-factors obtained for a linear peptide fragment and those from the full, folded model. We term this measure the fold factor (F), and it should be negative if the model explains long-range contacts well. Figure 4 shows that the average fold factor (F_{mean}) is a sensitive overall measure of correct folding, while the sequence profile can localize misfolded or poorly defined regions. Owing to this independent measure of convergence, any routine can be used to build initial models. Here we employ either a Rosetta-based protocol

(Materials and Methods) or a purpose-designed molecular dynamics routine: Simulated Annealing Replica *Seilschaft* (SARS) for initial model building (Materials and Methods).

In order to construct high-resolution ensembles with accurate description of the underlying microstates, the experimental information contained in the conformational maps can be encoded in two possible representations. The first is a temporal equilibrium ensemble derived from molecular dynamics simulations. Here the conformer probability distribution is imposed via a grid-based dihedral energy correction term [15] (CMAP; Figure 3 and Materials and Methods). These override the standard force field CMAPs with bespoke ones on a per-residue basis, providing an experimentally augmented ensemble representation. The other is to aggregate a set of frames from a generalized ensemble using the standard force field, based on an R-factor selection criterion, providing a wider coverage of the phase space.

We name this method of *de novo* structure determination CoMAND (for Conformational Mapping by Analytical NOESY Decomposition). In addition to hUb, we present examples for four structure determination projects from our Institute. U3Sfl (125 amino-acids) is a protein designed as a chimera of sub-domain sized fragments, KH-S1 (170 amino-acids) is a fusion construct of the KH and S1 domains of E. coli exosomal polynucleotide phosphorylase. MlbQ is a protein implicated in self-resistance to endogenous lantibiotics in actinomycetes [16]. The final example, polb4, is a protein designed to reconstruct the polymerase beta N-terminal domain using two unrelated peptide fragments and is presented here as a *de novo* structure determination.

For all five proteins, we first built starting models. For U3Sfl, KH-S1 and hUb, we extracted backbone dihedral angles from the factorization minima and used these for fragment picking in a Rosetta protocol (Materials and Methods). For MlbQ and polb4 we applied SARS, supplementing the CHARMM36 forcefield with bespoke conformational maps, starting from completely extended chains. For MlbQ, folding was accelerated by the addition of 10 unambiguous NOE distance restraints. We used the average R-factor across the full length of the protein as a criterion for selecting models, choosing a single Rosetta decoy or a single frame from the SARS runs. These models were very similar to respective reference structures (Figure 5). For U3Sfl this was a structure we had previously solved by manual analysis (RMSD over backbone atoms 1.98 Å). For KH-S1,

crystal structures are available for homologues of the individual domains (4AM3; 1.48 Å and 4NNG; 1.67 Å). For MlbQ, this was the published solution structure (2MVO; 1.92 Å). As no structure has previously been solved for polb4, we used the design target as a reference structure (RMSD over backbone atoms 1.48 Å).

For refinement we conducted unrestrained molecular dynamics simulations in explicit solvent for microsecond timescales, seeded by the starting model. This was followed by a frame-picking procedure that employs a greedy optimizer to minimize the average R-factor across the ensemble. Given the wealth of structural and dynamics data available for hUb, we compiled such a refined ensemble and compared it to the reference ensembles (2KOX, 2MJB and 2NR2). The resulting ensemble of 20 conformers shows better correlation to experimental NH order parameters than the literature ensembles. It is also comparable in predicting experimental scalar couplings and RDCs to ensembles that have been built on one or more of these observables plus thousands of NOE restraints (Figure 6 and Supplement). This demonstrates the depth of the structural information captured when NOESY spectra are analysed holistically.

Discussion

NOESY spectra can be seen as an encoding of a proton-proton contact map with an approximate upper distance limit of 5 Å. If correctly decoded as a set of distance restraints, this information is sufficient to solve the structure with high accuracy and precision. However, crowded spectra and the consequent spectral overlap mean that the encoding is ambiguous. Spectral editing, for example via additional frequency dimensions, can only partially alleviate this problem, often at considerable cost in experiment time [17]. The consequences of this ambiguity are not only that individual cross-peaks cannot be uniquely assigned – i.e. attributed to a specific proton-proton contact – but also that cross-peaks may comprise significant intensity from several contacts. Conventional NMR structure determination protocols interpret NOESY spectra through peak-picking, assignment and conversion into distance restraints under a paradigm of one peak; one assignment; one restraint. Even automated routines that specifically consider ambiguities will resolve to a single effective restraint per picked peak. This represents a compromise that is not justified by the underlying nature of the data, affecting either the accuracy or precision of distance estimates. In contrast, the

CoMAND method makes no interpretation of cross-peaks, and thus stands outside this conventional assignment paradigm.

Given the ambiguity of NOESY data, the incorporation of unambiguous data from other sources is advantageous in NMR structure determination protocols. Particularly useful are data that define local dihedral angles (e.g. scalar couplings), as these are poorly defined by imprecise NOE-based distance estimates. Backbone dihedral angle predictions derived from chemical shift heuristics using the program TALOS are very widely used [18]. For example, the CS-Rosetta approach exploits this data to build structural models within the Rosetta framework [19]. In contrast to previous methods, we show that direct signal decomposition can yield backbone dihedrals unambiguously and without heuristics. Moreover, we demonstrate that the NOESY data can be leveraged to map the underlying conformational landscape in a systematic fashion. A further key advantage over existing methods is that the whole process of structure determination, including resonance assignment, model building, refinement and conformational mixture elucidation can be objectively assessed by the R-factor as a single metric.

In recent years development of analysis methods in solution NMR of proteins has been driven by the need to make automation more reliable, while using less data and extending the range to larger proteins and more difficult cases, such as membrane proteins. CoMAND contributes to this effort in that it leverages a small set of spectra on a single sample into a high-resolution structure and is therefore applicable where protein concentration or stability are limiting. As the method involves minimum user intervention after the resonance assignment stage, it is also intrinsically suited to automation. However, the most unique feature of the method lies in the power to obtain accurate descriptions of protein conformational ensembles. We therefore anticipate that the method can be applied to studying ligand binding and allosteric processes, promising to elucidate subtle conformational changes in an unprecedented level of detail.

Materials and Methods

NMR Spectroscopy

Backbone and sidechain assignments for *de novo* structure determinations were obtained using standard triple resonance experiments. For human Ubiquitin literature values were used. Slight correction of ^{13}C shifts against the respective CNH-NOESY spectra was necessary to account for calibration differences between spectrometers and spectrum types. 3D CNH-NOESY spectra were acquired at 800 MHz on a Bruker AvanceIII spectrometer equipped with room temperature probehead. Indirect ^{13}C dimensions were typically acquired with ~100 time increments and processed with linear prediction and zero filling to 256 data points. The ^{13}C sweep width was set to cover aliphatic carbon resonances; i.e. ~10-73 ppm, resulting in a resolution of ~30 Hz per point. At this resolution, $^{1}J_{CC}$ couplings are unresolved and the spectra were run in non-constant time mode. Broadband ^{13}C pulses were used to excite aromatic resonances and these were folded into the aliphatic window without phase inversion.

CNH-NOESY spectra were analysed by extracting one-dimensional ^{13}C sub-spectra chosen from a search area centred on assigned ^{15}N-HSQC positions (typically 1-3 points in each dimension). As these sub-spectra contain only cross-peaks to a specific amide proton, choosing the strip with highest integral maximises the signal-to-noise. Residues with overlapping search areas were examined separately. In most cases strips with acceptable separation of signals could be obtained. Where this was not possible the residues were flagged as overlapped and a joint strip constructed by summing those at the estimated maxima of the respective components. A set of strips well separated from assigned HSQC positions were averaged to define a global noise level for the spectrum.

NOESY back-calculation

In order to back-calculate 3D CNH-NOESY spectra we modified the program SPIRIT [20] by porting it to C++ and extending it to accommodate any combination of proton and heteronuclear dimensions. We name this program SHINE, for Simulation of Hetero-Indirect NOESY Experiments. The calculations are based on a full relaxation matrix and thus account for spin diffusion in static structures. Internal motion of the protein is treated by ensemble averaging over *n* contributing microstates, effectively applying an *n*-

state jump model of motion, where the life-time of a microstate is assumed to be long compared to the interconversion time. This does not account for true time-averaged phenomena, such as motion-mediated spin-diffusion, which are treated as negligible for the current application. Inputs for the calculations are a chemical shift list, a test structure and a set of simulation parameters. The latter are largely spectral details, such as spectrometer frequencies and sweep widths, which are extracted automatically from the corresponding experimental files (Bruker format), but also include an estimate of a global molecular correlation time. Resonances are modelled as gaussians, with ^{13}C linewidths assigned on a class basis, taking into account unresolved $^{1}J_{CC}$ couplings. Here it should be noted that the short acquisition times in an indirect ^{13}C dimension (<10 ms on an 800 MHz spectrometer) mean that effective lineshapes are largely governed by apodisation of the time domain. They are thus considerably more homogeneous than for a proton dimension.

The computational demand of NOESY back-calculation depends on the number of protons in the relaxation network. For this reason, we employ sub-structures containing <150 atoms. These can be linear peptides or fragments of a folded structure. Linear peptides are typically tri- or penta-peptides where the test residue is in the second last position. Fragments are compiled at the residue level; residues are included in the sub-structure if they contain a proton within a given radius (typically 5 Å) of a target residue proton. The output is a one-dimensional strip displaying contacts to a single backbone amide proton. In this mode, back-calculation typically takes less than 20 ms per conformer on a single processor core. The program also outputs a list of peak intensities that can be used to build multi-dimensional spectra suitable for viewing in SPARKY [21], with annotation of individual cross peaks.

Calculation of features sets is performed for each residue for which an experimental spectrum is available. The starting structures are linear peptide fragments extracted from an arbitrary structural model. In the current work these were tripeptides centred on the test residue. For back-calculation of contacts to the amide proton of residue *i* this peptide is modified through a set of torsion angles in a shifted Ramachandran space: angles ψ_{i-1} (here denoted υ_i) and ϕ_i and up to two sidechain χ angles. The backbone angles were searched at 10° granularity, while sidechain angles

sample all staggered rotamers. This results in a maximum of 11664 conformers. No checks for steric clashes are applied to test if a conformer is physically reasonable and bond lengths and angles remain constant throughout. An exception is for proline residues where the search space is restricted to a physically realistic range in ϕ_i. As proline residues lack an amide proton, their features sets are compiled by back-calculation of the amide proton of the previous residue, which is the most sensitive reporter on the υ angle of proline. The set of back-calculated spectra for each residue are stored as a single file. For residues flagged as overlapped, the features set files are concatenated to create a joint set.

Calculation of R-factors

We define the R-factor as the relative RMS residual between an experimental spectrum and the expectation spectrum back-calculated from a structural model. The use of RMS is in analogy to the quality factor (Q-factor) calculated for residual dipolar coupling data [22]. The use of RMS tends to emphasize large outliers relative to the R-factor used in crystallography, which averages absolute differences between experimental and back-calculated structure factors. It also provides a convenient definition of the optimum scaling factor for the back-calculated spectrum s_{calc}, which can be calculated as:

$$s_{calc} = \frac{\boldsymbol{v}_{exp} \cdot \boldsymbol{v}_{calc}}{\|\boldsymbol{v}_{calc}\|^2}$$

where $\boldsymbol{v}_{exp}$ and $\boldsymbol{v}_{calc}$ are the experimental and back-calculated spectral intensities vectors, respectively. The R-factor is then calculated as:

$$\boldsymbol{R} = \frac{\sqrt{\sum_i^n (\boldsymbol{v}_{exp_i} - \boldsymbol{s}_{calc}\boldsymbol{v}_{calc_i})^2}}{\sqrt{\sum_i^n \boldsymbol{v}_{exp_i}^2}}$$

The theoretical range of the metric is from 0 to 1, however the maximum value can only be reached if there is no correspondence between peaks in the experimental and

expectation spectra, in which case the scaling factor, s_{calc}, will approach zero. The practical minimum value is limited by the RMS noise on a per-residue basis.

Calculation of fold factors

Back-calculation for a structure can be carried out either on a linear peptide or as a fragment of the full structure. The R-factor can therefore be calculated on either basis. Calculation for peptides cannot explain any peaks outside the linear context, whereas all peaks should be explained in a fragment. The difference between fragment and peptide R-factors therefore reports on the fraction of cross-peak intensity that can only be explained in the folded structure. Here we report this difference as a per-residue fold factor, F:

$$F_i = R_i^{full} - R_i^{pentap}$$

where R_i^{full} is based on a fragment from the full model and R_i^{penta} is the penapeptide based R-factor for residue *i*. As R-factors should decrease as more of the cross-peak intensity is explained, the fold factor will be consistently negative for well-folded structures. High positive values are indications of misfolding, while continuous stretches of values close to zero should only be seen for unstructured regions.

Factorisation and CMAP construction

The dipeptide conformational space was sampled according to the following granularity: $\Delta\upsilon = 10°$, $\Delta\phi = 10°$ $\Delta\chi^1 = 120°$, $\Delta\chi^2 = 120°$. The experimental vector $\boldsymbol{v}$ consisted of m = 256 data points of the acquired CNH-NOESY strip (i.e. NOE intensities *vs.* ^{13}C chemical shift), while W contained all of the back-calculated spectra of the l conformers sampled. With the aim of solving for the positive factors vector $\boldsymbol{h}$ that weights each column of W to best explain $\boldsymbol{v}$. The principal solution can be defined as:

$$\min_{h \geq 0} \|\boldsymbol{v} - W\boldsymbol{h}\|^2$$

and $\boldsymbol{h}$ can be also derived directly once the Moore-Penrose pseudo-inverse of the back-calculated spectra matrix, W^+, is computed as:

$$\boldsymbol{h} = \sqrt{\boldsymbol{v}^2 W^{+2}}$$

The uniqueness of the solution in positive matrix factorization is limited by the ranks of the component matrices with the upper bound of *min(m,n)* [23]. Here, as rank(W) >> rank(***v***) = rank(***v***) = 1, the limiting factor is the rank of ***v***. Thus a 2-conformer block-wise factorisation was sought, being closest to this limit. The 2-conformer solution is also computationally tractable for handling the fine degree of conformational sampling described above. This solution offers a recovered spectrum that has a higher or equal weight against any single-conformer solution. The former two-component weight $\boldsymbol{h}_{ref}$ was used as the highest propensity reference state for estimating the relative normalised propensities of every other available conformer $\boldsymbol{h}_i$. A Boltzmann factor can be directly used to estimate the energy of every conformer according to:

$$\frac{h_{ref}}{h_i} = e^{\frac{\Delta G_{ref \to i}}{kT}}$$

The above procedure was performed across the (υ, ϕ) planes at every (χ^1, χ^2) combination. And lowest R-factor yielding plane was the one embedded into the CHARMM36 force field to generate MD ensembles with experimentally derived backbone dihedrals energy surfaces.

Model building using Rosetta

The Rosetta software package [24] was used to build structural models using backbone dihedral restraints derived from conformational mapping (version 3.6). First, a Rosetta dihedral angle constraint file (.cst) was compiled. For each residue position *i*, a MultiConstraint field was written to comprise both dihedral angles υ_i and ϕ_i. When multiple dihedral angles were possible for one residue position, an AmbigousConstraint field was used to include all possibilities. The Rosetta fragment picking program fragment_picker [25] was used to select 3mer and 9mer fragments satisfying the dihedral restraints from the PDB database. The DihedralConstraintsScore weight used was 500 and the minimum allowed 100. The SecondarySimilarity (weight 150, minimum 1.5) and RamaScore (weight 150, minimum 1.5) both used psipred. FragmentCrmsd was not used. Other fragment_picker parameters were defaults. The fragment database vall.jul19.2011 was searched and 200 3mer and 200 9mer fragments were picked for each position in the protein. The Rosetta ab initio folding program AbinitioRelax was then used to fold the

target proteins using these picked fragments (flag file: -ex1, -ex2 -use_input_sc, -flip_HNQ, -no_optH false, -silent_gz 1). Typically, 20,000 decoys were generated for each target protein.

SARS simulations

The SARS framework was designed to provide robust and efficient conformational sampling without relying on any bioinformatic data, with the only input being CNH-NOESY-acquired dihedral preferences encoded as residue-wise biasing potentials into a standard atomistic force field. The sampling acceleration scheme is based on the assumption that the native structure lies at the global minimum of the potential energy surface that is led to by successively deeper local minima. The algorithm initiates multiple replicas from the same fully extended peptide chain starting system. It then combines alternating rounds of simulated annealing between two temperature baths with conjugate gradient minimisation. Each such round constitutes a search step in a collective swarming behaviour that guides the configuration exchange between replicas whenever a new minimum is reached. In this way, all of the high-energy replicas follow the lead of the lowest energy one. The implementation details and convergence properties of the SARS method will be detailed in a separate publication.

Molecular dynamics of human ubiquitin

An initial low-resolution model generated by ROSETTA from the CNH-NOESY-acquired dihedral restraints was taken as the input coordinates for molecular dynamics simulations. The standard trajectories were acquired from 10 independent replicas conducted using the standard CHARMM36 force field [26]. In contrast, the guided trajectories were acquired from 10 independent replicas where the systems were built using an augmented CHARMM36 with bespoke CMAP potentials. The CMAP potentials were directly constructed on a per-residue basis, in the shifted Ramachandran space (υ_i, ϕ_i) from the energy maps as described above. The energy maps where scaled by an arbitrary factor depending on restraint level required, and the standard CMAPs were adopted wherever a bespoke one was not available. These residues with unmodified cross-terms were M1, Q2, G10, P19, E24, I30, P37, P38, G53 and G76.

All of the simulations were performed in explicit TIP3P water as solvent, containing 9226 solvent molecules each, and neutralised by 0.15 M Sodium Chloride in a cubic periodic unit cell. Energy minimisation was performed through 5000 steps of conjugate gradient minimisation, which was followed by 30 ns of NPT equilibration. The time step was set to 2 fs and a Langevin Piston was set to 1 Atm at oscillation period of 200 fs and damping period of 50 fs and a temperature of 298 K. A Langevin thermostat was accordingly set with damping coefficient of 1 ps^{-1}. A nonbonded interactions distance cutoff was set to 12.0 Å at a switching distance of 10.0 Å with all nonbonded force and pair list evaluations were performed every timestep, and long-range electrostatics were computed using the Smooth Particle Mesh Ewald method [27] as implemented in the NAMD engine [28]. Data was collected from the ensuing NVT trajectories, dumping coordinates every 5 ps for analysis. Frame picking was done from the 10 production trajectories of 30 ns each based on the standard force field that would represent a steady state canonical ensemble.

Ensemble building

To compile the final CoMAND ensemble for hUb, we performed frame picking from an equilibrium ensemble, such that the compiled conformers belong to microstates of minimal free energy and maximal entropy at the target temperature. This should provide more physically realistic final models compared to those collected from constrained tempering schemes with unrealistic Hamiltonians. To pick frames from the production trajectories we applied a greedy algorithm aimed at minimising the average R-factor for all residues where experimental CNH-NOESY strips were available. For consistency, residues lacking experimental strips (M1, Q2, G10, P19, E24, I30, P37, P38, G53, G76) were excluded from the following comparisons with other datasets.

Validation versus NMR observables

Expectation NMR observables, R-factors and fold-factors were calculated for the CoMAND and reference hUb ensembles. For uniformity, only the first 20 conformers from each ensemble were considered for the comparisons shown in the figures.

For backbone amide order parameters, frames were aligned into a singular molecular frame of reference that best fits backbone atoms of residues 1 through 70. The order

parameter, S^2_{NH}, was directly computed according to the method described by Nederveen and Bonvin [29] through the following equation:

$$S^2_{NH} = \frac{1}{2}\left(3 \sum_{i=1}^{3} \sum_{j=1}^{3} \langle \mu_i \mu_j \rangle^2 - 1 \right)$$

where μ_i is the normalised internuclear NH bond vector in the molecular frame of reference and $\langle * \rangle$ represents the ensemble-averaged value. Expectation H^N-H[?] scalar coupling constants were calculated according to the following Karplus function:

$${}^3J_{H^N H^\alpha} = 6.98 \cos^2 \theta - 1.38 \cos \theta + 1.72$$

where θ is the $H^N - N - C^\alpha - H^\alpha$ dihedral angle. These were compared to literature experimental values [30]. Deviations from experimental residual dipolar couplings and an overall Q-factor for the CoMAND ensemble were calculated using the program PALES [31] using 996 backbone coupling published for 2MJB across four alignment media [12].

Acknowledgments

We thank Birte Höcker and Remco Sprangers and their co-workers for providing samples for U3Sfl and KH-S1. Programming of the SHINE package was supported by Professors Alois Knoll and Horst Kessler at the Technical University, Munich. We thank Andre Noll (MPI for Developmental Biology, Tübingen) for helpful discussions.

Author contributions: MG designed and implemented spectral decomposition routines and the SARS protocol, performed molecular dynamics calculations and analysed their output. MC implemented back-calculation routines. MR programmed SHINE. HZ performed Rosetta calculations and analysis. VT and MC acquired and analysed NMR spectra. MC conceived and directed the project. The manuscript was written by MC and MG and edited by all the authors.

Declaration of Interests: This work was supported by institutional funds of the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability: Coordinates for the CoMAND ensemble for human ubiquitin have been deposited in the Protein Data Bank under the accession number (TBC).

Supplementary Materials

Fig. S1. CNH-NOESY based R-factors are highly sensitive to local conformation.

Fig. S2. The CoMAND ensemble independently reproduces NMR observables.

Fig. S3. Fold-factors identify well-folded models for hUb.

Fig. S4. The CoMAND ensemble for human ubiquitin.

Movie S1. The SARS folding trajectory of polb4.

References

1. O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, B. L. de Groot, Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471 (2008).

2. A. M. Bonvin, A. T. Brunger, Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? *J Biomol NMR* **7**, 72 (1996).

3. R. Burgi, J. Pitera, W. F. van Gunsteren, Assessing the effect of conformational averaging on the measured values of observables. *J Biomol NMR* **19**, 305 (2001).

4. K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, M. Vendruscolo, Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128 (2005).

5. B. Richter, J. Gsponer, P. Várnai, X. Salvatella, M. Vendruscolo, The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *Journal of Biomolecular NMR* **37**, 117 (2007).

6. A. Brunger, G. Clore, A. Gronenborn, R. Saffrich, M. Nilges, Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science* **261**, 328 (1993).

7. C. A. E. M. Spronk, S. B. Nabuurs, E. Krieger, G. Vriend, G. W. Vuister, Validation of protein structures derived by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **45**, 315 (2004).

8. Y. J. Huang, A. Rosato, G. Singh, G. T. Montelione, RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Research* **40**, W542 (2012).

9. T. Diercks, M. Coles, H. Kessler, An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. *Journal of Biomolecular NMR* **15**, 177 (1999).

10. K. Lindorff-Larsen, P. Maragakis, S. Piana, D. E. Shaw, Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *The Journal of Physical Chemistry B* **120**, 8313 (2016).

11. R. B. Fenwick, S. Esteban-Martín, B. Richter, D. Lee, K. F. A. Walter, D. Milovanovic, S. Becker, N. A. Lakomek, C. Griesinger, X. Salvatella, Weak Long-Range Correlated Motions in a

Surface Patch of Ubiquitin Involved in Molecular Recognition. *Journal of the American Chemical Society* **133**, 10336 (2011).

12. A. S. Maltsev, A. Grishaev, J. Roche, M. Zasloff, A. Bax, Improved Cross Validation of a Static Ubiquitin Structure Derived from High Precision Residual Dipolar Couplings Measured in a Drug-Based Liquid Crystalline Phase. *Journal of the American Chemical Society* **136**, 3752 (2014).

13. Q. Xu, J. R. Sachs, T.-C. Wang, W. H. Schaefer, Quantification and Identification of Components in Solution Mixtures from 1D Proton NMR Spectra Using Singular Value Decomposition. *Analytical Chemistry* **78**, 7175 (2006).

14. Q. Xu, J. R. Sachs, T.-C. Wang, W. H. Schaefer, Quantification and Identification of Components in Solution Mixtures from 1D Proton NMR Spectra Using Singular Value Decomposition. *Analytical Chemistry* **78**, 7175 (2006).

15. A. D. Mackerell, M. Feig, C. L. Brooks, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* **25**, 1400 (2004).

16. R. Pozzi, M. Coles, D. Linke, A. Kulik, M. Nega, W. Wohlleben, E. Stegmann, Distinct mechanisms contribute to immunity in the lantibiotic NAI-107 producer strain Microbispora ATCC PTA-5024. *Environmental Microbiology* **18**, 118 (2016).

17. P. Güntert, Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy* **43**, 105 (2003).

18. Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **44**, 213 (2009).

19. Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, A. Bax, Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences* **105**, 4685 (2008).

20. L. Zhu, H. J. Dyson, P. E. Wright, A NOESY-HSQC simulation program, SPIRIT. *Journal of Biomolecular NMR* **11**, 17 (1998).

21. SPARKY 3 (University of California, San Fransisco).

22. G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *Journal of the American Chemical Society* **120**, 6836 (1998).

23. Y. X. Wang, Y. J. Zhang, Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering* **25**, 1336 (2013).

24. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annual review of biochemistry* **77**, 363 (2008).

25. D. Gront, D. W. Kulp, R. M. Vernon, C. E. Strauss, D. Baker, Generalized fragment picking in Rosetta: design, protocols and applications. *PloS one* **6**, e23294 (2011).

26. J. Huang, A. D. MacKerell, CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **34**, 2135 (2013).

27. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577 (1995).

28. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781 (2005).

29. A. J. Nederveen, A. M. J. J. Bonvin, NMR Relaxation and Internal Dynamics of Ubiquitin from a 0.2 μs MD Simulation. *Journal of Chemical Theory and Computation* **1**, 363 (2005).

30. A. C. Wang, A. Bax, Determination of the backbone dihedral angles phi in human ubiquitin from reparametrized empirical Karplus equations. *Journal of the American Chemical Society* **118**, 2483 (1996).

31. M. Zweckstetter, A. Bax, Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. *Journal of the American Chemical Society* **122**, 3791 (2000).

Figures

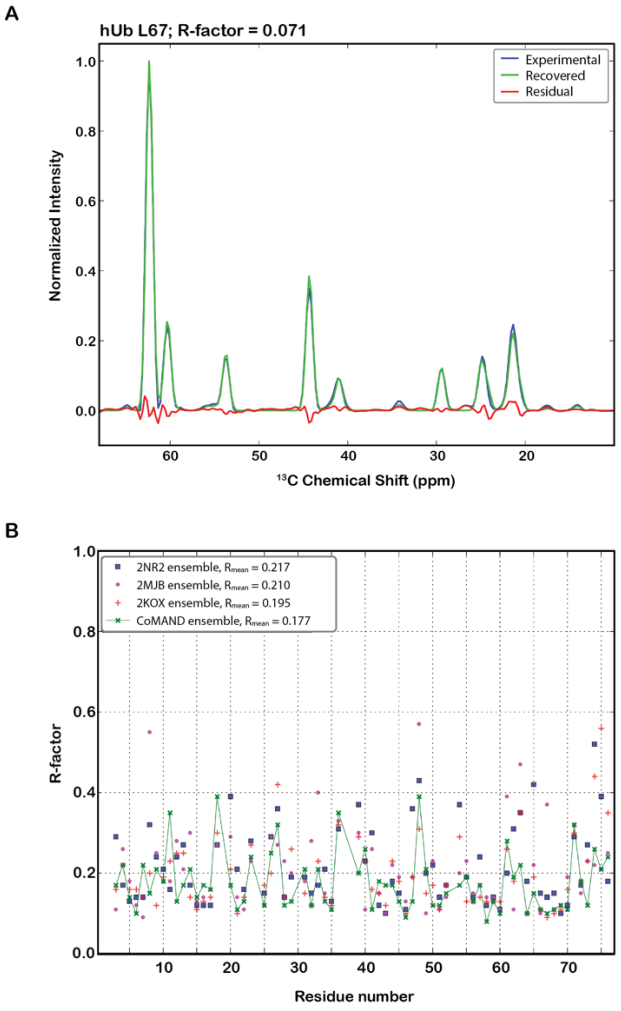


Fig. 1. NOESY back-calculation in the CoMAND method.

(A) An example comparison demonstrating the quality of back-calculation of CNH-NOESY spectra. The experimental spectrum for L67 in human ubiquitin (hUb) is in blue and the recovered spectrum back-calculated by averaging across the 640 models of the 2KOX structure ensemble [11] is in green. The residual signal is in red (R-factor = 0.071). **(B)** R-factors plotted across the sequence for three literature ensembles. These ensembles have been compiled to emphasise different aspects of the hUb structure: 2KOX to elucidate internal motions [11], 2NR2 via a minimal under-restraining, minimal over-restraining procedure [5] and 2MJB to represent a static average pose [12]. The average structures for all three ensembles are very similar and differences in R-factors are therefore attributable to the different representations of conformational diversity (see Supplemental Figure S1 for specific examples). These are compared to the CoMAND ensemble (green line).

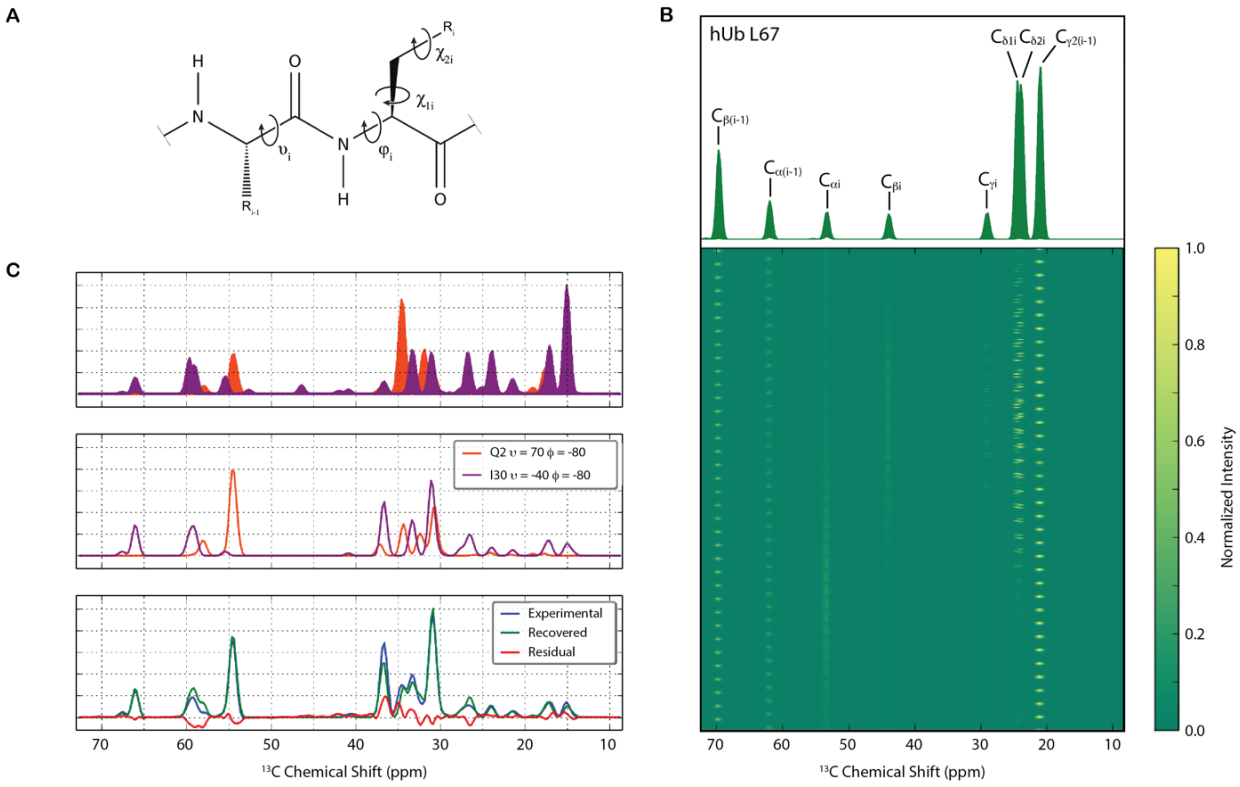


Fig. 2. Conformational mapping in the CoMAND method.

(**A)** The definition of the conformational search space. Note that in the shifted Ramachandran space, υ_i is equivalent to $\psi_{(i-1)}$. **(B)** The features matrix for L67 in hUb shown as a stacked plot where each row is a spectrum back-calculated via systematic conformational sampling, resulting in periodic intensity patterns. The order of sampling, from fastest to slowest, is χ^2, χ^1, υ, ϕ with 10° steps for backbone and 120° steps for sidechain angles. The intensity of each peak in the spectrum displays a different dependency on the dihedrals, underlining the power of the data to discriminate individual conformations. The projection of this plot – i.e. all members of the features set overlaid - is shown above with individual peaks assigned. **(C)** Decomposing overlapped spectra. The top panel shows all members of the concatenated features set for Q2 (orange) and I30 (purple) in hUb. Two-component factorization successfully decomposes the completely overlapped experimental spectra, yielding the correct conformations of the respective residues (middle panel).

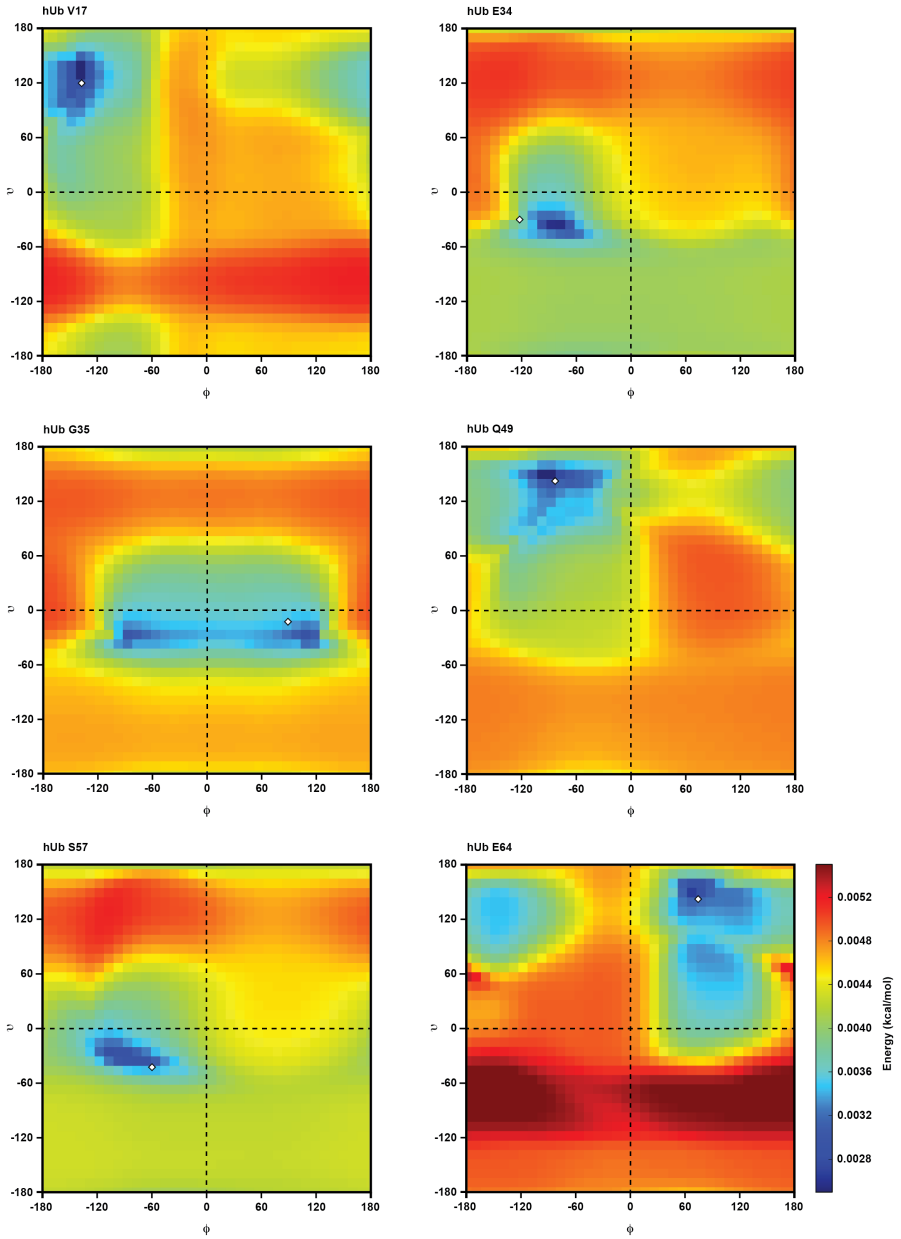


Fig. 3. Conformation maps from spectral decomposition.

Conformational maps are shown for six examples residues representing different secondary structural contexts in hUb. They are expressed as heat maps of conformer free energy change, relative to a two-conformer global minimum reference state. For non-glycine residues, a two-dimensional (υ, ϕ) slice through the full three- or four-dimensional map at the minimum χ^1/χ^2 position is shown. The map for G35 displays typical pseudo-symmetry about the ϕ=0 axis due to the achiral nature of glycine residues. In each map, the minima agree very well with the corresponding crystallographic conformations (1UBQ; white diamonds).

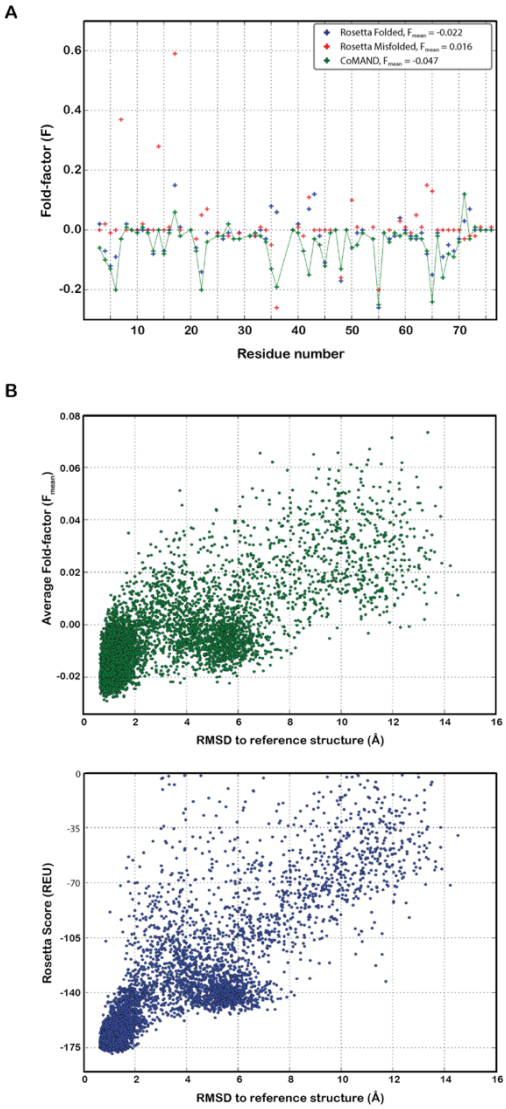


Fig. 4. The R-factor as an objective target function.

(A) Fold factors plotted across the sequence for hUb. Fold factors (F) are calculated on a per-residue basis as the difference between the R-factor calculated for a linear peptide and that calculated for the full structure. This isolates the component of the R-factor not explained by local contacts. Negative values indicate residues in well-folded environments. Values are shown for an initial folded model from Rosetta runs plus a Rosetta structure misfolded by a strand swap in the N-terminal α-hairpin. These are compared to a representative of the final CoMAND ensemble (green line). **(B)** Comparison of the average fold factor (F_{mean}) versus the Rosetta score (Rosetta Energy Units) as selection criteria for well-folded hUb models. Both measures are plotted against the RMSD to the reference structure (1UBQ) for the same set of 7215 Rosetta decoys with sub-zero score. Structures with low F_{mean} are consistently close to the reference structure.

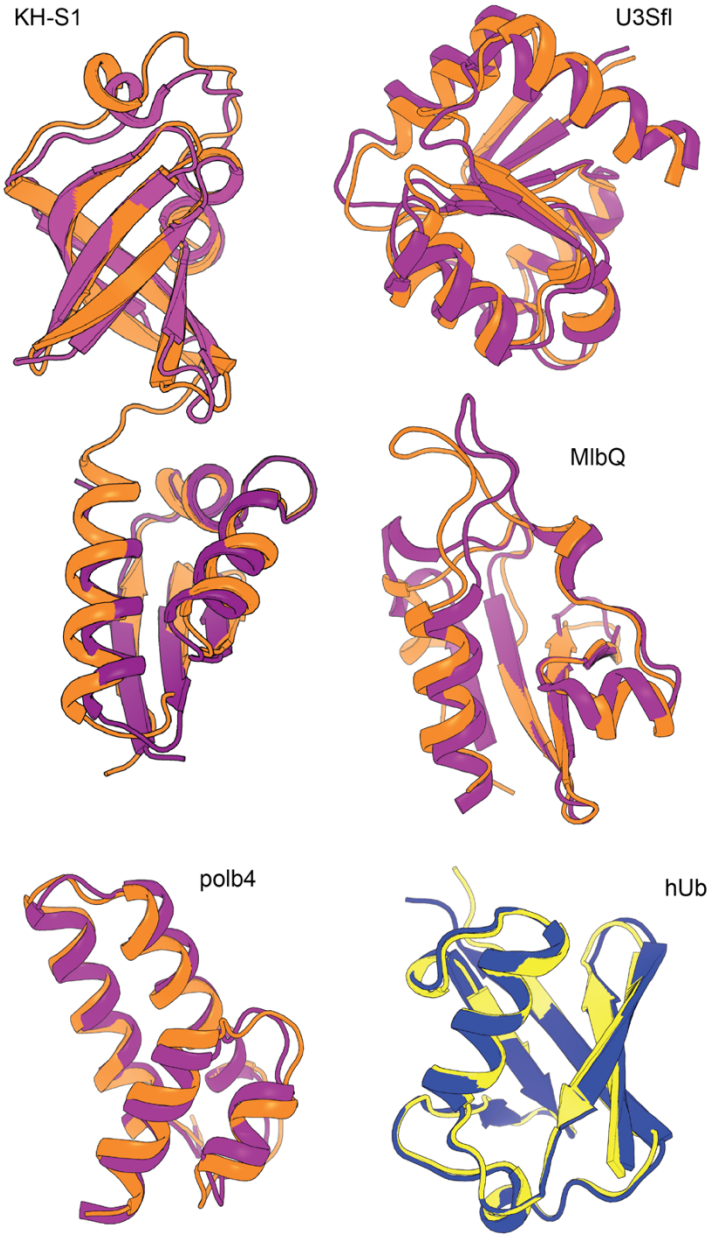


Fig. 5. The CoMAND structure gallery.

(A) The CoMAND structure gallery. Models (orange) are shown superimposed on their respective reference structures. For U3Sfl and MlbQ the reference structures are previously solved solution structures. For KH-S1, the KH domain reference structure is 4AM3 (light purple) and the S1 domain reference structure is 4NNG (dark purple). The *de novo* structure determined for polb4 is compared to the design target. A single model from the refined CoMAND ensemble for hUb is shown in yellow and the reference structure (1UBQ) in blue (RMSD over backbone atoms 0.49 Å).

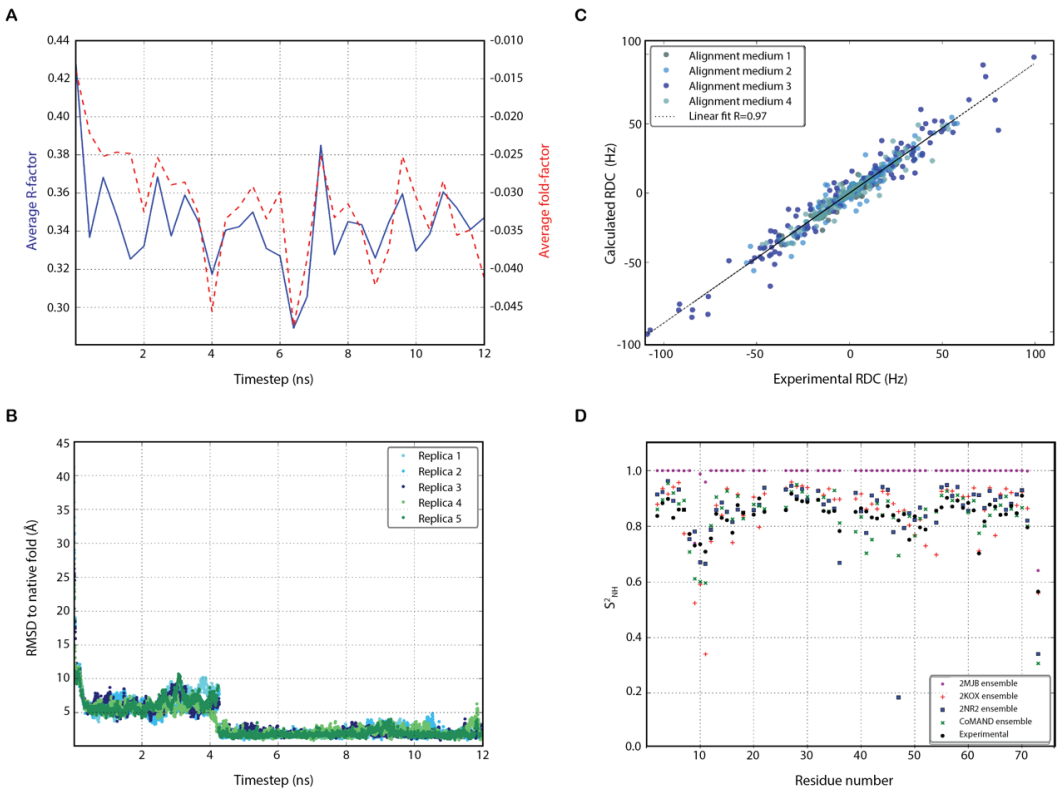
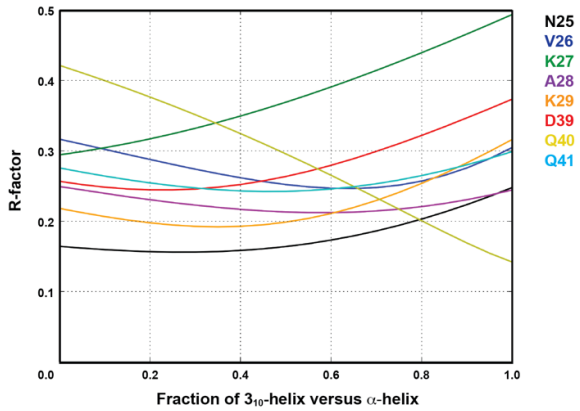


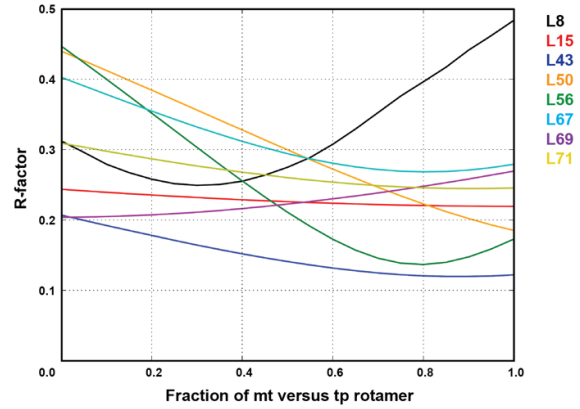
Fig. 6. Validation of CoMAND ensembles.

(A) Evolution of the average sequence R-factor and fold factor along the first replica of the SARS folding trajectory of polb4. The lowest R-factor structure (time point of 6.4 ns) was chosen as a low-resolution model (See also Supplemental Movie S1) **(B)** The RMSD from the native fold along the trajectory of the same SARS folding simulation. Traces for all five replicates are shown, illustrating the convergence of the protocol. **(C)** The CoMAND ensemble independently reproduces NMR observables. The correlation between residual RDC values back-calculated from the CoMAND ensemble and experimental values in four different alignment media is shown [12]. The Q-factor expressing the agreement between prediction and experiment for this data set is 0.24. Similarly good agreement is obtained between back-calculated and experiment $^{3}J_{HNH\alpha}$ coupling constants (correlation coefficient = 0.95; Supplementary Figure S2). RDC values report on the orientation of various bond vectors to an external molecular alignment medium and are thus sensitive to both local and global structure. $^{3}J_{HNH\alpha}$ coupling constants report on local ϕ angles. Neither parameter was used in compiling the CoMAND ensemble. **d)** Calculated S^2_{NH} order parameter values across the sequence of hUb using the first 20 models of the 2MJB, 2KOX, 2NR2 and CoMAND ensembles. The CoMAND ensemble best reproduces experimental values derived from NMR relaxation analysis [29] (correlation coefficient=0.82). Correlations for the reference ensembles are shown in Supplementary Figure S2).

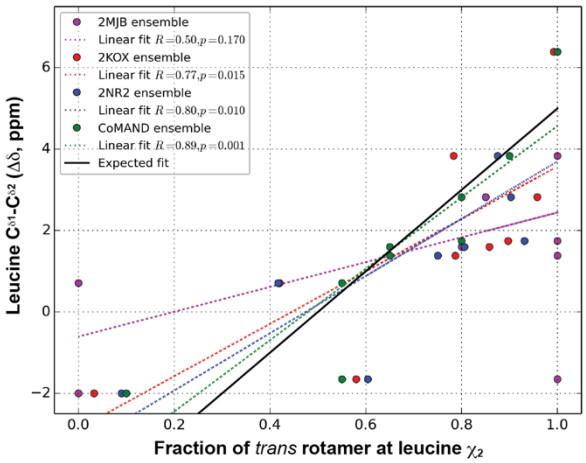
Supplementary Materials

A





B



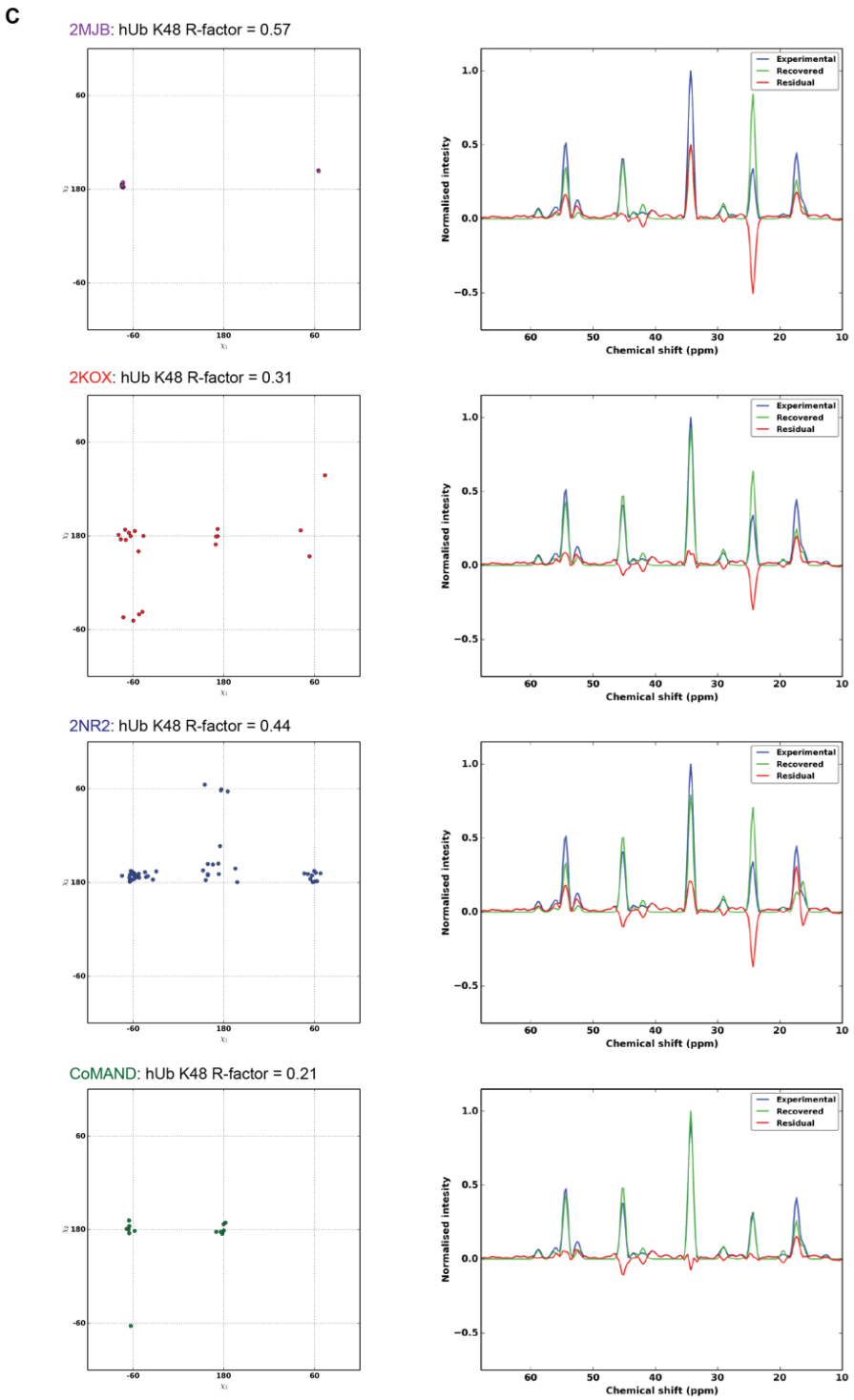


Figure S1: CNH-NOESY based R-factors are highly sensitive to local conformation.

(A) Plots are shown of R-factors versus composition for idealised two-state conformational mixtures. The upper plot shows linear mixtures between α-helical (υ = -40°, ϕ = -62°) and 3_{10}-helical (υ = -10°, ϕ = -95°) conformations, modelling a conformational transition by helical

unwinding. Traces are shown for selected residues in helical regions of human Ubiquitin (hUb). Some residues are best explained by pure conformers, e.g. K27 and Q40, while others have R-factor minima for mixtures. The lower panel models transition between the two most populated sidechain rotamers for leucine residues: mt (χ^1 = -60°, χ^2 = 180°) and tp (χ^1 = 180°, χ^2 = 60°). Traces are shown for all leucine residues in hUb for which data is available. Residues sampling multiple conformations are clearly identified. **(B)** Validating conformational mixtures in the CoMAND ensemble. The chemical shifts of leucine $C^{\delta 1}$ and $C^{\delta 2}$ carbons are sensitive to the χ^2 rotamer due to a "γ-gauche effect" [32]. The difference in these shifts correlates with the proportion of *trans* rotamer in leucine residues and thus provides an independent estimate of the conformational mixtures described in the lower plot of panel A. The plots show the shift differences ($\Delta\delta$) versus the proportion of *trans* rotamer for the CoMAND and three reference ensembles for hUb. These ensembles have been compiled according to different metrics: 2KOX to elucidate internal motions, 2NR2 according to minimum under-restraining, minimum over-restraining criteria and 2MJB, which represents a static structure close to the average structure and is therefore not expected to explain conformational diversity well. The CoMAND ensemble best explains the observed chemical shifts. The expected shift difference (solid line) is based on the equation derived by Mulder [32]. **(C)**. Literature ensembles for hUb over- and under-estimate conformational diversity. The panels on the left show the distribution of χ^1/χ^2 rotamers for K48 in hUb in the CoMAND and reference ensembles. For 2KOX only the first 20 models of the ensemble are shown for clarity. The panels on the right show the comparison between experimental spectra and spectra back-calculated over the whole ensemble. The R-factors for these comparisons demonstrate its sensitivity to accurate representation of conformational diversity.

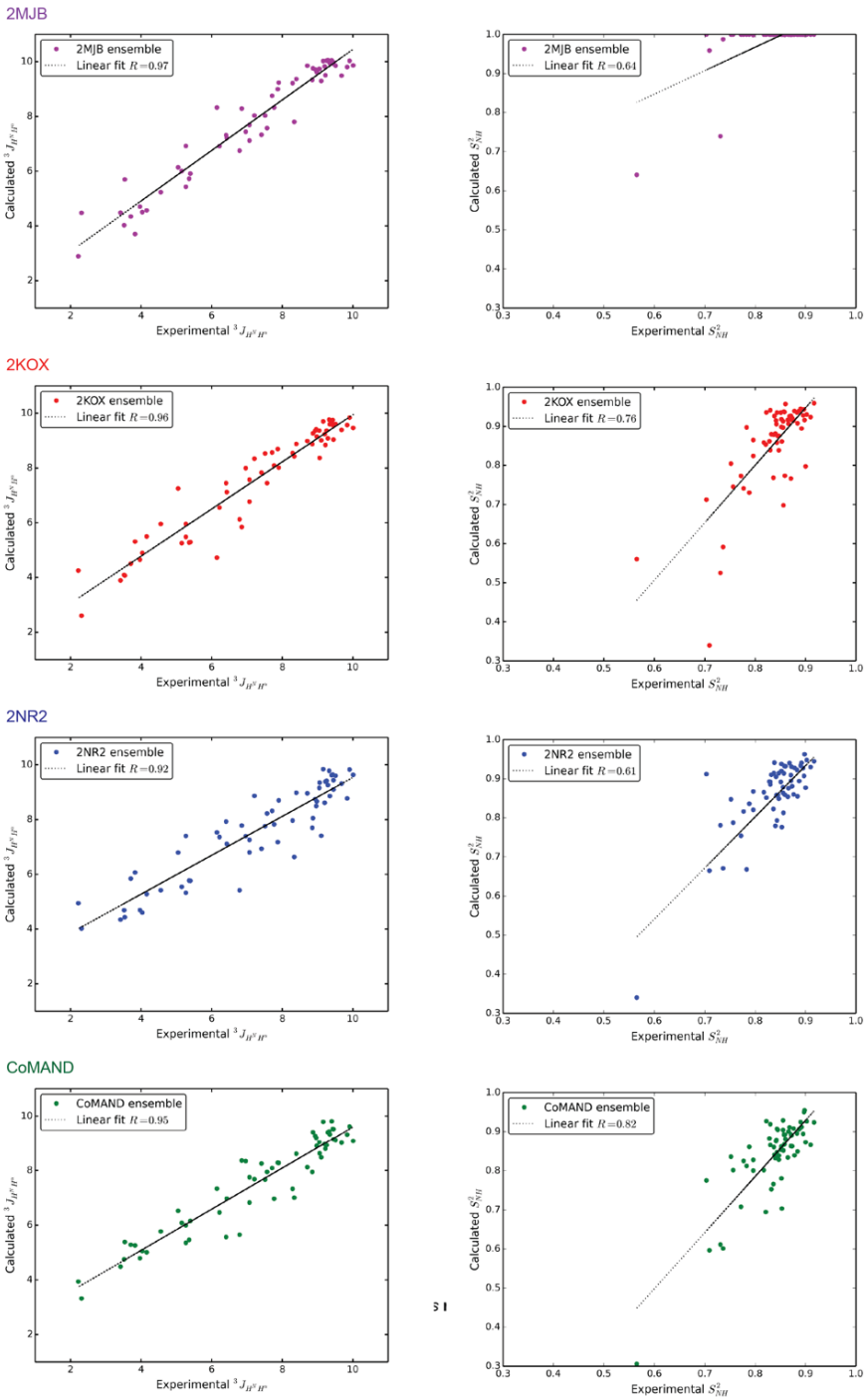


Figure S2: The CoMAND ensemble independently reproduces NMR observables.

Correlations are shown between experimental and back-calculated $^3J_{HNH\alpha}$ coupling constants [12] and backbone S^2_{NH} order parameters [29] for the CoMAND and reference ensembles. Note that the 2MJB ensemble was refined against $^3J_{HNH\alpha}$ couplings, while 2NR2 was refined against order parameters.

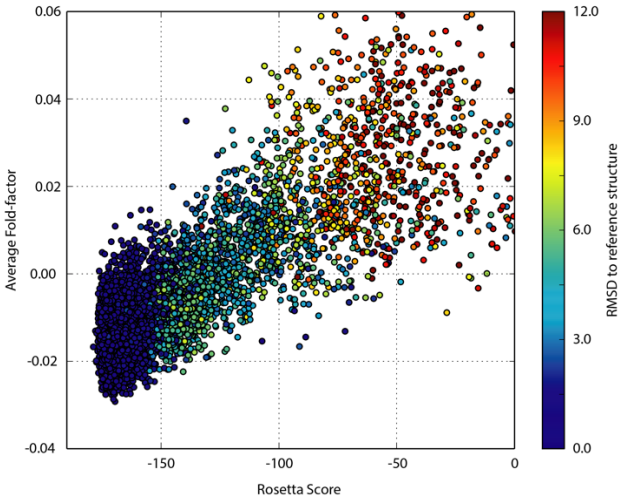


Figure S3: Fold-factors identify well-folded models for hUb.

Correlations are shown between the average fold-factor (F_{mean}) and Rosetta Score (Rosetta Energy Units) for a set of 7215 decoys with sub-zero score calculated for hUb. Each point is coloured according to the RMSD to the reference crystal structure (1UBQ). Agreement between the two measures is a very good predictor of well-folded decoys.

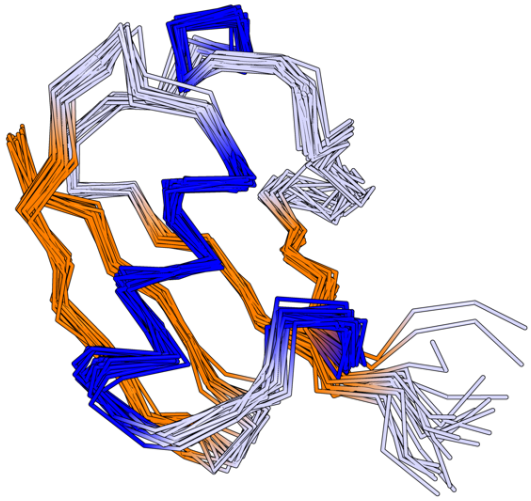


Figure S4: The CoMAND ensemble for human ubiquitin.

The refined ensemble for hUb (20 models) is shown superimposed over backbone atoms. Helices are in blue and β-strands in orange. The backbone RMSD to the average structure for the ensemble is 0.64 Å. The ensemble has been compiled by frame-picking structures from an unrestrained molecular dynamics simulation employing a greedy algorithm to minimise the overall average R-factor.

Movie S1

The SARS folding trajectory of polb4. The movie shows the time evolution of the communicating replica performing a *seilschaft* search for lower energy minima. The inset shows the backbone RMSD from the design as a function of time.

Conclusions

Success of protein design can be defined by atomic-level agreement between the design model and determined structure. Although differing from case to case, success rates have been consistently low. This can be attributed to deficient sampling due to the intractable dimensionality of the search space and inaccurate scoring due to the simplistic bases of the scoring routines. Through this research, I aimed to localise the search spaces, which both allows exhaustive sampling to be carried out and more rigorous scoring schemes to evaluate the sampled states. On the sampling side, I show that the step-wise interface-driven approach utilising fragments drastically reduces the sampling required to reach a good solution, while maintaining topological control. On the scoring side, the work-based PMF and perturb-probe schemes had been tested for convergence against benchmarks before they were deployed to filter the design candidates for experimental testing. Combined, these advanced have allowed me to experimentally test a small number of designs and were reflected in very high success rates. Likewise, performing localised systematic sampling of the conformational space combined with analytical decomposition of the experimental NMR spectra could yield

local free energy distribution maps for individual residues. This systematic factorisation not only provided a means for building initial models *de novo* using a purpose-built accelerated molecular dynamics routine (SARS), but also enables highly detailed refinement of the final ensemble. In contrast to present NMR methods, the quantitative information extraction from spectra allows for resolving overlapped spectra, explaining split-peaks, and accounting for multi-stable conformational distributions.

The ultimate goal of protein design is creating novel proteins with bespoke functions. This is a challenging goal given the extra level of complexity relative to the design of structure alone. To this end, designing structures at atomic precision and high success rates, combined with accurate description of protein conformational dynamics in solution, offers decisive advantages. Learning from that, I am currently applying such approaches to design functional proteins, where the high success rate is already feeding into the discovery of highly active molecules.

References

[1] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[2] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[3] F. Weinhold and R. A. Klein, "What is a hydrogen bond? mutually consistent theoretical and experimental criteria for characterizing h-bonding interactions," *Molecular Physics*, vol. 110, no. 9-10, pp. 565–579, 2012.

[4] Levinthal, Cyrus, "Are there pathways for protein folding?," *J. Chim. Phys.*, vol. 65, pp. 44–45, 1968.

[5] P. S. Kim and R. L. Baldwin, "Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding," *Annual Review of Biochemistry*, vol. 51, no. 1, pp. 459–489, 1982.

[6] A. L. Serrano, M. M. Waegele, and F. Gai, "Spectroscopic studies of protein folding: Linear and nonlinear methods," *Protein Science*, vol. 21, no. 2, pp. 157–170.

[7] A. J. Wirth, Y. Liu, M. B. Prigozhin, K. Schulten, and M. Gruebele, "Comparing fast pressure jump and temperature jump protein folding experiments and simulations," *Journal of the American Chemical Society*, vol. 137, no. 22, pp. 7152–7159, 2015.

[8] S. W. Englander, L. Mayne, Z.-Y. Kan, and W. Hu, "Protein folding—how and why: By hydrogen exchange, fragment separation, and mass spectrometry," *Annual Review of Biophysics*, vol. 45, no. 1, pp. 135–152, 2016.

[9] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[10] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: challenges in the simulation of protein folding," *Current Opinion in Structural Biology*, vol. 23, no. 1, pp. 58 – 65, 2013.

[11] H. Meirovitch, S. Cheluvaraja, and R. P. White, "Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding," *Current Protein Peptide Science*, vol. 10, no. 3, pp. 229–243, 2009.

[12] M. C. Thielges and M. D. Fayer, "Protein dynamics studied with ultrafast two-dimensional infrared vibrational echo spectroscopy," *Accounts of Chemical Research*, vol. 45, no. 11, pp. 1866–1874, 2012.

[13] C. Baiz, Y.-S. Lin, C. Peng, K. Beauchamp, V. Voelz, V. Pande, and A. Tokmakoff, "A molecular interpretation of 2d ir protein folding experiments with markov state models," *Biophysical Journal*, vol. 106, no. 6, pp. 1359 – 1370, 2014.

[14] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia, "Scop: a structural classification of proteins database," *Nucleic Acids Research*, vol. 25, no. 1, pp. 236–239, 1997.

[15] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "Cath – a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093 – 1109, 1997.

[16] M. Sadowski and W. Taylor, "On the evolutionary origins of "fold space continuity": A study of topological convergence and divergence in mixed alpha-beta domains," *Journal of Structural Biology*, vol. 172, no. 3, pp. 244 – 252, 2010.

[17] R. Kolodny, L. Pereyaslavets, A. O. Samson, and M. Levitt, "On the universe of protein folds," *Annual Review of Biophysics*, vol. 42, no. 1, pp. 559–582, 2013.

[18] W. R. Taylor, "Evolutionary transitions in protein fold space," *Current Opinion in Structural Biology*, vol. 17, no. 3, pp. 354 – 361, 2007.

[19] M. I. Sadowski and W. R. Taylor, "Protein structures, folds and fold spaces," *Journal of Physics: Condensed Matter*, vol. 22, no. 3, p. 033103, 2010.

[20] W. R. Taylor, V. Chelliah, S. M. Hollup, J. T. MacDonald, and I. Jonassen, "Probing the "dark matter" of protein fold space," *Structure*, vol. 17, no. 9, pp. 1244 – 1252, 2009.

[21] D. N. Woolfson, G. J. Bartlett, A. J. Burton, J. W. Heal, A. Niitsu, A. R. Thomson, and C. W. Wood, "De novo protein design: how do we expand into the universe of possible protein structures?," *Current Opinion in Structural Biology*, vol. 33, pp. 16 – 26, 2015.

[22] Z. Li and H. A. Scheraga, "Monte carlo-minimization approach to the multiple-minima problem in protein folding," *Proceedings of the National Academy of Sciences*, vol. 84, no. 19, pp. 6611–6615, 1987.

[23] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1, pp. 141 – 151, 1999.

[24] J. Kästner, "Umbrella sampling," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 6, pp. 932–942.

[25] C. Zhang and J. Ma, "Enhanced sampling and applications in protein folding in explicit solvent," *The Journal of Chemical Physics*, vol. 132, no. 24, p. 244101, 2010.

[26] A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, and B. Kuhlman, "Chapter six - scientific benchmarks for guiding macromolecular energy function improvement," in *Methods in Protein Design* (A. E. Keating, ed.), vol. 523 of *Methods in Enzymology*, pp. 109 – 143, Academic Press, 2013.

[27] D. A. Kofke, "Free energy methods in molecular simulation," *Fluid Phase Equilibria*, vol. 228, pp. 41–48, 2005.

[28] J. Huang and A. MacKerell, "Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data," *Journal of Computational Chemistry*, vol. 34, no. 25, pp. 2135–2145.

[29] G. Grigoryan, "Absolute free energies of biomolecules from unperturbed ensembles," *Journal of Computational Chemistry*, vol. 34, no. 31, pp. 2726–2741.

[30] D. E. Koshland, "Application of a theory of enzyme specificity to protein synthesis," *Proceedings of the National Academy of Sciences*, vol. 44, no. 2, pp. 98–104, 1958.

[31] J. Monod, J. Wyman, and J.-P. Changeux, "On the nature of allosteric transitions: A plausible model," *Journal of Molecular Biology*, vol. 12, no. 1, pp. 88 – 118, 1965.

[32] D. E. Koshland, G. Némethy, and D. Filmer, "Comparison of experimental binding data and theoretical models in proteins containing subunits*," *Biochemistry*, vol. 5, no. 1, pp. 365–385, 1966.

[33] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, p. 964, Dec 2007.

[34] D. M. LeMaster, "Nmr relaxation order parameter analysis of the dynamics of protein side chains," *Journal of the American Chemical Society*, vol. 121, no. 8, pp. 1726–1742, 1999.

[35] A. G. Palmer, "Nmr characterization of the dynamics of biomacromolecules," *Chemical Reviews*, vol. 104, no. 8, pp. 3623–3640, 2004.

[36] I. R. Kleckner and M. P. Foster, "An introduction to nmr-based approaches for measuring protein dynamics," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1814, no. 8, pp. 942–968, 2011.

[37] G. A. Khoury, J. Smadbeck, C. A. Kieslich, and C. A. Floudas, "Protein folding and de novo protein design for biotechnological applications," *Trends in Biotechnology*, vol. 32, no. 2, pp. 99 – 109, 2014.

[38] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T.

Montelione, D. Baker, and A. Bax, "Consistent blind protein structure generation from nmr chemical shift data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4685–4690, 2008.

[39] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, H. L. Axelrod, D. Das, S. M. Vorobiev, H. Iwaï, P. R. Pokkuluri, and D. Baker, "Improved molecular replacement by density- and energy-guided protein structure optimization," *Nature*, vol. 473, p. 540, May 2011.

[40] R. Y.-R. Wang, M. Kudryashev, X. Li, E. H. Egelman, M. Basler, Y. Cheng, D. Baker, and F. DiMaio, "De novo protein structure determination from near-atomic-resolution cryo-em maps," *Nature Methods*, vol. 12, p. 335, Feb 2015.

[41] J. Söding and A. N. Lupas, "More than the sum of their parts: On the evolution of proteins from peptides," *BioEssays*, vol. 25, no. 9, pp. 837–846.

[42] D. U. Ferreiro and P. G. Wolynes, "The capillarity picture and the kinetics of one-dimensional protein folding," *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 9853–9854, 2008.

[43] F. Parmeggiani and P.-S. Huang, "Designing repeat proteins: a modular approach to protein design," *Current Opinion in Structural Biology*, vol. 45, pp. 116 – 123, 2017.

[44] F. Parmeggiani, P.-S. Huang, S. Vorobiev, R. Xiao, K. Park, S. Caprari, M. Su, J. Seetharaman, L. Mao, H. Janjua, G. T. Montelione, J. Hunt, and D. Baker, "A general computational approach for repeat protein design," *Journal of Molecular Biology*, vol. 427, no. 2, pp. 563 – 575, 2015.

[45] T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie, and L. Regan, "A new folding paradigm for repeat proteins," *Journal of the American Chemical Society*, vol. 127, no. 29, pp. 10188–10190, 2005.

[46] B. T. Porebski and A. M. Buckle, "Consensus protein design," *Protein Engineering, Design and Selection*, vol. 29, no. 7, pp. 245–251, 2016.

[47] V. Alva, J. Söding, and A. N. Lupas, "A vocabulary of ancient peptides at the origin of folded proteins," *eLife*, vol. 4, p. e09410, dec 2015.

[48] T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski, and B. Kuhlman, "Design of structurally distinct proteins using strategies inspired by evolution," *Science*, vol. 352, no. 6286, pp. 687–690, 2016.

[49] C. F. Wright, S. A. Teichmann, J. Clarke, and C. M. Dobson, "The importance of sequence diversity in the aggregation and evolution of proteins," *Nature*, vol. 438, pp. 878–881, Dec 2005.

[50] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, pp. 1364–1368, 2003.

[51] D. E. Kim, B. Blum, P. Bradley, and D. Baker, "Sampling bottlenecks in de novo protein structure prediction," *Journal of Molecular Biology*, vol. 393, no. 1, pp. 249 – 260, 2009.

[52] D. Ringe and G. A. Petsko, "The 'glass transition' in protein dynamics: what it is, why it occurs, and how to exploit it," *Biophysical Chemistry*, vol. 105, no. 2, pp. 667 – 680, 2003.

[53] B. Zagrovic and V. S. Pande, "How does averaging affect protein structure comparison on the ensemble level?," *Biophysical Journal*, vol. 87, no. 4, pp. 2240 – 2246, 2004.

[54] B. Richter, J. Gsponer, P. Várnai, X. Salvatella, and M. Vendruscolo, "The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins," *Journal of Biomolecular NMR*, vol. 37, pp. 117–135, Feb 2007.

[55] L. Zhu, H. J. Dyson, and P. E. Wright, "A noesy-hsqc simulation program, spirit," *Journal of Biomolecular NMR*, vol. 11, pp. 17–29, Jan 1998.

List of abbreviations:

NMR	Nuclear Magnetic Resonance
FTIR	Fourier-transform Infrared Spectroscopy
CD	Circular Dichroism
MD	Molecular Dynamics
MSM	Markov State Model
GPU	Graphics Processing Unit
k_B	Boltzmann constant
Z_β	Partition function
K_U	Unfolding equilibrium constant
R	Gas constant
k_f	Folding rate constant
n_{ss}	Number of secondary structural elements
T	Temperature in Kelvins
SCOP	Structural Classification Of Proteins database
CATH	Class-Architecture-Topology-Homology classification database
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
∇f	Gradient of function f with respect to 3D Cartesian position
δt	Simulation time step
v	Velocity vector
r	Position vector

k_{term} Force constant of the respective *term*

ϵ_o Permittivity of medium

q Atomic charge

ϕ Backbone phi dihedral angle

ψ Backbone psi dihedral angle

VALOCIDY Valuation of Local Configuration Integral with Dynamics

FT-NMR Fourier-transform Nuclear Magnetic Resonance

γ Nuclear gyromagnetic ratio

B_o Magnetic field strength in Hz

λ linewidth; full peak width at half maximum height

T_1 Longitudinal relaxation time constant

T_2 Transverse relaxation time constant

S^2 Generalised order parameter

CPMG Carr-Purcell Meiboom-Gill relaxation dispersion pulse sequence

HhH helix-hairpin-helix

TPR Tetratricopeptide repeat

dRP deoxyribophosphate

PMF Potential of Mean Force

SMD Steered Molecular Dynamics

RFD Rotational Force Dissipation

PDB Protein Data Bank

NOESY Nuclear Overhauser Effect Spectroscopy

CoMAND Conformational Mapping via Analytical NOESY Decomposition

CMAP Corrective backbone cross-term

SARS Simulated Annealing Replica Seilschaft

J Scalar coupling constant

RDC Residual Dipolar Coupling constant