NORTHWESTERN UNIVERSITY

Dissecting the stability determinants of a challenging de novo protein fold and assay development to identify cell-penetrating miniproteins

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Driskill Graduate Program in Life Sciences

By
Tae-Eun Kim

Evanston, IL

September 2023

TABLE OF CONTENTS

ABSTRACT

This dissertation focuses on quantifying protein folding stability determinants and presenting initial experiments that can guide the development of a novel assay that identifies cell-penetrating miniproteins.

First, despite over a century of scholarship on protein folding stability, applying this knowledge to design proteins computationally remains limited. Usually, protein designers generate many protein structures, ranging from dozens to thousands, but only a fraction of them will successfully express in *E. coli* and remain soluble in solution. This suggests that we still lack a fuller understanding of the determinants of protein folding stability and incorporating this knowledge into the protein design process. Addressing this challenge could increase the success rate in designing stable proteins for various therapeutic and biomedical applications, such as creating new binders and biosensors.

To better understand the determinants of protein folding stability, I used a miniprotein (αββα topology, 43-residues) that was previously difficult to design as a model system. By combining computational protein design, high-throughput experimentation, and machine learning, I designed stable αββα miniproteins with greater success than in previous work. Then, I quantified how individual biophysical forces uniquely contribute to folding stability and propose a “recipe” for designing future iterations of stable αββα miniproteins.

The second focus of this dissertation is to provide preliminary work that can guide the creation of a novel high-throughput screen for cell-penetrating miniproteins.

Many protein-protein interactions that are implicated in disease occur in the cell cytosol, but many small molecule drugs (currently the most common class of pharmaceutical drug) are not always effective. This is because small molecules require a deep binding pocket in a protein to bind, but this is not a characteristic feature of protein-protein interfaces.

An alternative to small molecule drugs is a protein-based therapeutic, but proteins do not readily cross the cell plasma membrane given the hydrophilic surfaces of proteins and the nonpolar lipid bilayer of the cell plasma membrane. Only a few proteins with cell-penetrating capabilities have been characterized (e.g. histone proteins, PTEN, zinc-finger 5.3), but this nonetheless lends credence to the hypothesis that there exists determinants for cell-penetration. Thus, a high-throughput screen could help identify cell-penetrating proteins from which we could discover general design rules for cell-penetration.

Here, I show that combining intein splicing, synthetic transcription factors, and reporter gene activation in mammalian cells can signal the presence of a protein-of-interest in the cell cytosol. I propose that this system can serve as the foundation for the design of a “reporter” cell, which expresses a fluorescent protein only if a cell-penetrating protein has entered the cell cytosol. This “reporter” cell, in conjunction with a “secretor” cell, can serve as two core components in a high-throughput droplet microfluidic screen for detecting novel cell-penetrating miniproteins. I will describe in the final chapter possible directions for making this assay a reality.

ACKNOWLEDGEMENTS

This is the most important section in this dissertation. So many people have supported my path to and through graduate school. Every person, whether they knew it or not, contributed to my success through their words and actions and in their own unique way. They have taught me not only how to be a better scientist, but to be a more attentive, kind, patient, thankful, curious, and analytical person. When I look at all of these names, I realize that they came into my life at key moments, and it seems as if the pieces fell in place at the right time. I am grateful for this great cloud of witnesses, and I hope to emulate them and pay it forward. If I have missed anyone, that is my error.

My advisor, Gabe Rocklin, was willing to take me as his first PhD student, and I am grateful for this opportunity. I am thankful for his mentorship, his patience when my experiments would not work or when my code still needed debugging, his willingness to allow me to test new ideas, and that he introduced me to Nikolai Medtner. I am happy to have spent lab lunches, games, and picnics at Olive Park with my labmates, Suggie Dixit, Kotaro Tsuboyama, Jane Thibeault, Állan Ferrari, Wes Ludwig, Cydney Martell, Claire Phoumyvong, Will Corcoran, Tanu Priya, and Yulia Gutierrez.

The bulk of my PhD research was made possible by collaborations inside and outside of Northwestern. Jimmy Casey at Northwestern University's Keck Biophysics Facility and Sam Halabiya at the University of Washington's BioFab performed crucial experiments. Hugh Haddox from the University of Washington created a computational pipeline that computes metrics for the protein models used in my research, and Scott

Houliston and Alex Lemak from the University of Toronto worked on solving the structure for two protein designs.

My thesis committee members, Neil Kelleher, Josh Leonard, and Neha Kamat helped me fine-tune research ideas and be precise with my experimental design. Our committee meetings helped me be more thorough with my research and improve my presentation skills. I am thankful for Neil’s advice of “bottom line up front,” for Josh tracking me down at the CSB retreat because he had some experimental ideas that I could test, and for Neha reminding me about my control experiments.

Successful research depends on a lot of administrative support, and I am thankful for Pam Carpentier, Toni Gutierrez, and Grace Musante who make the DGP PhD program run as smoothly as it can be. Lexi Smith, Nate Will, and Melissa Daley from the Pharmacology Department were helpful in organizing the departmental student-led symposium and retreats, registering for conferences, and lunches for journal clubs. Steve Anderson gave sound advice and often checked in on how I was doing during my rotations. And Cathy Prullage from IBiS encouraged and helped me navigate my first year in graduate school.

Meeting Joe Muldoon and Patrick Donahue during my first week in graduate school at the IBiS retreat was a serendipitous moment as it was my first introduction to synthetic biology. They mentored me as I rotated in Josh Leonard’s lab and taught me how to think about biology from an engineering perspective (and man, the daily morning trek to the IBiS office for diluted coffee!). I am grateful for their help, especially when they spent late evenings with me in the lab. They taught me the importance of being available to help others. Moreover, the rotation in the Leonard lab led me to enroll in

ChBE 478, which made a lasting impact on how I think about the kinds of research I want to do in the future.

I had the opportunity to contextualize and broaden my view of science outside the lab, and I am thankful for Julie Gertz who organized the Management for Scientists and Engineers Program, and the faculty at Kellogg School of Management for opening the world of management and business strategy to me. Kate Flom Derrick, Katie Pierson, and Grace Bellinger in The Teaching Certificate Program at Northwestern's Searle Center taught me best teaching practices, and Julius Lucks was my teaching mentor to whom I am thankful as he shared his experiences and insights on teaching. Dave Van Winkle allowed me to teach at Missio Dei Uptown where I could test the teaching practices that I had learned, and I am thankful for the opportunity. Kiki Zissimopoulos organized the Research Communication Training Program where I had the opportunity to learn from Byron Stewart, Beth Bennett, and Barbara Shwom. Members of the Rosetta Commons Justice, Equity, Diversity, and Inclusion Committee, especially Parisa Hosseinzadeh and Una Nattermann, taught me the importance of establishing scientific environments where everyone can experience a sense of safety and belonging. I am a more well-rounded scientist and person because of all of them.

The reason why I wanted to go to graduate school was largely because of my experience at the Ann Romney Center for Neurologic Diseases at Brigham and Women's Hospital. I am thankful for getting to know Dennis Selkoe and learning from him during our lab meetings and seminars. Herr Professor Ulf Dettmer taught me to always be critical and that the devil is in the details. If I stay in academia, I want to emulate a lot of the ways he runs his lab. It was an honor to work with Baron von Glehn, and I am thankful to have done research with Thibaut Marguerite, Nagou Legend, Pusheen Kim, Stormy Rovere, Gloria von Stirtz, Svenja Terry-Kantor, and Camilla Hoesch.

My first introduction to hands-on research was at the Ragon Institute of Mass General, MIT and Harvard. Stephanie Jost gave me a chance and let me join her lab. Being mentored by Wilfredo Gracia Beltran was a pivotal moment, and I am so thankful that I learned from him. The way he taught me how to do research, his patience, and his positive energy helped create a welcoming, comfortable, and fun learning environment. I decided I wanted to do research because of him.

When I took classes at Harvard Extension School to gain additional science class credits, Dr. Titan and Christina Fan financially supported me and sent care packages. I would not have been able to apply to graduate school had it not been for their generosity. I still have the gifts from their care packages and I think of them often. During this time at 20 Banks St., having great and supportive housemates in Kenny Fan, Lue Qin, John Yang, and Annie Wang made the transition between college and graduate school meaningful.

I got excited about science, molecular and protein biology in particular, because of Ms. Roberts' AP Biology class at Whitefish Bay High School and her leadership with Ms. Weiss in advising the SMART team. Through this program, I had my first experience learning about protein modeling and what research is like beyond the classroom. I am thankful for Shannon Colton for running the SMART team program at the Milwaukee School of Engineering and learning from Jack Gorski at the Blood Center of Wisconsin.

My parents created a home that highly valued education and learning, and I am grateful for that space. Reading books and watching shows with my brother and sister about science, nature, and wildlife ignited my fascination for biology and the natural world. I cannot think of anyone who did that more than Chris and Martin Kratt in *Kratts' Kreatures.* In many ways, that was where it all started. I also had teachers like Ms. Cook, Mrs. McCormick, and Mrs. Bartell who fostered engaging learning environments and encouraged me to be curious. It is their kindness that I remember the most, and if I become a teacher, I have their examples to follow.

Finally, my wife and daughter have been loving, supportive, flexible, and patient through and through. John Choi showed me the importance of mental health, and long conversations with friends were great counter-balances to my life in the lab: Tyler Parker, Xavier du Maine, Yvon Woappi, Henry Li, Jonathan Holmes, Aaron and Jihye Gyde, Christian and Isabel Espinoza-Schatz, and Nate and Molly Otey. Nick Nowalk, Don Weiss, Dave Van Winkle, and the various faith communities (Missio Dei Uptown, Trinity Church, Holy Trinity Church, Chicago Sinai, and Immanuel Anglican) all provided a space and sounding board as I wrestled with larger questions on life, purpose, faith, and science. They were regular reminders and encouragers to love God and all people, raise my ebenezer, that all other ground is sinking sand, and as Neander wrote: "Lobe den Herren, der dich auf Adelers Fittichen sicher geführet, der dich erhält." Soli deo gloria.

LIST OF FIGURES AND TABLES

CHAPTER 2

CHAPTER 3

APPENDIX 1

APPENDIX 2

CHAPTER 1: Introduction

1.1 Motivation

The study of identifying and measuring the biophysical determinants of protein folding stability has a long history. Almost a century ago, simple protein denaturation and digestion studies on individual proteins, such as albumin and hemoglobin, led to identifying the importance of nonpolar residues and hydrogen bonding in holding proteins in their native conformations.[1,2] Since then, our understanding of these two biophysical forces have been refined,[3–8] and additional determinants for protein folding stability have been studied, such as helix capping, surface charges, loops, salt-bridges, and electrostatics.[9–14]

With this knowledge, many groups over the past few decades have sought to design proteins from scratch (*de novo)* to evaluate our understanding of protein stability and build novel proteins for various biomedical and biotechnological applications.[15–17] Both endeavors are synergistic – a more accurate understanding of the determinants of protein folding stability informs how to better design novel protein structures with user-defined functions; and building new proteins can reveal gaps in our understanding, especially when they fail to fold and function.

Despite advances in protein design efforts, there are two challenges that are the focus of this dissertation. First, many protein designs fail to fold as designed. They are often insoluble or form oligomers when expressed *in vitro*,[15] which suggests that we do not fully understand the determinants of protein folding stability or are unable to incorporate this knowledge into the design process. Overcoming this limitation could enable the design of stable protein therapeutics and protein-based medical diagnostics. Second, there is very little published work [18,19] on designing *de novo* proteins that can cross cellular membranes because current experimental methods have limitations. This is a promising area of research because identifying general biophysical principles ("design rules") for cell-penetration could enable us to design a new class of cell-penetrating protein drugs.

This chapter first provides historical and modern contexts to the challenges that protein designers have faced in generating computational protein models that fold as designed. I introduce how predictive modeling, empirical re-weighting of an energy function used to evaluate computational protein models, and a high-throughput protease stability assay can be used to better understand the determinants of protein folding stability. Ultimately, the goal is to improve the success rate in making stable protein designs.

The second portion of this chapter discusses the promise of cell-penetrating miniproteins (CPMPs) as a potential new class of therapeutic drug as well as current limitations in assaying for cell-penetration. Before we can begin designing cell-penetrating miniproteins, however, we need an experimental assay that can screen for such proteins. Here, I propose how a novel approach of combining split-inteins with a synthetic transcription factor could be used to identify CPMPs.

1.2 Analyzing protein folding stability determinants using protein design

The challenge of designing stable proteins using Rosetta

Some of the earliest protein designs were made in the 1980s. They were simple folds like α-helical coiled coils and β-sheets whose amino acid sequences were manually selected and chemically synthesized.[20,21] Today, a wide range protein designs (e.g. binders,[22–24] biosensors,[25,26] bioswitches,[27,28] enzymes,[29–31] and assemblies [32–34]) have been made computationally using programs like Rosetta that automate many aspects of the design process, such as assembling the backbone, designing the sequence, and calculating the structure's energy using an energy function.[35,36] In addition, high-throughput experimental testing is now possible as thousands of design sequences can be genetically encoded, expressed *in vitro*, and identified by next-generation sequencing.[37]

Despite these advances, the challenge of designing proteins that fold and remain folded as designed is largely still with us today. One of the earliest designs in 1984, a β-bellin, was soluble and showed well-resolved 2-D nuclear magnetic resonance (NMR) spectra only after nine iterative attempts of design and testing.[21] Over the past decade, the success rate (i.e. designs that fold as designed by biophysical characterization / all designs tested) of protein design efforts *in vitro* varied, irrespective of the scale of the experiment – 9% (1 successful / 11 tested) for a βαββαβ protein,[38] 11% (2/19) for a β-sheet protein,[39] 18% (15/83) for an αα-repeat protein,[40] and 25% (5/54) for three small folds.[41]

To generate more stable proteins, one could computationally sample more configurations of a design's backbone and sequence to increase the pool of structures with lower calculated energies. However, this would take more time and require more computing resources.[15,42] Even if a protein designer succeeds in generating more stable structures, the success rate may not necessarily increase.

Machine learning and re-weighting the energy function

Two approaches that could improve the success rate for designing stable proteins are building predictive models and optimizing the weighted terms of the Rosetta energy function used for evaluating protein structures.

First, machine learning models have recently been developed to predict changes in protein folding stability and thermal stability due to mutation, but they include dozens to hundreds of features that make it difficult to interpret the model.[43–46] Simple physics-based models, however, are more interpretable and have been shown to explain mutational effects on binding energy [47] and stability.[48] What has been unexplored is utilizing large-scale protein folding stability data to build an interpretable low-resolution physical model that can quantitatively explain how each feature contributes to a protein's folding stability. This could enable protein designers to better understand how specific determinants play a role in the stability of a protein structure. Then, they can select which features to add or remove in order to design stable proteins for various applications.

Second, re-weighting the terms of the Rosetta energy function using large-scale data has been shown to improve the identification of small molecules that disrupt protein-protein interactions [49] and predict the effect of coordinated water molecules at protein interfaces.[50] However, it has not yet been demonstrated whether optimizing the function's weights based on large-scale folding stability data can lead to generating stable miniproteins with a higher success rate. Because previous research has shown that one can successfully re-weight the energy function for a specific task, it may be possible to do so with designing stable miniproteins.

αββα miniprotein as a model system

To test these two hypotheses, it is important to select a protein that could be used as a model system and has low folding stability so that we can study and improve it. Additionally, the ability to test for folding stability in a high-throughput manner is also crucial because this will allow us to generate large datasets that can be used to build interpretable models and optimize the Rosetta energy function.

We recently combined computational protein design with a high-throughput protease stability assay to test the folding stability of thousands of miniproteins (43-residues) bearing one of four folds.[37] Individual members in a miniprotein library were displayed on the surface of yeast, tagged with a fluorophore, and simultaneously subjected to varying concentrations of protease (trypsin or chymotrypsin). Yeast cells were sorted by flow cytometry, and the frequencies of each miniprotein were identified by next-generation sequencing. This high-throughput and massively parallel assay allowed us to quantify EC_{50} protease values and calculate a "stability score" for each miniprotein.

There are several strengths to this method over other approaches that measures folding stability, such as calorimetry, UV-absorbance, and X-ray scattering.[51] First, the massively parallel approach reduces experimental bias, and the high-throughput nature of the assay enables us to analyze biophysical features related to stability with statistical significance. In addition, the diversity of miniprotein sequences in each library allows us to devise general principles.

Surprisingly, it was difficult to design stable miniproteins with an αββα fold. When using the protease stability assay, the success rate was 0% in the first round of design and test, and 2% by the fourth round.[37] αββα miniproteins, therefore, seemed to be a good model system to study folding stability because there was something about this topology that we did not understand. As “de novo design provides a rigorous test of our understanding of protein structure,”[42] Chapter 2 of this dissertation builds on this previous work by designing, testing, and evaluating a fifth and sixth round of new αββα miniproteins that are more stable than previous rounds. This provided us with large-scale data from which we could analyze determinants of folding stability.

1.3 Towards developing a method to identify cell-penetrating miniproteins

Need for cell-penetrating protein therapeutics

One application for *de novo* protein design is to confer them a function that can meet unsolved needs, specifically the ability to penetrate the cell plasma membrane and target cytosolic proteins.

Currently, over 85% of the human proteome is undruggable because 80% of all FDA-approved drugs are small molecules, which are designed to target well-defined binding pockets of proteins.[52] However, many proteins function through protein-protein interactions (PPIs), which are challenging to target using small molecules because PPIs occur between shallow interfaces.[53] A growing class of protein therapeutics like antibodies,[54] antibody mimetics,[55] stapled peptides,[56] and miniproteins[23] are promising alternatives as they can fold into a rigid conformation and bind to a protein target and small enough to be thoroughly characterized.[53] However, most of these proof-of-concept and FDA-approved protein therapeutics mainly target extracellular proteins or cell-surface receptors due to their inability to cross cellular membranes, escape endosomes, and enter the cytosol, where many disease-implicated proteins exist.[57]

There are a few examples of eukaryotic and engineered proteins that are reported to cross the cell plasma membrane, such as histone proteins,[58] the phosphatase PTEN,[59] and the engineered ZF5.3.[60] But, this is the exception rather than the rule because the hydrophilic surfaces of proteins readily do not pass through the nonpolar lipid bilayer that comprise cellular membranes.

Cell-penetrating peptides (CPPs) have long been reported to cross cell membranes, but they are unstable and degrade in the presence of extracellular and intracellular proteases.[61] To overcome this limitation, CPPs are often linked to a protein or peptide (often referred to as “cargo”) so that the CPP delivers the cargo across the plasma membrane.[62] However, this lowers the rate and efficiency of cell penetration (up to 4-8 fold),[63,64] and covalently-attached CPPs can alter the bioactivity of the cargo protein.[61] Alternative vehicles for delivery, such as gold nanoparticles, quantum dots, and carbon nanotubes also have limitations, namely requiring chemical modification, poor delivery, and low biocompatibility, respectively.[61]

Towards designing novel cell-penetrating miniproteins

A new approach is to develop cell-penetrating miniproteins (CPMPs) whose single small domain (< 10 kDa) has all of the sequence and structure requirements for cell-penetration, endosomal escape, and target binding. Such a miniprotein has the potential to overcome the limitations of CPP-conjugated proteins or other delivery systems. First, the engineered ZF5.3 (3.5 kDa) has been shown to efficiently escape endosomes and enter the cytosol,[60,64] suggesting that there may be a “goldilocks” range for size — larger than a CPP so that it can fold into a rigid structure and be resistant to proteolysis, but small enough to cross cell membranes. Second, a purine nucleoside phosphorylase with a CPP-motif grafted onto its solvent-facing loop was demonstrated to enter the cell cytosol without affecting its function.[65,66] This suggests that conjugating a protein to a CPP may not be necessary. Instead, it may be possible for a single protein fold to have both functional and cell-penetrating capabilities.

However, another engineered CPMP, aPP5.3, showed very little ability to enter the cell cytosol, despite a similar structure and the same number of arginine mutations as ZF5.3.[60,67] This suggests that our understanding of the determinants for cell-penetration is limited. What is needed is large-scale design, testing, and analysis of protein structures and sequences that could reveal general principles for cell-penetration.

Towards building a novel assay that can screen for cell-penetrating miniproteins

Currently, there is no high-throughput screen for CPMPs. When we look to *in vitro* CPP studies, the scale of experimentation is low to medium-throughput in scale, ranging from two peptides[19] to a library of 128 peptides.[68] To test for cell-penetration, CPPs are often tagged or linked to a dye, peptide, or chemical so that one can identify and measure a fluorescence intensity or a phenotypic readout.[69] However, these methods have limitations. Solely relying on fluorescent dyes with cell-fixation or split-fluorescent tags that reconstitute a fluorescent protein require microscopy analysis, which can lead to artificial intensity signals.[19,70,71] In addition, incorporating compounds, such as chloroalkane or azide, into CPPs so that they bind to cytosolic HaloTag and lead to a fluorescent or luminescent readout [68,72,73] can result in false negatives as the compound can potentially be degraded.[69]

A proposed alternative is to build a “reporter” cell line that activates a fluorescent protein only when a CPMP enters the cytosol via an intein-mediated splicing reaction. Split-inteins are protein fragments that spontaneously undergo a trans-splicing reaction, resulting in the reconstitution of the inteins and the ligation of the exteins (regions flanking the inteins).[74] In this new system, a candidate CPMP is fused to a small intein fragment, and a synthetic transcription factor that drives the gene transcription of a fluorescent protein is fused to the complementary intein fragment. In addition, the transcription factor and its linked intein are sequestered away from the nucleus by being bound to a transmembrane domain (TMD). Only if the CPMP enters the cytosol, the trans-splicing reaction between the two split-inteins will occur, thereby releasing the transcription factor and initiating the transcription of a gene encoding a fluorescent protein. A recent study demonstrated that split-inteins can be combined with synthetic transcription factors to regulate gene expression,[75] and one report electroporated the split-intein VidaL into mammalian cells to release a histone protein tethered to a TMD.[76] However, there is currently no study that has combined split-inteins and synthetic transcription factors to test for CPMPs.

There are several advantages for building this kind of reporter system. First, any fluorescence is a true signal for cell-penetration as all components are orthogonal to the cell, and all are necessary for gene expression to occur. Second, gene expression as a readout enables the identification of low-efficiency CPMPs because the reporter cell can amplify weak signals. This will allow us to identify biophysical properties that correlate with weak cell-penetration and the subsequent opportunity to improve those CPMPs. Third, this assay can be generalizable to any CPMP candidate as long as it is tagged to an intein. Finally, it could be possible for this system to be scaled to screen thousands of CPMP candidates using droplet microfluidics. A recent study demonstrated that cytokines could be screened by passing droplets of co-encapsulated single “secretor” yeast cells and single “reporter” mammalian cells through a microfluidic device.[77] However, this approach has not yet been applied to screen for CPMPs, and thus this presents an opportunity to build a new technology. Chapter 3 of this dissertation provides initial experiments demonstrating that combining the synthetic transcription factor ZF1-VP64[75,78] with the split-intein VidaL[76] could be used towards building a reporter cell. Finally, Chapter 4 describes future directions for optimizing the reporter cell and creating a secretor cell.

CHAPTER 2: Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation

The text of Chapter 2 is a reprint of the material from:

> Kim, T.E. et al. Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. *Proceedings of the National Academy of Sciences*. **119** (41), e2122676119. 2022

2.1 Abstract

Designing entirely new protein structures remains challenging because we do not fully understand the biophysical determinants of folding stability. Yet some protein folds are easier to design than others. Previous work identified the 43-residue αββα fold as especially challenging: the best designs had only a 2% success rate, compared to 39-87% success for other simple folds.[37] This suggested the αββα fold would be a useful model system for gaining a deeper understanding of folding stability determinants and for testing new protein design methods. Here, we designed over ten thousand new αββα proteins and found over three thousand of them to fold into stable structures using a high-throughput protease-based assay. Nuclear magnetic resonance, hydrogen-deuterium exchange, circular dichroism, deep mutational scanning, and scrambled sequence control experiments indicated that our stable designs fold into their designed αββα structures with exceptional stability for their small size. Our large dataset enabled us to quantify the influence of universal stability determinants including

nonpolar burial, helix capping, and buried unsatisfied polar atoms, as well as stability determinants unique to the αββα topology. Our work demonstrates how large-scale design and test cycles can solve challenging design problems while illuminating the biophysical determinants of folding.

2.2 Introduction

Improving our understanding of the determinants of protein stability [3,79,80] would accelerate biological, biomedical, and biotechnology research. In particular, computational models of protein stability are commonly used for a range of applications, including protein design,[15,81,82] stabilizing naturally occurring proteins,[83,84] and predicting the effects of point mutants.[85–87] However, all of these models have important limitations. For example, most computationally designed proteins made by experts fail to fold and function.[23,88,24] Non-experts avoid computational design techniques because they are not reliable. These challenges stem from our incomplete understanding of the biophysical determinants of folding stability and from the difficulty of encoding these determinants into computational models for practical applications.

Recently, we introduced a high-throughput approach to study protein folding stability that is particularly helpful for improving computational modeling and design. In our approach, we designed thousands of de novo proteins and measured their folding stabilities using a yeast display-based proteolysis assay coupled to next-generation sequencing.[37] Several new studies have applied our methodology [89–91] as it has several advantages. First, measuring folding stability for thousands of proteins makes it possible to statistically quantify biophysical features that contribute to stability. Second, examining diverse sequences makes it easier to derive principles that are not specific to a particular protein context. Finally, assaying computationally designed proteins focuses the experimentation on the regions of sequence and structural space that are predicted to be low energy according to a particular computational model, which is especially useful for improving that model.

We previously used this approach to increase the success rate (i.e. fraction of designs that form stable, folded structures) of de novo miniprotein designs from 6% to 47%.[37] Three different protein topologies could be designed very robustly (39-87% success), but a fourth topology (αββα, 43 residues) proved very challenging. Only 2% of αββα designs folded into stable structures despite the simplicity of the structure and four repeated efforts to improve the design procedure (Fig. 1A). This suggested that our design procedure and stability model were missing something fundamental about the αββα topology, and that this particular fold could be a useful model system for building a deeper understanding of folding stability. Here, we investigated this by asking two main questions. First, how can we improve our design procedure to obtain a large number of stable αββα proteins for further analysis? Notably, there are no naturally occurring examples of the 43-residue αββα fold for us to learn from, although this architecture is similar to the unusual 55-residue αββα fold of the gpW protein from bacteriophage lambda.[92] Second, how do the biophysical and topological features of different αββα designs combine to determine each protein's folding stability? We investigated these questions by designing and experimentally testing over ten thousand new αββα miniproteins using our high-throughput approach. We also examined whether the structure prediction model AlphaFold 2 [93] could be applied to differentiate stable and unstable designs.

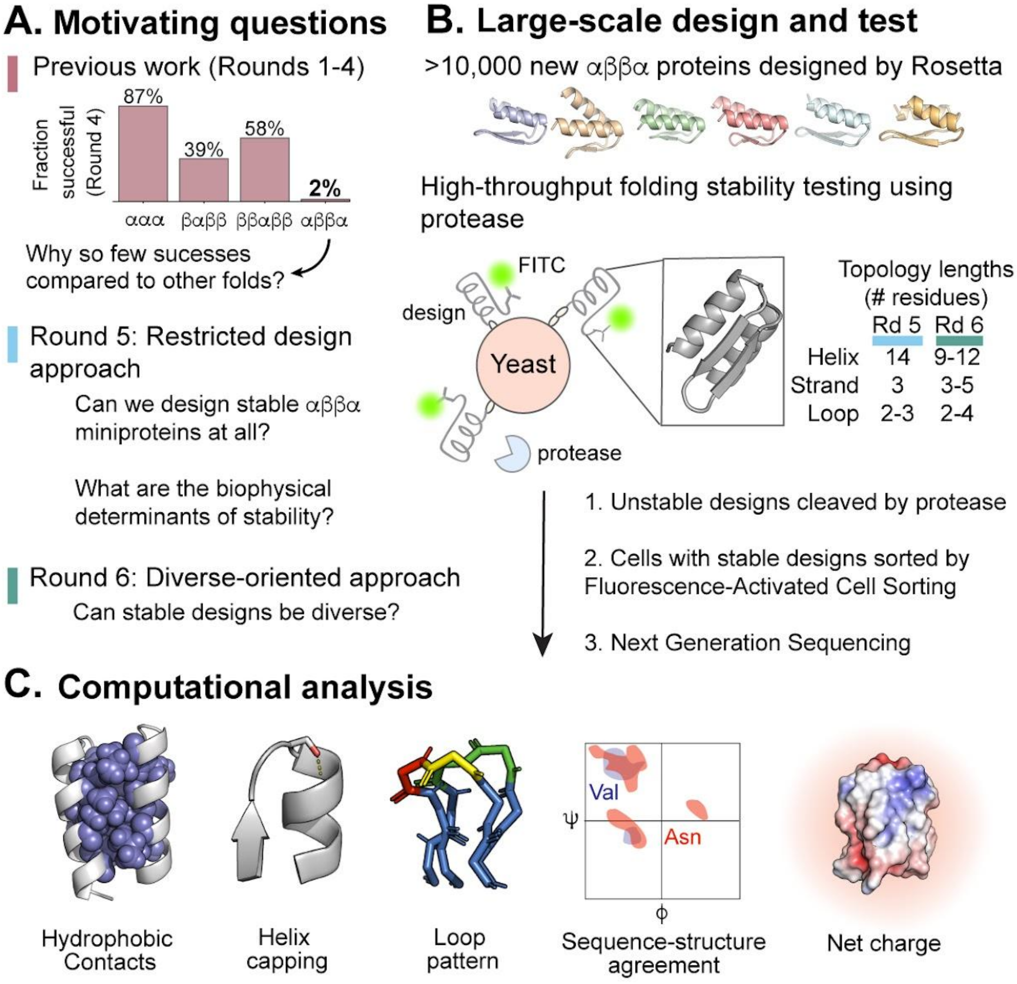

A. Motivating questions
Previous work (Rounds 1-4)
Fraction successful (Round 4)
87%
39%
58%
2%
ααα
βαββ
ββαββ
αββα
Why so few sucesses compared to other folds?
Round 5: Restricted design approach
Can we design stable αββα miniproteins at all?
What are the biophysical determinants of stability?
Round 6: Diverse-oriented approach
Can stable designs be diverse?
B. Large-scale design and test
>10,000 new αββα proteins designed by Rosetta
High-throughput folding stability testing using protease
FITC
design
Yeast
protease
Topology lengths (# residues)
Rd 5
Rd 6
Helix 14 9-12
Strand 3 3-5
Loop 2-3 2-4
1. Unstable designs cleaved by protease
2. Cells with stable designs sorted by Fluorescence-Activated Cell Sorting
3. Next Generation Sequencing
C. Computational analysis
Val
Asn
ψ
ϕ
Hydrophobic Contacts
Helix capping
Loop pattern
Sequence-structure agreement
Net charge

Fig. 1. Design strategy for generating and testing αββα miniproteins. (A) Previously, we performed four iterative design-test-analysis cycles to generate stable αββα miniprotein designs, but only achieved a 2% success rate.[37] (B) Here, we designed thousands of new αββα miniproteins using Rosetta (6,000 designs in Round 5 and 5,307 designs in Round 6) and experimentally tested them for their folding stability using a combined yeast display and protease sensitivity assay. (C) We then performed computational analysis to identify and understand the relative importance of key stability determinants (e.g. hydrophobic contacts, helix capping, loop patterning, local sequence-structure agreement, and net charge).

2.3 Designing αββα miniproteins using a restricted design strategy

We first computationally designed thousands of new αββα miniproteins ("Round 5") based on lessons learned from our previous four rounds of design.[37] All designs were based on a single protein architecture [94] that previously led to the greatest number of stable designs (Fig. S1A). This architecture restricted our new αββα miniproteins to 14-residue α-helices, 3-residue β-strands, and a specific loop structure (Fig. 1B). In addition, we ensured our designs met strict criteria for buried nonpolar surface area, Rosetta energy, and predicted secondary structure (Fig. S1B). Finally, we required the middle loop to have a hydrophobic residue, required solvent-facing residues on the β-strands to be polar or charged, set a minimum threshold for the total number of hydrophobic residues, and eliminated Gly, Thr, and Val in helices (Fig. S1C-D) (see Methods). We hypothesized these restrictions would increase the success of our new designs because these constraints would enforce the overall αββα topology and build a larger hydrophobic core. However, this would reduce the potential sequence and structural diversity.

Based on this "restricted" design strategy, we generated 28,000 αββα miniproteins using an improved version of the Rosetta score function. This score function was previously parameterized to correlate with our earlier high-throughput data on miniprotein folding stability.[95] In addition, we used an improved sequence sampling procedure that minimizes over-compaction and produces more native-like protein cores containing bulky residues.[96] Our final set of 6,000 αββα designs were chosen by ranking the predicted stabilities of all 28,000 αββα designs using a linear regression model trained on previous large-scale αββα stability data.[37] Because this regression model included a low number of stable designs (60/2830), we used this model for the practical task of selecting designs, but we did not expect reliable performance. After we ranked our designs, we eliminated designs that were more than 31/43 residues identical to a higher-ranking design. Within our final set of 6,000 designs, the median backbone RMSD between any two designs was 2 Å (Fig. S2) and the median sequence identity was 35% (Fig. S3). Each design based on this restricted strategy is named HEEH_TK_rd5_####, where HEEH indicates the pattern of α-helices (H) and β-strands (E), TK indicates the designer (author TEK), rd5 indicates these designs follow our four previous efforts,[37] and #### is the design number.

2.4 Biophysical characterization of αββα miniproteins using a restricted design strategy

We measured the folding stabilities of our newly designed αββα miniproteins using the high-throughput protease sensitivity assay introduced previously (Fig. 1B).[37] Briefly, all sequences were synthesized as DNA oligonucleotides in a pooled library. We then used S. cerevisiae to express and display our sequences on their cell surface, along with a c-terminal myc tag. Next, we subjected the yeast cells to varying concentrations of trypsin and chymotrypsin (tested separately) (Fig. S5A-B) and fluorescently labeled the cells displaying protease-resistant sequences. Finally, we sorted the fluorescently labeled cells by flow cytometry and identified the protease-resistant sequences by deep sequencing (Fig. 1B). Out of the 6,000 designs, only 5,662 designs had sufficient sequencing counts to precisely determine their protease sensitivity, and we used this set of 5,662 designs for our analysis. As previously, we assigned each design a “stability score”, defined as the difference between that sequence’s observed protease sensitivity and the predicted sensitivity of that sequence in its unfolded state. Each one-unit increase in stability score indicates a 10-fold higher amount of protease required to cleave that sequence under assay conditions, compared with the predicted protease concentration required to cleave that sequence in its unfolded state.[37] To conservatively identify stable designs, each design’s overall stability score is the minimum of the stability scores observed separately with trypsin and chymotrypsin. We previously observed that sequences of scrambled amino acids (not designed sequences) rarely have stability scores above 1, and so we classify designs as stable when their stability score exceeds 1.

Our set of 5,662 designs had an average stability score of 0.81, and we classified 38% of these designs as stable (stability score > 1, Fig. 2A). The stable set had a median pairwise sequence identity of 37% (Fig. S3). This greatly exceeded our previous success rate of 2% (Fig. 1A).[37] We also included control sequences in our library whose residue compositions matched our αββα designs, but with the ordering of the residues scrambled in a specific manner: polar residues remained polar, nonpolar residues remained nonpolar, and proline and glycine residues remained in their identical positions. In contrast to our designs, almost all scrambled sequences had stability scores < 1 with an average stability score of -0.86 (Fig. 2A). This suggests that the protease resistance observed for a subset of designs can be attributed to the folding stability of their designed structures, rather than generic properties of their sequences such as residue composition or patterning. In addition, stability scores measured using trypsin and chymotrypsin were correlated with each other despite the differing specificities of the proteases (Fig. S5A-B). This further indicates that our measured stability scores reflect folding stability rather than protease-specific factors.

We next sought to verify that stable αββα miniproteins folded as designed using several orthogonal approaches. First, we selected six stable αββα designs with varying hydrophobicity values [97] and individually purified them from E. coli (Fig. 3A; Table S1) for circular dichroism (CD) and thermal denaturation. Protein purification by size-exclusion chromatography revealed that three of the six miniproteins (HEEH_TK_rd5_0420, HEEH_TK_rd5_0614, and HEEH_TK_rd5_0958) predominantly eluted at the expected molecular weight of a monomer, whereas the other three showed both monomeric and dimeric peaks (Fig. 3B).

CD spectra exhibited helical secondary structure and reversible folding after heating to 95°C (Fig. 3C), but the initial 25°C and cooled 25°C return measurements for HEEH_TK_rd5_0341 and HEEH_TK_rd5_3711 were not superimposable, possibly due to aggregation that altered the signal intensity. None of the designs showed a clear melting transition, although designs HEEH_TK_rd5_0958 and HEEH_TK_rd5_3711 lost much of their helical character at 95°C. In contrast, design HEEH_TK_rd5_0420 was minimally perturbed during melting (Fig. 3D), indicating extreme thermostability.

Next, to spot-check the accuracy of our designed structures, we solved the structures of HEEH_TK_rd5_0958 and HEEH_TK_rd5_0341 by nuclear magnetic resonance (NMR). For HEEH_TK_rd5_0958, the average backbone root-mean-squared deviation (RMSD) of the design model compared to all 20 structures in the NMR ensemble was 1.26 Å (Fig. 3F). For HEEH_TK_rd5_0341, both monomers in the dimeric structure were also very close to the designed monomeric model: the average backbone RMSD of the design model compared to all 40 structures in the NMR ensemble was 1.65 Å (Fig. 3H). The structure was symmetrical so only one HSQC peak was visible for each residue, although 15N NMR relaxation measurements were consistent with the dimeric state. The two monomers come together near the β-hairpin and designed N- an C-termini, burying hydrophobic residues in that region (Fig. 3I).

To analyze the structural differences between the design models and NMR structures, we quantified the number of contacts a residue in the design model gained or lost in the NMR model (Fig. S6). Most of the residues gained or lost zero or one contact, indicating the close structural similarity between the design model and NMR ensemble. For HEEH_TK_rd5_0341, the protein was designed to form a monomer. So, residues at the dimeric interface (the N- and C-termini and the hairpin turn) all gained new contacts, changing the environments of these residues (Fig. S6A). However, as shown by the overall RMSD, these changes did not affect the overall structure of each monomeric subunit.

We also examined the local stability of designs HEEH_TK_rd5_0958 and HEEH_TK_rd5_0341 by hydrogen deuterium exchange (HDX) NMR. The HDX opening free energies revealed differences in local stability in different regions of the topology (Figs. 3G, 3J). The most stable secondary structure was Helix 2 for both miniproteins, with opening energies around 4 kcal/mol at 15°C (compared to ~2-3 kcal/mol in Helix 1). The central β-hairpin was the least stable structure in HEEH_TK_rd5_0341 (Fig. 3J). Four residues in this hairpin (I21, G23, I24, and V26) form intramolecular hydrogen bonds that should protect those amides from exchange (Fig. S7A) but three of these residues exchanged too quickly in HEEH_TK_rd5_0341 to be measured by NMR. In contrast, three of the four hairpin residues that form intramolecular hydrogen bonds in HEEH_TK_rd5_0958 had measurable protection from exchange (Fig. 3G, Fig. S7B) and were similarly stable to Helix 1. Overall, the hierarchy of stabilities between Helix 2, Helix 1, and the central β-hairpin suggests the folding energy landscape is not fully cooperative.

The highest opening energy in the monomeric HEEH_TK_rd5_0958 was 4.5 kcal/mol, observed at I35 (Figs. 3J, 3G). This highest opening energy typically indicates the global stability of the protein,[98] making HEEH_TK_rd5_0958 almost 2 kcal/mol more stable than the previous highest stability observed for a designed αββα structure.[37] However, this higher stability was observed at a lower temperature (15°C instead of 25°C in ref. (35)) and in the presence of D_2O, which typically stabilizes proteins.

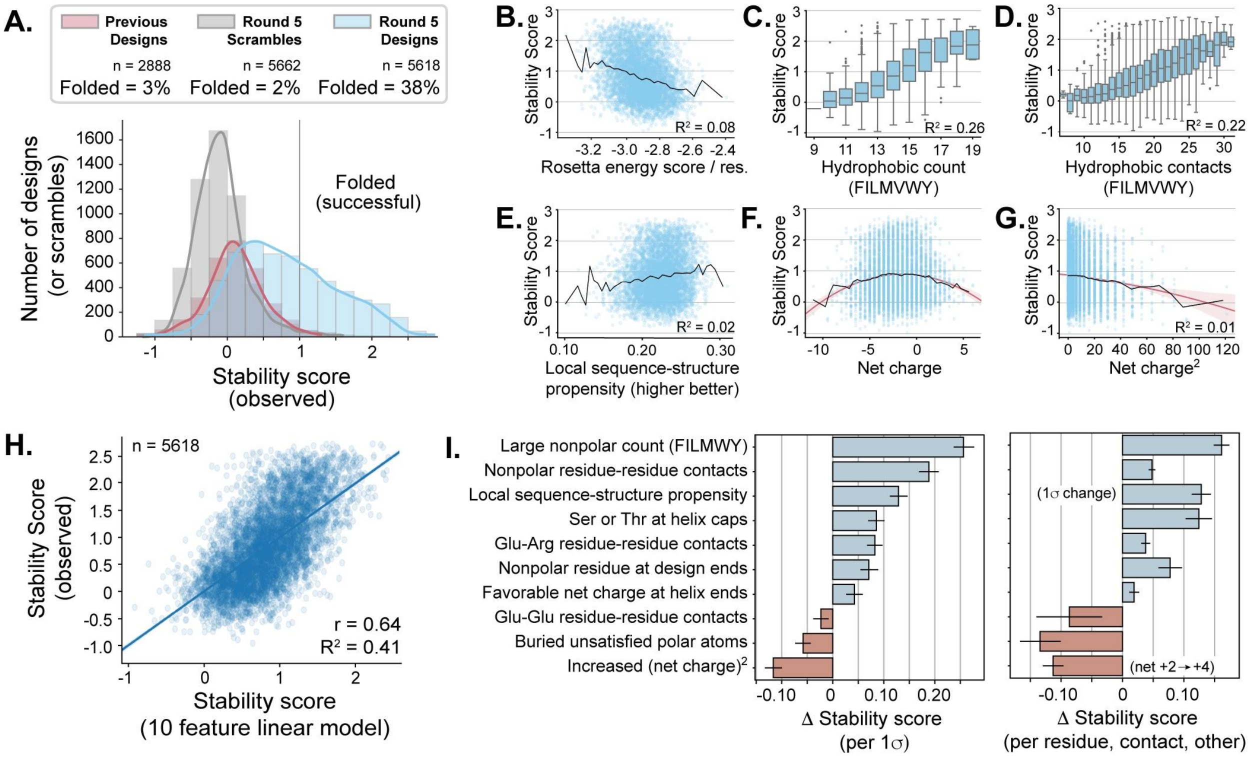


Fig. 2. Experimental testing and analysis of αββα stability determinants from a restricted design strategy. (A) The stability score distributions of designed αββα miniproteins (blue), scrambled sequences (gray) and previously published αββα miniproteins (red);[37] the vertical line at stability score = 1 denotes the threshold above which we consider a design to be stable. (B-G) The relation between Individual protein features and stability score. For Rosetta energy, lower values indicate favorable energies, and for local sequence-structure propensity, higher values indicate favorable propensity. Black lines show moving averages; red lines show fits to quadratic (F) and linear (G) models. (H) A ten-feature linear regression model was built using normalized data, and the experimental stability scores are compared to the model's predicted stability scores. (I) The magnitudes of the coefficients from the model based on their importance in the dataset (left) and their biophysical strength (right). Error bars indicate 95% confidence intervals from bootstrapping.

2.5 Stability determinants of αββα designs from a restricted design strategy

We next investigated which design features correlated with folding stability. To this end, we computed over a thousand structural and sequence-based metrics for each design and analyzed whether particular metrics correlated with stability. Several of the strongest individual correlations are shown in Fig. 2. Designs were generally more stable if their Rosetta energy scores were lower (Fig. 2B) and had more hydrophobic residues and hydrophobic sidechain contacts (Fig. 2C-D). Hydrophobic residue count correlated more strongly with stability than Rosetta energy. Stability also increased if a design's sequence was highly compatible with its local backbone structure (see Methods and Fig. 2E). Finally, increased net charge destabilized our designs, although the optimal net charge was slightly negative (Fig. 2F-G). This stability change was approximately linear with the square of the net charge, as expected.[99]

We also explored whether specific residues could individually have large influences on the stabilities of the designs. Because all designs are based on an identical architecture, each position in the sequence shares an identical structural role in all designs. Using the binomial test, we identified positions where specific amino acid identities had large and significant changes on the success rates of the designs (Fig. S8). Two positions near the N- and C-termini stood out as particularly important. Positions 2 and 39 are near the tips of each helix and contact each other in space (Fig. S8D). Across the design set, leucine residues at these positions increased the success rate of the designs by 25-39%, whereas other residues such as glutamate and tryptophan decreased the success rate by similar amounts. These differences in success rates were highly significant (adjusted p-value < 10-18) (Fig. S8C). The importance of these residues suggests that termini of the helices play an especially important role in the overall stability of designed αββα miniproteins.

To further examine individual residue contributions to stability, we performed deep mutational scanning analyses (Fig. S9) on the six αββα designs whose structures we verified by CD (Fig. 3C). Using our protease sensitivity assay (Fig. S5C-D), we measured the folding stability changes for all single mutants of each design (Fig. S9). Four of the six mutational scans showed many destabilizing mutations from replacing nonpolar residues in both the helices and the strands. A fifth design (HEEH_TK_rd5_0420) showed a similar pattern, but the helical residues seemed less sensitive to mutations than the strands. The high stability score of HEEH_TK_rd5_0420 (at the peak of our assay's dynamic range) may have limited us from resolving the stability effects of other mutations in the helices (Fig. S9C). The sixth design (HEEH_TK_rd5_0018) showed many destabilizing substitutions at nonpolar sites in the helices, but only a small number could be observed in the β-hairpin, suggesting the hairpin may be less structured in this design (Fig. S9A). Overall, the positions that were most sensitive to mutations (change in stability < 1) were found in the buried hydrophobic core (Figs. S10-11), and in particular large hydrophobic residues (Fig. S11). In contrast, hydrophobic residues, as well as polar and charged residues, that were more solvent exposed in the design models were less sensitive to mutation (Fig. S11). The specific sequence-stability relationships shown in the mutational scanning data suggest that the designs fold into specific structures. Furthermore, the consistency between a nonpolar residue’s burial in the designed models and its sensitivity to mutation (Fig. S11) provides support that the stable designs fold into their designed structures.

Charged and polar residues also contributed to folding stability, although they were less important than buried hydrophobic residues. The top three polar positions that were most sensitive to mutation (average change in stability < -0.5) were positions 15 (end of first helix), 28 (helix-capping position) and 31 (start of second helix that forms hydrogen bonding with the backbone) (Fig. 3E, Fig. S12). These positions indicate the importance of polar interactions toward stabilizing our designs and also support that the designs fold into their specific designed structures.

However, our mutational data also revealed some unexpectedly stable mutants (Fig. S9). For example, we expected that mutants to G23 would be highly destabilizing because G23 should be critical for forming the central β-hairpin. However, in four of the six designs, mutants to G23 could actually increase folding stability (Fig. S9B-D, F). To investigate this, we predicted the structures of all mutant sequences using AlphaFold 2.[93] Although most mutants were predicted to have similar structures to the original design, some predictions (including mutants of G23) suggest the possibility of alternative, compact structures (Fig. S13).

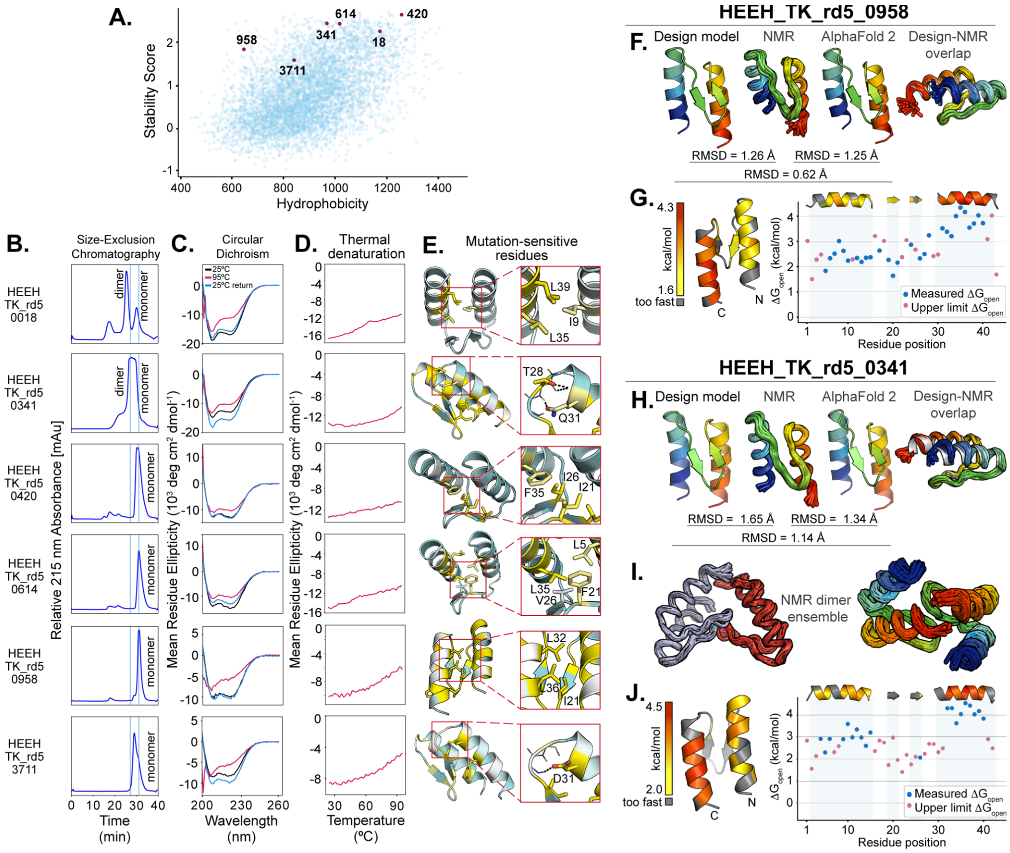


Fig. 3. Biophysical characterization of αββα miniproteins made using a restricted design strategy. (A) The stability scores of all αββα miniproteins made using a restricted design strategy are plotted by their hydrophobicity values.[99] We selected six miniproteins (red dots) with varying hydrophobicity and (B) purified them by size-exclusion chromatography; vertical lines indicate expected dimeric and monomeric forms of the miniprotein based on a calibration curve (see Methods). (C) Far-ultraviolet circular dichroism spectra are shown at 25°C (black), 95°C (red), and 25°C after melting (blue). (D) Thermal denaturation was measured at 220 nm at every 1°C from 25°C to 95°C. (E) Design models highlight positions that are most tolerant (teal) or least tolerant (yellow) to mutations. Key residues that stabilize the protein are shown in stick representation. Each miniprotein's color scale is different to highlight the relative stabilizing or destabilizing effects within each protein; see Fig. S5 for complete data. (F) Comparison of HEEH_TK_rd5_0958 Rosetta design model, NMR ensemble, and AlphaFold 2-predicted structures; overlay of the Rosetta design model (gray) and NMR ensemble (rainbow). (G) Opening energies determined by hydrogen-deuterium exchange for HEEH_TK_rd5_0958. Observed measurements are colored red-yellow on a cartoon model and plotted in blue. For residues that exchanged too quickly to measure, the upper limit of ΔG_{open} is plotted in red. (H) Comparison of HEEH_TK_rd5_0341 Rosetta design model, NMR ensemble, and AlphaFold 2-predicted structures; overlay of the Rosetta design model (gray) and NMR ensemble (rainbow). (I) NMR dimer structure shown in two different perspectives. (J) Opening energies determined by hydrogen-deuterium exchange for HEEH_TK_rd5_0341. Observed measurements are colored red-yellow on a cartoon model and plotted in blue. For residues that exchanged too quickly to measure, the upper limit of ΔG_{open} is plotted in red.

2.6 Modeling relative contributions of biophysical determinants on folding stability

Our previous analysis identified individual determinants of stability without considering how various features relate to each other. Hence, we next analyzed which protein features were the most important contributors to stability and how they compared to each other. Instead of prioritizing predictive accuracy, we used linear regression to build a parsimonious, interpretable, low-resolution model. Our moderately accurate model (r = 0.64, r2 = 0.41; Fig. 2H, Table S3) included ten features chosen for either their large individual contributions to stability or their biophysical interest. Adding all 25 additional Rosetta energy terms provides only a minimal improvement to this low-resolution model (Table S4).

To analyze the strengths of the different features, we compared the different coefficients both in terms of their importance within our dataset (e.g. the impact of a one standard deviation change in each term, Fig. 2I left) and in terms of their biophysical strength (e.g. the impact of one additional residue, contact, charge, etc., Fig. 2I right). By representing the features in these two ways, we were able to observe how each feature contributes to a design’s stability while holding all other features constant. Relative to the variance in the features, the count of large nonpolar residues is the largest contributor to folding stability (Fig. 2I). Additional biophysical determinants known to stabilize globular proteins,[3,9,100,101] such as contacts between adjacent nonpolar residues and Ser/Thr helix capping, contribute to folding stability as well (Fig. 2I). However, our model also points to the stabilizing role of nonpolar residues at the design ends, which is a feature specific to the αββα topology (Fig. 2I). Whereas previous studies on the relative importance of stability determinants were based on assays that changed one feature on individual proteins,[6,7] our large-scale testing enabled us to analyze over a thousand protein features on several thousand proteins in parallel. This, in turn, allowed us to develop a model that offers criteria for designing even more stable αββα miniproteins.

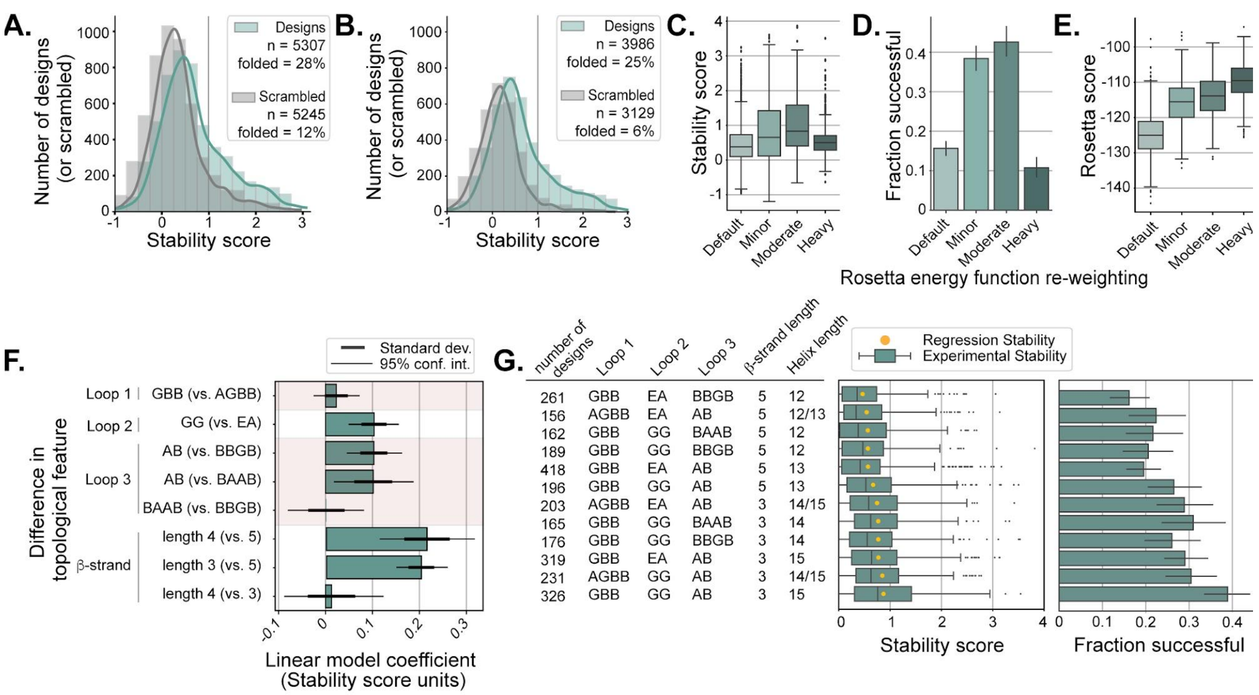
2.7 Designing αββα miniproteins using a diversity-oriented design strategy

Our restricted-design strategy (Round 5) focused on improving the success rate of designing stable αββα miniproteins but at the cost of reducing their structural diversity. Because we were now able to successfully generate stable αββα designs, we next investigated whether we could loosen the design restrictions that we had imposed, increase the diversity of our αββα miniproteins, and identify additional determinants of stability. Hence, we designed a new round of "diversity-oriented" (Round 6) αββα miniproteins based on fourteen different protein architectures instead of one. This allowed designs to have a greater variety of helix, β-strand, and loop lengths, while keeping the overall size of the protein to 43 residues (Fig. 1B). In addition, we did not impose residue restrictions on β-strands or in the middle loop and permitted a greater number of hydrophobic residues.

Importantly, we used our Round 5 stability data to directly re-weight the Rosetta energy function. Using ridge regression, we adjusted the weights on the Rosetta energy terms to create the best correlation with our measured Round 5 αββα stabilities, while regularizing the regression to penalize large deviations from the original weights. With this approach, we created three new energy functions labeled "Minor," "Medium," and

“Heavy” based on how much the weights deviated from the original weights. We used these three energy functions (and the original weights) to design our Round 6 designs (Fig. S14).

We generated ~ 20,000 designs and chose our final set of over five thousand αββα designs for experimental testing by identifying designs that had the greatest structural diversity, varied sequence identity (no closer than 28/43 residues), and an αββα topology as determined by the computer program PSIPRED (33). Notably, we prioritized structural diversity (Fig. S2) in our final selection instead of prioritizing the expected success rate. The median sequence identity across all pairs of sequences was 28% (42% if only nonpolar residues are considered) (Fig. S3). However, the diversity in amino acid composition (overall and nonpolar only) is lower than several known protein domains of similar sizes (Fig. S4). Each design is named HEEH_KT_rd6_####, in which KT indicates the designer (author K.T.), rd6 indicates these designs constitute a new “Round 6” following the previous rounds of αββα design, and #### is a design number.



number of designs	Loop 1	Loop 2	Loop 3	β-strand length	Helix length
261	GBB	EA	BBGB	5	12
156	AGBB	EA	AB	5	12/13
162	GBB	GG	BAAB	5	12
189	GBB	GG	BBGB	5	12
418	GBB	EA	AB	5	13
196	GBB	GG	AB	5	13
203	AGBB	EA	AB	3	14/15
165	GBB	GG	BAAB	3	14
176	GBB	GG	BBGB	3	14
319	GBB	EA	AB	3	15
231	AGBB	GG	AB	3	14/15
326	GBB	GG	AB	3	15

Fig. 4. Experimental testing and analysis of αββα stability determinants from a diversity-oriented design strategy. (A) Stability score distribution of αββα miniproteins (green) and scrambled sequences (gray). (B) As in A, filtered to eliminate designed and scrambled sequences that may fold into non-designed structures; see text and Fig. S7. The vertical line at stability score = 1 denotes the threshold above which we consider a design to be stable. (C-D) Stability scores and success frequencies of designs made with differently-weighted Rosetta energy functions; "Heavy" indicates the largest amount of reweighting. (E) Rosetta scores (using the unmodified score function) of designs made using different weighting; the more positive scores of the designs from the re-weighted energy functions indicate these designs are less favorable according to the default energy function. (F) Stability contribution of the most common loop patterns (using ABEGO notation) and β-strand lengths based on a linear regression model. (G) The most common unique structure combinations (loop pattern, β-strand and helix lengths) are listed (left) followed by the distribution of observed stability scores (middle, with the expected stability from the linear regression model as a yellow dot). At right, the fraction of stable designs for each unique structure. All error bars indicate 95% confidence intervals from bootstrapping.

2.8 Stability determinants of αββα designs based on diversity-oriented design strategy

We tested the stabilities of our “diversity-oriented” αββα miniproteins (and matching scrambled sequences) using the high-throughput protease sensitivity assay.[37] Surprisingly, 12% of our scrambled sequences had stability scores above 1, compared to 2% or fewer in previous rounds (Fig. 4A). We further found that scrambled sequences were most likely to be stable when they were very hydrophobic and when their sequences had high helical propensity as determined by DSSP[102,103] (Fig. S15). This suggested that designed sequences might also be stabilized by these properties alone, even if they did not fold into their designed structures. To remove these potential “false positive” designs from our analysis, we restricted our analysis to designs with a lower nonpolar residue count and lower helical propensity (Fig. S15). Restricting our analysis in this way removed 25% of our total designs, while lowering the fraction of stable scrambles from 12% to 6% (Fig. 4B). The overall fraction of stable designs was 26% - still substantially above the “success rate” of the scrambled sequences (Fig. 4B).

We then analyzed the impact of differently weighted Rosetta energy functions on folding stability. On average, designs made using the reweighted energy functions had higher stability than designs made with the default energy function (Fig. 4C-D). However, some regularization (restraining the weights near their original values) was critical to successful re-weighting: the “Heavy” energy function, where the changes to the weights were the largest, performed much more poorly than the energy functions with “Minor” and “Moderate” changes to the weights (Fig. 4C-D). The success of the re-weighted energy functions suggests that empirical re-weighting could be an efficient practical tool for protein design in situations where large-scale data is available for a specific task. The designs created by the re-weighted energy functions would not have been favored under our previous design procedure, with larger changes to the weights leading to designs that appear less and less favorable according to the default energy function (Fig. 4E). These re-weighted “Minor” and “Moderate” energy functions also showed better correlation with previously published stabilities for other miniprotein topologies compared to the default score function (Table S5).

Next, we investigated how topological features (loop, β-strand, helix) of the designs affect folding stability. We selected the seven most common loop structures found in our designs (represented using ABEGO notation) (36) and the three most common β-strand lengths as inputs to another linear regression model (Fig. S16, Fig. 4F). The explanatory strength of this model is weak (95% conf. int. from bootstrapping, mean r = 0.167 , mean R2 = 0.028). This is due to the simplicity of the model and because the topology-only model excludes critical stability determinants such as hydrophobic residue count. Despite these shortcomings, this model still enables us to examine the relative importance of different topological components. The largest structural contributors to stability are the lengths of β-strands and helices, with shorter β-strands (and corresponding longer helices) as the most favorable topological parameter (β-strand and helix lengths are inversely related because all designs have a fixed length of 43 residues) (Fig. 4F). Secondarily, particular structures in loops 2 and 3 influenced folding stability as well. A loop structure of GBB in the first loop, GG in the second loop, and AB in the third loop increases the stability of a design more than other loop structures (Fig. 4F).

Based on this topology-focused model, we would expect αββα miniproteins with a GBB-GG-AB loop patterning, β-strands that are 4 residues long, and helices that are 14-residues long to be more stable on average than αββα miniproteins with any other loop, strand, and helix combination (Fig. 4F). Although designs with a β-strand length of 4 residues were not common in our dataset, a very similar design structure (GBB-GG-AB with a β-strand length of 3 residues) had the highest average stability score and the highest success rate in our dataset (Fig. 4G), which is in agreement with a previous study on loop patterning and stability.[104] In fact, this design pattern is the protein architecture that we used to generate all the Round 5 αββα miniproteins (Fig. S1A). However, the high success of this architecture in Round 6 may be due to using re-weighted energy functions that were optimized based on Round 5 designs with this specific architecture. Nonetheless, when we subset our Round 6 designs to identify αββα miniproteins with a GBB-GG-AB loop pattern and features that we previously determined to promote stability, these designs are diverse in their sequence identity and highly stable (81% successful) (Fig. 5). This provides a “recipe” for designing new stable αββα miniproteins in the future.

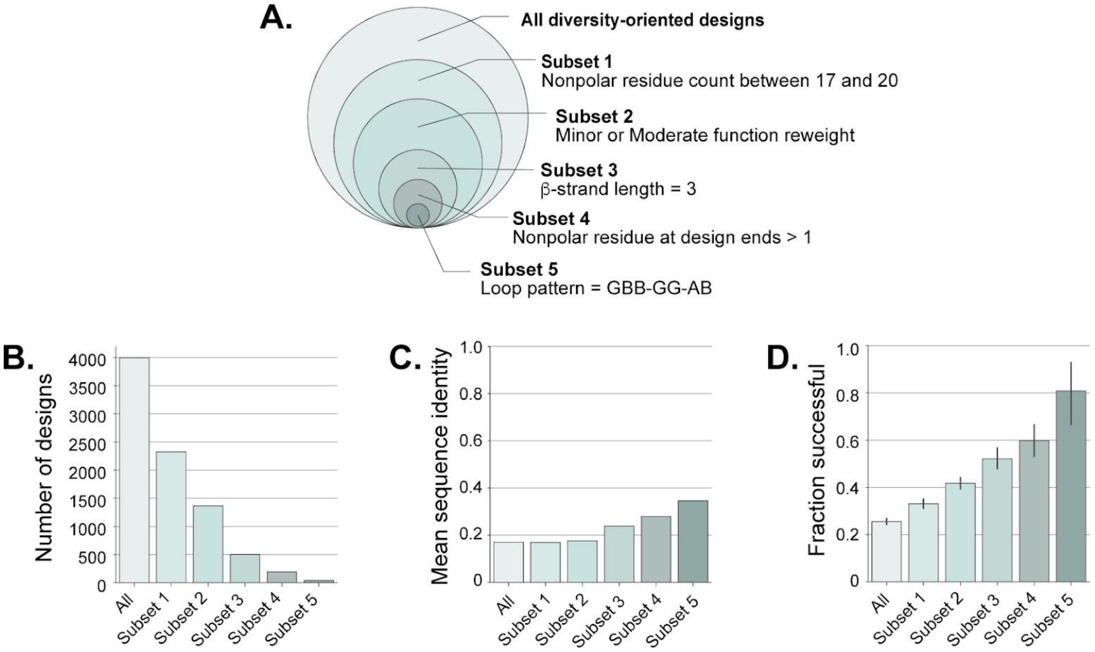


Fig 5. A recipe for building diverse high-stability αββα designs. (A) Designs made from a diversity-oriented strategy are grouped into subsets based on five features that we identified to be important for stability (Fig. 2I, Fig. 4F). (B) The number of designs that comprise each subset; (C) the mean sequence identity between any two designs in each subset; (D) the fraction of successful designs in each subset, with error bars indicating 95% confidence intervals from bootstrapping. Ideal designs (those with the parameters of Subset 5) are 80% successful with under 40% sequence identity between pairs of designs.

2.9 Predicting stable de novo αββα miniproteins by AlphaFold 2

When we designed and tested αββα miniproteins for their folding stability, AlphaFold 2 was not yet available. With its recent release,[93] we wondered whether AlphaFold 2 could discriminate between stable and unstable miniproteins. We explored this possibility even though AlphaFold 2 is intended for structure prediction and not stability prediction. Out of the ~5,600 and ~4,000 restricted and diversity-oriented designs, respectively, we found that 78% of the former and 20% of the latter had at least one predicted structure within 2 Å RMSD to the designed model. These predictions were equally in agreement with design models regardless of whether a design was experimentally unstable, moderately stable, or stable, indicating that AlphaFold 2 did not discriminate stable from unstable designs (Fig. S17). We also examined whether the Rosetta energy scores of the AlphaFold 2-predicted models were better correlated with experimental stability scores than the scores of the original design models. The AlphaFold 2-predicted models did not improve the correlation with experiment for the Round 5 design set, but provided a small improvement for Round 6 (Fig. S18A-D). Neither RMSD nor AlphaFold 2's average confidence measure (pLDDT) showed much ability to enrich for stable designs (Fig. S18E), indicating that AlphaFold 2 is currently unable to determine the folding stability of these designed miniproteins.

2.10 Discussion

Understanding the biophysical determinants that enable proteins to fold and remain stable is important in protein design, drug development, and other areas. Here, we examined the stability determinants of the αββα miniprotein fold, which was previously identified as unusually challenging to design.[37] We took advantage of an improved Rosetta design protocol [95,96] to design over ten thousand αββα miniproteins using both restrictive and diversity-oriented design strategies. Our two design strategies led to over three thousand new stable designs (~2,100 restricted and ~1,000 diversity-oriented designs) and a much higher success rate (38%, Fig. 2A) than the 2% success previously reported.[37] Our designed proteins also had a much higher success rate than control sequences with identical residue composition and polar-nonpolar patterning. This suggests that their stability was conferred by their designed three-dimensional structures. Supporting this, NMR structures of two designs closely matched the designed models (below 2 Å backbone RMSD, Fig. 3), circular dichroism spectra of six designs were consistent with the designed structures (Fig. 3C), and deep mutational scanning analysis of 5/6 designs showed specific sequence-stability relationships that were consistent with the designed structures. However, the lower resolution of circular dichroism and mutational scanning cannot directly demonstrate the atomic accuracy of the designs.

Our large dataset of stable designs enabled us to quantify determinants of stability for the previously-challenging αββα fold (Figs. 2I, 4F-G). Most of the stability determinants were common across globular proteins,[3,9,100,104,6] and similar to those previously observed in large-scale de novo design experiments.[37] We also identified that designing hydrophobic residues near the termini was especially important for the αββα miniprotein fold (Fig. 2I). Our design success rate improved substantially when we used our large dataset to re-weight the Rosetta energy function specifically for αββα design (Fig. 4C-E). These observations largely explain the low success of previously designed αββα proteins: previous designs frequently employed non-optimal loop patterns, helix capping residues, and residues near the design termini, and typically had 13-16 nonpolar residues rather than the 17-20 used here (Fig. S1). Notably, the total number of nonpolar residues in each design is influenced by the design energy function and by parameters that restrict the amino acids that are sampled at each position according to the solvent accessibility of that position.[37,38] These restrictions are manually tuned to balance stability and solubility, as well as to reduce the search space of sequences. Designing proteins with too few nonpolar residues can thus be considered a failure of manual tuning as well as a failure of the design energy function.

Our study has several notable limitations. First, some fraction of "stable" designs are likely stable for non-designed reasons, such as folding into an alternative structure, forming a compact "molten" state, or aggregating on the surface of yeast. In our diversity-oriented set, 6% of our scrambled sequences met our stability threshold, compared with 25% of designs (Fig. 4B). Naively, this suggests that one in five stable designs could be stable for non-designed reasons. In addition, three of the six designs based on the restrictive protocol exhibited some oligomeric species when purified from E. coli (Fig. 3B), suggesting designs might be stabilized by intermolecular interactions. Because our regression analysis assumes that each design's stability (or lack thereof) is due to its designed monomeric structure, our analysis will be unreliable if non-designed structures or interactions played an important role in our observed stabilities. Still, our regression analysis was able to identify specific three-dimensional features as stabilizing or destabilizing, such as buried unsatisfied polar atoms and attractive or repulsive ion pairs (Fig. 2I).

Secondly, our findings regarding the determinants of stability are limited to the specific context we examined: αββα miniproteins designed by a particular computational procedure. The samples of designs that we tested were not random: they were designed to be high stability and showed variation across some dimensions but not others. If a biophysical property (such as backbone torsional strain or higher polarity) varied only minimally across our design set, we would not be able to identify the contribution of that feature to stability. An alternative design procedure might also generate structures in a different region of “property space,” permitting high stability designs that are different from the recipe described in Fig. 5. Constructing a fully general model of folding stability will ultimately require a broad sampling of sequences, structures, and biophysical properties. Our work here investigating a specific design space suggests that this should be possible.

Despite these limitations, our study demonstrates how large-scale experimental testing can be applied to solve a challenging design problem and to quantify the biophysical features that influence design stability. In contrast to other studies that use mutagenesis to study determinants of folding stability,[105–108] our method examines the strengths of different biophysical features across thousands of different protein contexts, although these contexts are all related by the αββα fold and design procedure. Simplified low-resolution models like our linear regression are valuable for building biophysical intuition about the strengths of different interactions [109,110] as well as for guiding the construction of high-resolution models like the Rosetta energy function, which is also an additive model.[111] Our stable αββα designs (and our recipe for generating more) may also be valuable scaffolds for engineering binding functionality for therapeutic, diagnostic, and synthetic biology applications.[23,112,113]

2.11 Materials and Methods

Computational protein design

We designed αββα miniproteins using Rosetta based on our previous work.[37] Briefly, we used fragment assembly to build backbones according to protein architectures specified in a blueprint file,[94] as in[37,38]. For the restricted design strategy (Round 5), we chose the protein architecture that previously led to generating the greatest number of stable (defined only here as stability score ≥ 0.8) αββα miniproteins (Fig. S1A). This architecture restricted the αββα miniprotein structure to have two helices that are 14-residues long, two β-strands that are 3-residues long, and three loops with an ABEGO pattern of GBB, GG, and AB, respectively. We also applied several design constraints. We forced the first residue in the middle loop (position 22) to be nonpolar (AFILMVWY), the solvent-facing residues in the β-strands (positions 20 and 25) to be polar or charged (QNSTDEHKR), and any helical positions were prevented from being designed as Gly, Thr, or Val as they are known to destabilize helix formation (51–53). We also required all designs to possess at least 15 hydrophobic residues (AFILMVWY) and no more than 21 hydrophobic residues. Finally, we filtered out designs with low Rosetta total energy scores or low buried nonpolar surface area (Fig. S1B-D).

Sequence design was performed using the Rosetta protocol FastDesign,[96] the beta_nov16_protease version of the full-atom energy function, and a recently-improved sampling method designed to prevent over-compaction.[50] In order to select αββα miniproteins for experimental testing, we ranked each αββα design by their predicted stability scores, which was determined by a Lasso regression model that we built using previous αββα miniprotein structural metrics and experimental stability scores.[37] Based on this ranking, we selected the top ~5,600 designs with a threshold of 67% sequence identity for experimental testing.

Round 6 designs were designed as above with several changes. First, we utilized 15 different protein architectures. Moreover, we removed the hydrophobic restriction in the middle loop, were more permissive on non-helical residues (GDNST) inside the helices, and allowed hydrophobic residues to appear on the protein surface. We further specified a penalty for a protein's net charge outside the range of -5 or 3. Upon generating ~20,000 αββα designs, we took several steps to select over 7,000 designs for experimental testing. We first built Lasso and XGBoost regression models (54) using experimental data from Round 5 to identify ~3,000 designs with significantly different predictions between the two models (predicted stability scores were at least 0.25 scores away from the best-fit line between the models). We next independently performed principal component analysis to identify ~9,000 designs that were most distant from each other. From the combined ~12,000 αββα designs, we selected ~7,400 designs for experimental testing whose sequence identity was no closer than 66% to any other design.

Although all ~7,400 designs were experimentally tested, we determined afterwards that many of these structures either diverged away from the αββα topology during design or were not predicted to fold into an αββα structure by Rosetta's ab initio algorithm. We further found that scrambled sequences could form secondary structure according to psipred.[114] To focus our analysis on designed αββα structures, we restricted our analysis to ~5,300 designs meeting these criteria: distance between the C-terminus to the middle loop < 22 Å, distance between the N- and C-termini < 20 Å, β-strand lengths according to DSSP ≤5 residues, loop lengths ≤ 5 residues, and unbroken αββα secondary structural elements according to DSSP.[102,103]

Energy function re-weighting

The Rosetta energy function is a weighted sum of individual, independent score terms.[111] To test whether our experimental data could directly optimize the energy function for αββα miniprotein design, we sought to re-weight these terms in Round 6 to produce the best correlation with our experimentally measured stability scores from Round 5. In re-weighting, we also sought to bias our new weights to be as close to the original weights as possible by using ridge regression.[115] However, because the L2 regularization in ridge regression biases coefficients to be near zero, we used ridge regression to identify optimal perturbations to our original weights, rather than directly optimizing the weights themselves. To determine the appropriate perturbations, we first regressed our set of experimentally measured stability scores against the original Rosetta (computational) total scores of the designs. We then used the residuals from this regression (i.e. the error in the prediction of experimental stability score) as the target values for our ridge regression. We used scikit-learn's implementation of Ridge regression[115] to determine new weights on the 25 unweighted Rosetta score terms that best fit the residuals of the first regression (Fig. S7). The coefficients in this second regression are effectively perturbations to the original Rosetta weights that minimize the error in predicting experimental stability scores (subject to the regularization constraint). After performing ridge regression, the new score function weights were determined based on the formula:

$\text{NewWeight}_i = \text{OriginalWeight}_i * (1 + \text{Coefficient}_i)$

where NewWeighti is the new weight on score term i, OriginalWeighti is the original weight for term i in the beta_nov16_protease energy function, and Coefficienti is the coefficient on score term i in the ridge regression.

We tested three new weight sets (“Minor,” “Medium”, and “Heavy”) in addition to the default weights. These new weight sets were determined using three different regularization strengths in the ridge regression and are named based on the magnitude of the change. The “Minor” set used regularization strength alpha=200,000; “Moderate” used alpha=20,000, and “Heavy” used alpha=0.1. “Heavy” corresponds to the value of alpha from a cross-validated ridge regression using scikit-learn’s RidgeCV method.[115] In all weight sets, the score term fa_intra_rep_xover4 was maintained at its default value to avoid favoring extended structures. These weight sets are all provided in the supporting information.

Miniprotein library generation

We reverse translated the residue sequences and optimized the codons (based on E. coli codon frequencies) of all αββα miniproteins that we selected for experimental testing using DNAworks 2.0.[116] We also included scrambled sequences (while preserving locations of P and G residues as well as nonpolar/polar patterning) for each corresponding αββα sequence (following [117]). Both oligo libraries (Round 5, and Round 6 + mutational scanning) were purchased from Agilent.

Yeast display and protease stability assay

DNA amplification, yeast display proteolysis, sorting, and next-generation sequencing were all performed by research contract to the University of Washington BioFab [118] according to the protocol of [37]. Yeast display was performed using a display vector with improved protease resistance.[96]

Computing stability scores

We calculated a "stability score" for each design based on a probabilistic model described previously.[37] The model determines the EC_{50} (the protease concentration at which half of the yeast cells pass selection during flow cytometry) for each design. The difference between the experimental EC_{50} in the folded state and predicted EC_{50} in the unfolded state (based on the identical model from [37]) is what we call a "stability score." The overall stability score for each sequence is the minimum of the independent stability scores measured by trypsin and chymotrypsin. As previously,[37] data were filtered based on the confidence interval of the EC_{50} estimate: only sequences where the 95% confidence interval was smaller than 2.0 (meaning the equivalent of two selection rounds, or 9x protease concentration) were retained for analysis. However, the mutational scanning data were not filtered based on the EC_{50} confidence interval; however, only 6/4650 (0.1%) sequences had low confidence stability estimates.

Computing metrics and Regression modeling

The Rosetta models used for computing structural features were the lowest energy structures from at least 1,000 ab initio trajectories and 200 relax trajectories starting from the design models. Rosetta design models were scored using the Rosetta score function, and we computed structural and biophysical features pertaining to secondary structure, dipeptides, hydrophobicity, hydrogen-bonding, and fragment quality using the score_monomeric_designs package (https://github.com/Haddox/score_monomeric_designs).

For regression modeling, we performed linear regression by bootstrapping (sampling with replacement 1000 times) using Python scikit learn [115] and selected the 95% confidence interval for each variable's coefficient for analysis. For the restricted design strategy, we first used stepwise linear regression to identify eight features (large nonpolar count, nonpolar residue-residue contacts, local sequence-structure propensity, Ser/Thr at helix caps, Glu-Arg residue-residue contacts, nonpolar residue at design ends (which we define as positions 1, 2, 42, and 43 of the 43-residue-long protein structure), Glu-Glu residue-residue contacts, and increased net charge) that increased the correlation coefficient between predicted and experienced stability scores. We also selected two features (favorable net charge at helix ends and buried unsatisfied polar atoms) to determine their relative contributions to stability. For the diversity-oriented topology-focused linear regression model, we selected the seven most common loop patterns and the three most common β-strand lengths found in our dataset as inputs to a linear regression model.

Calculation of local sequence-structure agreement

The compatibility of each protein sequence with its local backbone structure (Fig. 2E) was computed using the abego_res_profile method from.[37]

Protein expression and purification

We purchased six αββα designs whose nucleotide sequences were optimized for E. coli expression and encoded in the pET-28a(+) vector (that contains an N-terminal His-tag and thrombin cleavage) from Twist Bioscience. The plasmid vectors were transformed in BL21(DE3) competent cells (Invitrogen or Sigma Aldrich) and grown overnight in a starter culture of 50 mL LB media (Fisher Bioreagents) and 50 µg/mL kanamycin at 37°C while shaking at 225 rpm. 16-18 hrs later, we inoculated 500 mL of LB media and 50 µg/mL kanamycin with 10 mL of the starter culture and allowed the competent cells to grow until OD_{600} ~0.6-0.8.

In preparation for NMR analysis, we transformed one αββα design encoded in pET-28a(+) into BL21(DE3) competent cells (Sigma Aldrich) and grown in an LB media starter culture (as stated above). After 16-18 hrs, we pelleted the cells by centrifugation, replaced the LB media with M9 media (40 mM Na_2HPO4, 8.5 mM NaCl, 20 mM KH_2PO_4, 60 mM d-Biotin, 55 mM Thiamine, 0.1 mM $CaCl_2$, 0.01 mM $ZnSO_4$, 2 mM $MgSO_4$, 50 ug/mL kanamycin) that included 15 mM $^{15}NH_4Cl$ and 10 mM ^{13}C glucose

(Cambridge Isotopes) and resuspended the pellet. We then inoculated 500 mL of LB media with M9 media (including 15 mM $^{15}NH_4Cl$, 10 mM ^{13}C glucose and 50 µg/mL kanamycin) with 10 mL of the resuspended starter culture and allowed the competent cells to grow until OD_{600} ~0.6-0.8.

Afterwards, for both labeled labeled and unlabeled competent cells, we induced protein expression by adding a final concentration of 500 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) (Fisher Bioreagents) to the LB media and allowing the cells to grow overnight at 15°C while shaking at 225 rpm. We then harvested the cells by centrifugation at 4°C and lysed the cells in 30 mL of lysis buffer (20 mM Tris, 500 mM NaCl, 30mM imidazole, 0.25% CHAPS, 1mM PMSF, pH 8.0), which included 60 mg of chicken lysozyme (Sigma), 1.5 µL of benzonase nuclease (Sigma Millipore), and 1 tablet of Pierce protease Inhibitor EDTA-free (ThermoFisher) followed by sonication (QSonica SL-18).

Next, we separated insoluble bacterial material by centrifugation (10,000 x g for 30 min) and purified the αββα miniproteins by immobilized metal-affinity chromatography (IMAC), which involved transferring the supernatant onto Econo-pac columns (Bio-Rad) that were previously prepared with Ni-NTA (Qiagen), washing the column with 15 mL Wash Buffer (20 mM Tris, 500 mM NaCl, 30 mM imidazole, 0.25% CHAPS, 5% glycerol, pH 8.0), and eluting the samples in 10 mL of Elution Buffer (20 mM Tris, 300 mM NaCl, 500 mM imidazole, 5% glycerol, pH 8.0). We initially verified the size and purification of the miniproteins by SDS-PAGE electrophoresis and Coomassie stain gel analysis. Then, we concentrated them by a centrifugal filtration system (Amicon Ultracel-15 or Amicon Ultra-0.5).

We further purified both labeled and unlabeled miniproteins by size-exclusion chromatography (Bio-Rad NGC Chromatography System) using a Superdex 75 10/300 GL column (GE Healthcare) and eluted in PBS buffer. A mixture of proteins with known molecular weights (BSA, Ovalbumin, Ribonuclease A, Aprotinin, and Vitamin B12) were also separated by SEC to determine a calibration curve, which was used to designate expected miniprotein monomeric and dimeric fractions (see Fig. 3B). Miniprotein size and purification were verified first by Coomassie stain gel analysis and then by mass spectrometry (Synapt G2 Si, Waters).

Circular dichroism

Far-ultraviolet circular dichroism measurements were performed on six αββα designs (HEEH_rd5_0018, HEEH_rd5_0341, HEEH_rd5_0420, HEEH_rd5_0614, HEEH_rd5_0958, and HEEH_rd5_3711) using a Jasco J-815 spectrophotometer. All analysis was performed on the unmodified expression constructs including a 21-residue N-terminal linker (MGSSHHHHHHSSGLVPRGSHM). We measured the concentration samples with a Qubit 4 Fluorometer (Invitrogen) and diluted them to a final concentration of ~0.1-0.4 mg/mL in PBS buffer. Wavelength scan measurements were made using a 1 mm path-length cuvette from 195 to 260 nm at 25°C and 95°C. We also measured temperature melts at 220 nm for every 1°C from 25°C to 95°C. For temperature melt analysis, we smoothed the data with a Savitsky-Golay filter of polyorder = 3.

Nuclear Magnetic Resonance

NMR spectra for HEEH_TK_rd5_0341 and HEEH_TK_rd5_0958 structure calculations were acquired at 288 K, on Bruker spectrometers operating at 600 or 800 MHz, equipped with TCI cryoprobes with the protein buffered in 20 mM sodium phosphate (pH 7.5, 150 mM NaCl) at concentrations of ~ 0.5 to 1.0 mM. Resonance assignments were determined for 15N/13C-labeled protein using FMCGUI [119] based on a standard suite of 3D triple and double-resonance NMR experiments collected as described previously.[120] All 3D spectra were acquired with non-uniform sampling in the indirect dimensions and were reconstructed by the multi-dimensional decomposition software qMDD,[121] interfaced with NMRPipe.[122] Peak picking was performed manually using NMRFAM-Sparky.[123] Torsion angle restraints were derived from TALOS+.[124] Automated NOE assignments and structure calculations were performed using CYANA 2.1.[125] However, for HEEH_TK_rd5_0341 NOEs identified manually with high confidence, that were not consistent with a monomeric structure, were added as initial restraints for dimeric structure calculation. The best 20 of 100 CYANA-calculated structures were refined with CNSSOLVE[126] by performing a short restrained molecular dynamics simulation in explicit solvent.[127] The final 20 refined structures comprise the NMR ensemble. Structure quality scores were performed using Procheck [128] and the PSVS server.[129] Longitudinal (T1) and transverse (T2) ^{15}N relaxation rates were determined for the constructs at 288 K (at 600 MHz)[130] using the integrated signal from the structured amide regions (>8.5 ppm) of 1D ^{15}N-edited spectra, fitted to an exponential decay as a function of delay times. Rotational correlation times (τc) were estimated based on T1-T2 ratios[131] and hydrodynamic radii (rH) calculated using the Stokes-Einstein equation.

Hydrogen-deuterium exchange and analysis

NH/D amide exchange rates were determined for HEEH_TK_rd5_0341 and HEEH_TK_rd5_0958 (including the 21aa N-terminal linker MGSSHHHHHHSSGLVPRGSHM) by performing an exchange series at 288 K (at 600 MHz), by monitoring the decay rate of amide peak intensities in 1H-15N HSQC spectra collected over the course of 24 hrs. HEEH_TK_rd5_0958 dissolved in MES buffer at ~ 500 μM (pH 5.8, 150 mM NaCl) was lyophilized, and exchange was initiated by solubilizing in an equal volume of D_2O; each HSQC time point was acquired in ~ 5 minutes. For HEEH_TK_rd5_0341, exchange was initiated by mixing phosphate buffered protein (at ~ 1.0 mM, pH 7.5) with D_2O at a ratio of 1:19, and each HSQC time point was acquired in ~ 16 minutes. The first time points for the series were started ~ 5 minutes following the initiation of exchange. Peak intensities were fitted to a single exponential decay (with an offset due to the presence of residual 5% H_2O for HEEH_TK_rd5_0341). Opening free energies were calculated from these rates as previously described.[114,132,133] For residues where exchange was too fast to quantify, we calculated an "upper limit" ΔG_{open} based on an exchange rate of 0.1 min^{-1} (the fastest quantifiable rate was 0.066 min^{-1}).

AlphaFold analysis

All αββα miniprotein structures were predicted from their primary sequences using AlphaFold 2[93] without using multiple sequence alignment (MSA) information because αββα miniproteins have low similarity to natural proteins. Five models were generated for each sequence and the lowest RMSD model to the designed structure was used for the analysis in Figs. S13, S17, and S18.

Analysis of residue composition difference

The residue composition of all Round 5 and 6 were compared by taking all possible combinations of protein pairs, and first identifying the total number of unique residues for each protein. Then, we took the absolute difference for each unique residue, taking the sum of all absolute differences, and then dividing the sum by the total number of residues being compared. Other protein domains were compared, LysM (https://pfam.xfam.org/family/LysM), PASTA (https://pfam.xfam.org/family/PF03793), and Cold Shock (https://pfam.xfam.org/family/CSD).

Analysis of structural agreement among Rosetta design, AlphaFold, and NMR models

Different models were aligned to each other using PyMOL. To analyze the RMSD between the Rosetta design model and AlphaFold model, both structures were aligned (super structure1, structure & name n+c+ca+o). To analyze the RMSD between the Rosetta design model (or AlphaFold model) with the NMR ensemble, all twenty states of the NMR ensemble were aligned to each other, and the design model (or AlphaFold model) was added to the NMR pdb file as the 21st state. Then, RMSD values for all NMR states to the design model (or AlphaFold model) were determined (intra_fit structure////n+c+ca+o, 21), and the average RMSD determined (Fig 3F, H).

CHAPTER 3: Towards developing a reporter cell to identify cell-penetrating miniproteins

3.1 Abstract

Cell-penetrating miniproteins (CPMPs) could become a new class of drugs as they would be better suited than small molecule drugs to target protein-protein interactions (PPIs). However, many PPIs occur in the cell cytosol, and proteins typically are unable to cross cell plasma and endosomal membranes. Cell-penetrating peptides (CPPs) have been well studied, but their small size, susceptibility for degradation, and low-efficient delivery make them non-ideal therapeutic candidates. Cell-penetrating proteins can address the limitations of CPPs, but only a few have been characterized. Thus, this limits our understanding of the biophysical features that enable cell-penetration. High-throughput screens and large quantitative data would enable us to formulate such general biophysical determinants, but no high-throughput assays currently exist.

Here, we present initial experiments that can guide future tests to successfully develop a high-throughput screen. We use the split-intein, VidaL, and the synthetic transcription factor, ZF1-VP64, to investigate whether the presence of a test miniprotein can be detected by a “reporter” mammalian cell. Our results by fluorescent microscopy suggest that ZF1-VP64 can successfully activate the gene expression of a fluorescent protein in the presence of the test miniprotein.

3.2 Introduction

There is an estimated 645,000 PPIs that occur in human cells.[134] Aberrations to these interactions due to mutations can lead to a wide range of diseases, such as cancer,[135,136] neurodegeneration,[137] cardiovascular disease,[138] and various metabolic diseases.[139]

Therapeutically targeting cytosolic PPIs has currently been challenging because small molecule drugs, which is currently the most common modality, are most effective when there is a deep binding pocket. However, many PPIs interact at shallow interfaces.[140] Protein-based therapeutics offer a promising alternative, but proteins readily cannot cross cell membranes, escape endosomes, and enter the cytosol because of the biological nature of proteins and membranes: proteins have hydrophilic surfaces, but the lipid bilayer of cell membranes consist of hydrophobic carbon tails.

Only a few proteins that can penetrate cell membranes have been characterized,[58–60] but this paucity makes it extremely difficult to generalize principles for cell-penetration and use that knowledge towards designing cell-penetrating proteins for drug targets. High-throughput screening for candidate cell-penetrating proteins could help address this gap in knowledge, but such methods currently do not exist.

Here, we propose a novel system, namely a “reporter” cell that drives the gene expression of a fluorescent protein when a protein enters the cytosol. The key reaction driving this readout is the release of a transcription factor mediated by split inteins, which are protein fragments that spontaneously ligate their flanking regions (exteins).[74]

Gene regulation and gene circuit design using synthetic transcription factors have been well studied [141–146] and applied to cell systems that respond to extracellular ligands.[147–152] For example, a recent study split the transcription factor, ZF1-VP64, into its two components (a zinc finger DNA-binding domain and VP64 activation domain), fused them to gp41-1 intein fragments, and demonstrated that an intein-mediated reconstitution of the transcription factor can lead to downstream gene activation (this was also possible by tethering the ZF1 domain onto a transmembrane domain, TMD).[75]

Here, we modify this system to test whether a 43-residue miniprotein fused to one intein can trigger a trans-splicing intein reaction with its complementary intein that is fused to a TMD and an unsplit ZF1-VP64. Using gp41-1 as our intein-of-choice seemed to be suboptimal because both inteins fragments are relatively large (88 residues for the N-terminal intein and 37 residues for the C-terminal intein),[75] and their use could result in complications with data interpretation in a future screen, especially as the intein is as large or larger than the protein-of-study.

Instead, we opted to test the split-intein, VidaL, because its N-terminal fragment is only 16 residues,[76] which is comparable to split-GFP complementation assays that tag CPPs with the 11th-strand of GFP (15 residues).[70,71] We also selected VidaL because of its fast splicing kinetics (reaction half-life of ~1 min) *in vitro*.[76]

Here, we show preliminary results suggesting it is possible to combine synthetic transcription factors with intein-splicing to report the presence of a protein-of-interest in the cytosol.

3.3 Design of constructs and expected products

To investigate an intein-mediated release of ZF1-VP64 and subsequent gene expression of a fluorescent protein, we initially designed and tested two components in

HEK 293T cells (Fig. 6, Table S6). First, we fused a test miniprotein, EHEE_rd2_0005, to the N-terminal VidaL intein (VidN), hereafter called EHEE-VidN. We selected this protein for its high folding stability.[37] Second, we replaced the C-terminal intein in the target chain of an engineered receptor (called MESA)[75,153] with the C-terminal VidaL intein (VidC). MESA was originally designed to sense extracellular ligands upon the dimerization of its target chain and protease chain,[148,150] but we repurposed it by only using the target chain to serve as a scaffold for sequestering ZF1-VP64 away from the nucleus. At the N-terminus of VidC is a TMD, and at the C-terminus of VidC is the transcription factor, ZF1-VP64, followed by a nuclear localization signal (NLS), hereafter called TMD-VidC-TF (Fig. 6, Table S6).

The predicted mechanism is the presence of EHEE-VidN in the cytosol will lead to a spontaneous and rapid intein trans-splicing reaction with VidC, which is tethered to the TMD and linked to ZF1-VP64. The reconstituted inteins are then linked to the TMD, and their exteins (EHEE_rd2_0005 and ZF1-VP64) combine and enter the nucleus. Then, gene transcription of the fluorescent protein, mKate2, will initiate (Fig. 6).

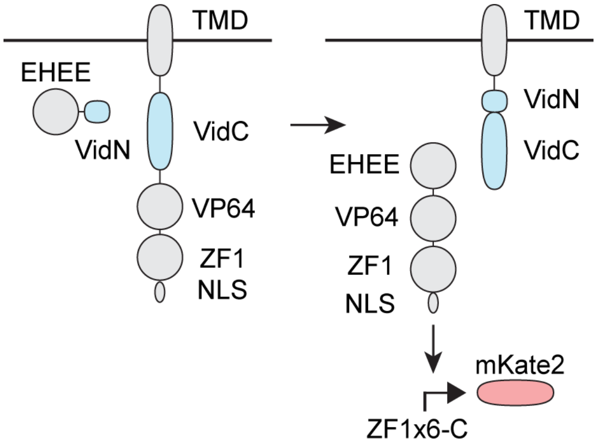


Fig. 6. Mechanism for gene activation based on ZF1-VP64 and VidaL. The miniprotein EHEE_rd2_0005 miniprotein (EHEE) fused to a 16-residue N-terminal intein (VidN) undergoes a trans-splicing reaction with its corresponding C-terminal intein (VidC). VidC is tethered to a transmembrane domain (TMD) at its N-terminus and to a transcription factor (VP64 activation domain and ZF1 DNA-binding domain) and a nuclear localization sequence (NLS) at its C-terminus. Once intein trans-splicing has occurred, the transcription factor is released and enters the nucleus where it binds to six ZF1 DNA-binding sites (ZF1x6-C) and activates the gene transcription of mKate2.

3.4 Testing intein splicing in mammalian cells

In HEK 293T cells, we co-transfected equal amounts of plasmids (by mass) encoding EHEE-VidN, TMD-VidC-TF, as well as mKate2 driven by a ZF1x6-C promoter (Fig. 7A). To quickly observe our results, we chose a qualitative approach by observing the activation of mKate2 by fluorescence microscopy (Fig. 7B). When both EHEE-VidN and TMD-VidC-TF were transfected, we observed reporter gene activation, initially suggesting to us that mKate2 expression may be due to intein-splicing and subsequent release of ZF1-VP64 (Fig. 7, bottom row). However, in the control experiment in which we transfected TMD-VidC-TF without its complementary intein, EHEE-VidN, we also observed fluorescence (Fig. 7, middle row).

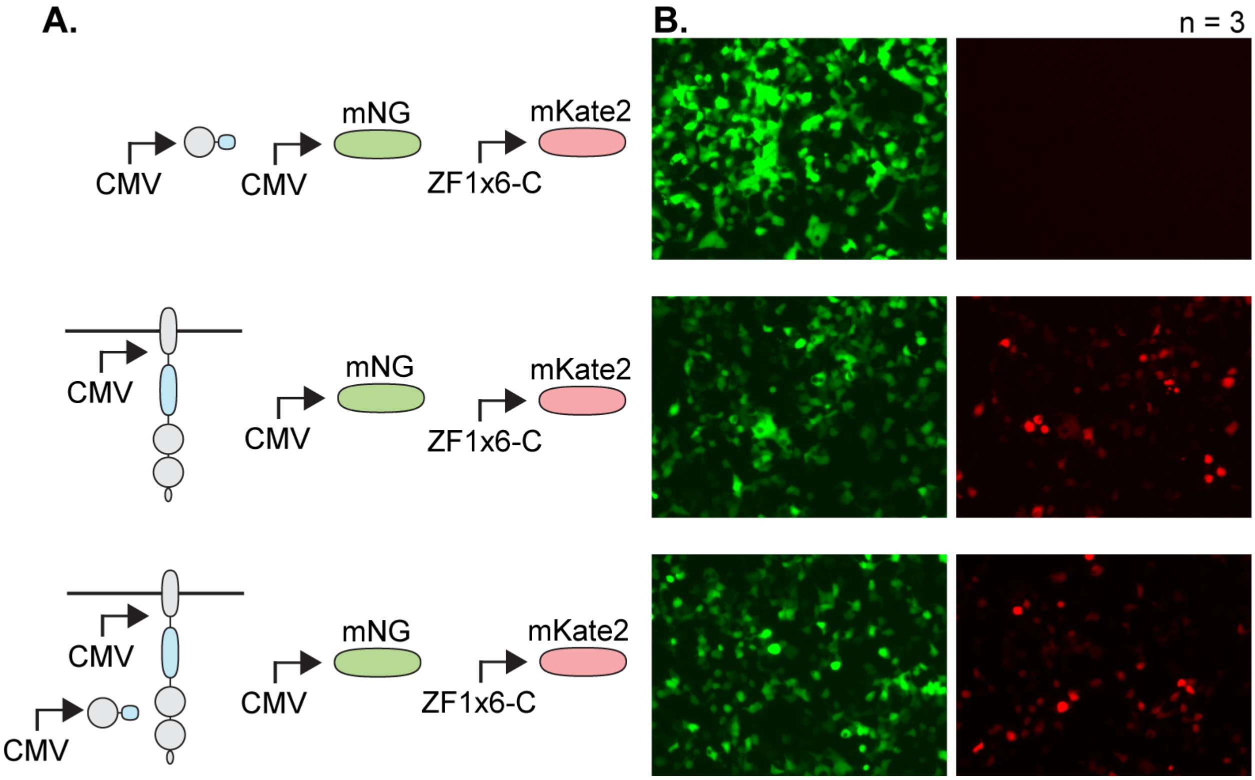


Fig 7. Reporter gene activation via intein constructs driven by a CMV promoter. (A) Cartoons of the gene constructs and the promoters that drive them that were transfected in HEK 293T cells: EHEE-VidN, TMD-VidC-TF, mNG, and mKate2. mNG (mNeonGreen) serves as a transfection control, and mKate2 is the reporter gene. (B) Fluorescent microscopy images are representative of three experiments at 20x resolution.

This suggests that ZF1-VP64 entered the nucleus to activate gene transcription of the reporter gene perhaps because either the transcription factor has self-spliced itself off the TMD, or the entire construct did not successfully get embedded into the cell plasma membrane. To check if self-splicing may be possible by VidC, we first identified the residue that VidC normally reacts with in VidN (Cys1) as well as the three extein residues immediately flanking Cys1 (Glu-Ser-Gly) (Table S6) because exteins have been shown to impact splicing kinetics.[76,154] We looked to see if TMD-VidC-TF had any Glu-Ser-Gly-Cys sequences that VidC could recognize and splice, but we did not find any. This suggested to us that VidC might not trigger a self-splicing reaction.

An alternative hypothesis is that over-expression of TMD-VidC-TF burdens the cell machinery that is responsible for trafficking membrane-bound proteins to the cell plasma membrane. Type 1 transmembrane proteins with signal peptides at their N-terminus are recognized by a signal recognition particle (SRP) as the nascent polypeptide chain is being translated from the ribosome. The SRP targets the protein to the protein channel Sec61 in the endoplasmic reticulum (ER) where the translating transmembrane protein gets translocated through the ER's lipid bilayer.[155] The plasmid that encodes TMD-VidC-TF has a CMV promoter and enhancer, β-globin intron, and a woodchuck hepatitis virus posttranscriptional regulatory element (WPRE), all of which enhance gene expression.[156,157] Thus, it may be possible that over-expression of TMD-VidC-TF leads to a subpopulation of this construct being processed by the SRP in a timely manner, and another subpopulation waiting in the cytosol to be trafficked into the ER. Because the TMD-VidC-TF has an NLS at the C-terminus, it may be possible for the NLS to be recognized by an import receptor [158] that trafficks the TMD-VidC-TF through the nuclear pore complex.

3.5 Testing intein splicing using a weaker promoter or masking the NLS

We therefore tested whether reducing the expression of the TMD-VidC-TF using a weaker promoter or removing the gene enhancing elements (β-globin intron and WPRE) could greatly reduce, if not eliminate, the TMD-VidC-TF from activating the reporter gene by itself. The EF1α and SV40 promoters have been correlated with less protein expression than CMV,[159] so we tested the first two promoters and their effects in reducing mKate2 signal (Fig. 8). We still observed by fluorescence microscopy activation of the reporter gene when the TMD-VidC-TF was transfected in HEK 293T cells without EHEE-VidN. However, mKate2 signal was qualitatively higher when both TMD-VidC-TF and EHEE-VidN were co-transfected (Fig. 8).

Although it might be possible to tolerate the background signal and use TMD-VidC-TF driven by either the EF1α or SV40 promoter when building a “reporter” cell, we wondered if we could improve the difference between true signal (mKate2 expression by both TMD-VidC-TF and EHEE-VidN) and background signal (mKate2 expression by TMD-VidC-TF alone).

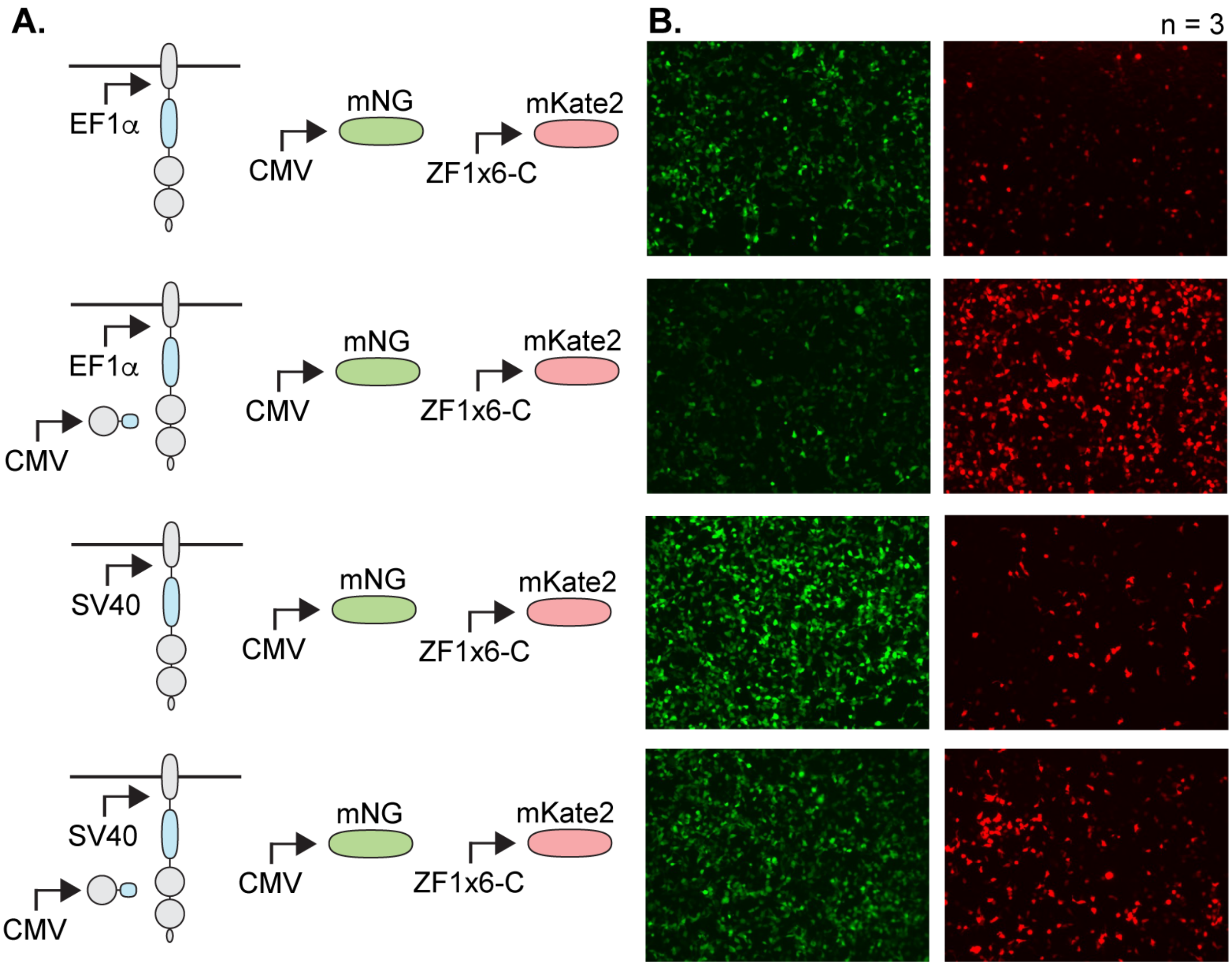


Fig. 8. Reporter gene activation from an EF1α or SV40 promoter. (A) Cartoons of the gene constructs and the promoters (EF1α and SV40) that drive them that were transfected for each of the three conditions. mNG serves as a transfection control, and mKate2 is the reporter gene. (B) Fluorescent microscopy images are representative of three experiments at 10x resolution.

We then investigated whether positioning the NLS in the middle of TMD-VidC-TF instead of at the C-terminus might “mask” the NLS from import receptors. Several studies have investigated how the NLS of a protein can be masked by the influence of its flanking regions,[160] by a binding event with another protein [161–163] or due to post-translational modifications.[164,165] AlphaFold 2 structure prediction of TMD-VidC-TF if the NLS was inserted between VidC and VP64 of the transcription factor suggest that the first lysine in the NLS interacts with the backbone of Thr33 in VidC (Fig. 9, red box). It could be that this interaction as well as the close proximity of the NLS to VidC are sufficient for the NLS to be inaccessible by an import receptor. Surprisingly, AlphaFold 2 was unable to accurately predict the structure of VP64 (Fig. S19), but it may be possible that both VP64, if predicted accurately, could also play a role in masking the NLS.

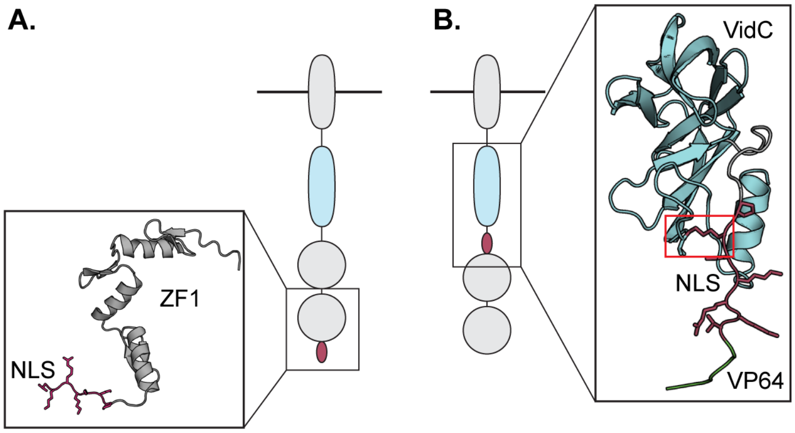


Fig. 9. Predicted structures of domains near NLS. (A) Cartoon of the original TMD-VidC-TF with the NLS (red) at the C-terminus and (B) TMD-VidC-TF with the NLS (red) shown in the middle of the structure. Inside each rectangle shows the best AlphaFold 2 predicted-structure for a specific region of the TMD-VidC-TF. Inside the rectangle in panel B shows a red rectangle, which highlights the lysine in the NLS that is predicted to form an interaction with the backbone of VidC.

To test this hypothesis, we generated a new TMD-VidC-TF construct in which the NLS was moved from the C-terminus to the middle of the construct (between VidC and VP64) (Fig. 9B). When this new construct was transfected in HEK 293T cells with or without EHEE-VidN, we observed by fluorescence microscopy that the mKate2 signal was higher when both constructs were expressed than just TMD-VidC-TF alone (Fig. 10).

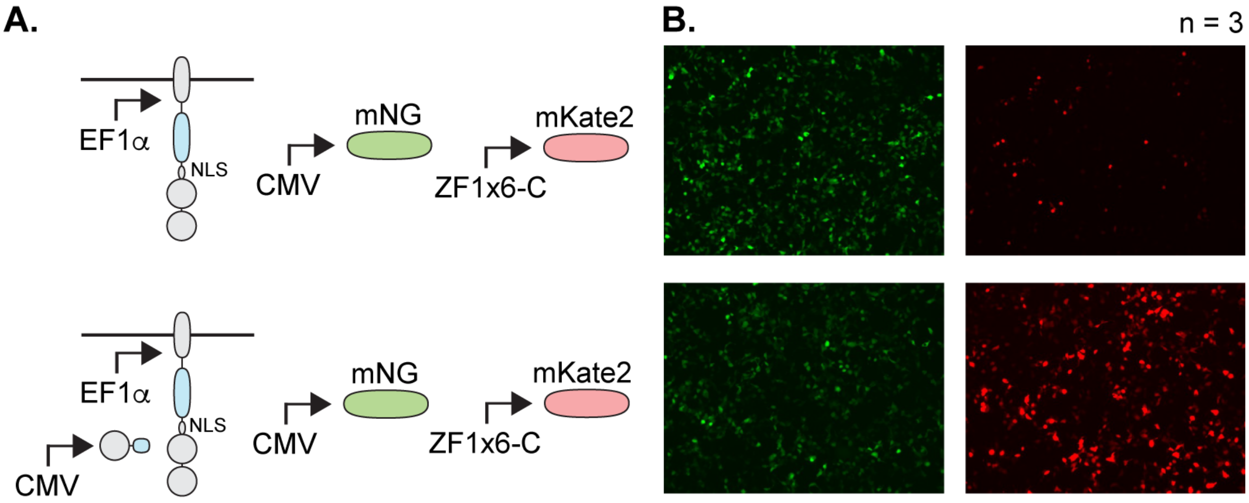


Fig. 10. Impact of masked NLS on reporter gene activation. (A) Cartoons of the gene constructs and the promoters that drive them that were transfected for each of the three conditions. mNG serves as a transfection control, and mKate2 is the reporter gene. (B) Fluorescent microscopy images are representative of three experiments at 10x resolution.

This observation was similar to the previous experiment in which we tested two different promoters (Fig. 8). Indeed, when we quantified the fluorescence intensity across both experiments, the true signal was always higher than the background signal. However, the difference between true and background signal is higher when the promoter is EF1α, although the true signal is more consistent when the NLS is masked (across n=3 experiments) (Fig. 11). These preliminary results indicate that a “reporter” cell could be built using either the TMD-VidC-TF driven under EF1α or EF1α with a masked NLS.

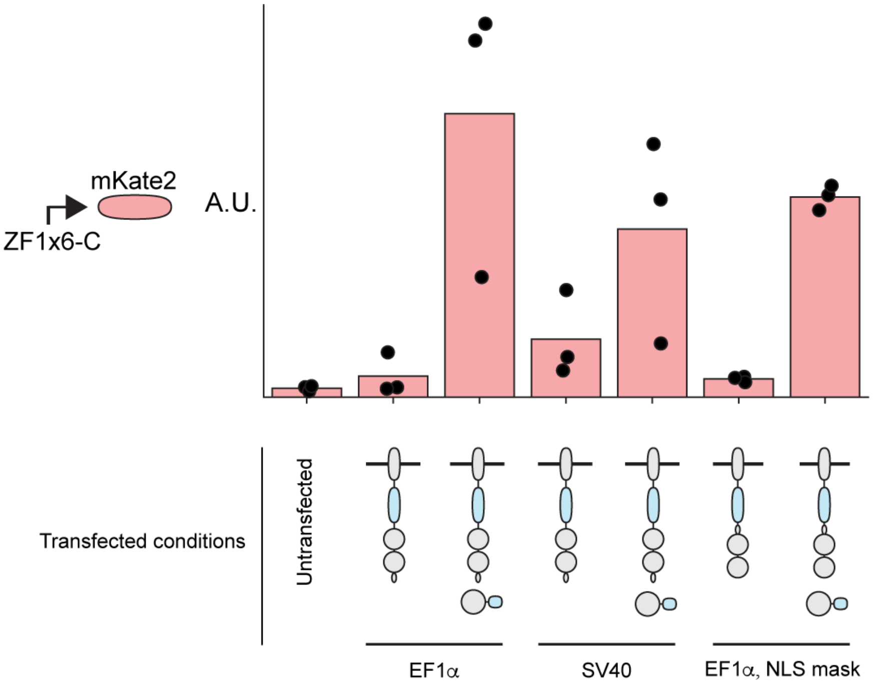


Fig. 11. Fluorescence intensity of reporter gene activation. Barplot showing mean normalized mKate2 signals for six conditions transfected in HEK 293T cells: TMD-VidC-TF without and with EHEE_VidN driven under the EF1α promoter, driven under the SV40p promoter, and driven under the EF1α promoter with the NLS masked. Black dots denote individual experiments (n = 3).

3.6 Methods and Materials

Mammalian Cell culture

HEK 293T cells (ATCC) were passaged using Trypsin (Gibco) and cultured using DMEM (Thermo Fisher) using 10% FBS (Corning) supplemented with 1% Penicillin and Streptomycin (Gibco). Cells were incubated at 37°C, 5% CO_2.

Gene synthesis and molecular cloning

Genes encoding EHEE_rd2_0005-FLAG-VidN and CD4-FLAG-FRB-FGFR4-VidC-VP64-ZF1-NLS were synthesized by Twist Bioscience as clonal genes in their stock pTwist vectors (promoters CMV and EF1α) (Table S6). ZF1x6-C-mKate2 was a gift from Josh Leonard. All three genes were transformed into DH5α competent cells (Sigma Aldrich), and plasmids were generated by mini-prep (Takara).

To generate the VidC construct (above) with an SV40 promoter, the gene that was synthesized from Twist and the vector pRL-SV40p (Addgene #27163) were linearized by PCR using custom primers (IDT) (Table S7). KOD Polymerase was used for PCR (Thermo Fisher). PCR products were purified (Qiagen), and ligated using the In-Fusion Cloning Kit (Takara). Ligated products were transformed into DH5α competent cells (Sigma Aldrich), selected by ampicillin, and the sequences confirmed by Sanger sequencing (ACGT Inc.).

Transient transfection

HEK 293T cells (ATCC) were plated in 96 well plates (Falcon) to a final confluency of 80-90% in DMEM (Thermo Fisher) + 10% FBS without antibiotics. We used the manufacturer's instructions for transfection using Lipofectamine 3000 (Thermo Fisher). pCAGEN-mNeonGreen was used as transfection control, and empty pCAGEN or empty pcDNA3.1 was used in order to ensure that all cells received the same total mass of plasmids. An equal ratio of plasmids were transfected. Cells were incubated in 37°C, 5% CO_2 and analyzed 24 hours after by microscopy.

Fluorescent microscopy and fluorescence plate reading

Microscopy images were taken using the Evos M5000 Imaging System (Thermo Fisher) at 10x or 20x resolution. Quantifying fluorescence intensity was done using the Synergy Neo2 multi-mode reader (Biotek). We measured green fluorescence using 487 nm for excitation and 528 nm for emission, and red fluorescence using 588 nm for excitation and 633 nm for emission. The gain was set to automatic.

To calculate the normalized fluorescence signal (Fig. 11), we first calculated the ratio of mKate2 and mNeonGreen signal intensities for each well. Nextt, we calculated the average ratio of mKate2/mNeonGreen for all wells. Then, we normalized the mKate2 signal intensity by dividing each mKate2 signal from each well by the average mKate2/mNeonGreen ratio.

AlphaFold2 structure prediction

Structures for TMD-VidC-TF-NLS and TMD-VidC-NLS-TF were predicted using ColabFold v1.5.2.[166] The sequences used as the input did not include the signal peptide, FLAG tag, the extracellular FRB domain, or the TMD as all of these would not interact with the intein, transcription factor, or NLS. mmseqs2_uniref_env and unpaired_paired mode were selected for MSA options, and the number of recycles was 6. Structure images were analyzed and visualized using PyMOL.

CHAPTER 4: Future Directions

This chapter highlights major themes and potential future directions for the αββα folding stability and reporter cell projects.

4.1 Further analysis for miniprotein folding stability

We used sequence-based and structural metrics as well as folding stability data to build an interpretable biophysical model that quantifies how ten determinants contribute to an αββα miniprotein's folding stability (Chapter 2, Fig. 2I). Based on this model, we found that 80% of stable αββα miniproteins in round 6 (which had not been used to build the model) had these determinants. Because we did not actually make αββα designs using our biophysical model, a follow-up set of experiments could be to create a seventh round of αββα miniproteins whose designs are fully guided by the model. If this new set of αββα miniproteins achieves a success rate greater than previous rounds, this would validate the strength of this simple model.

Multimerization

However, the model is not perfect. Biophysical characterization of several αββα miniproteins in round 5 of design revealed that two structures formed dimers (by SEC), one of which was also confirmed by NMR (Chapter 2, Fig. 3B). But, it might be possible that there are other αββα designs that are stable by forming dimers. So, it might be useful to use AlphaFold-Multimer[167] to predict the likelihood of designs forming higher ordered species, excluding multimer-forming αββα miniproteins from the dataset, and re-training the biophysical model.

However, if a considerable number of αββα designs have a likelihood of forming higher-ordered species, a new direction for the lab could be using dimeric αββα as a model system to quantify the determinants for miniprotein multimerization between the same design or among different designs. This has implications for biomedical research as many proteins exist as assemblies,[168–170] and disruption can lead to disease.[171–173]

A more generalizable model

Another approach to improve the model is to make it more generalizable to the αββα topology, or to miniproteins in general. Our analysis of the determinants of folding stability was based on a specific context – αββα miniproteins made from a specific computational approach. For round 5, designs were based on a particular protein architecture (14-residue helices and 3-residue β-strands), and round 6 designs were built off of fifteen protein architectures (with 9-12 residue helices and 3-5 β-strands). Because we limited this study to a certain region of sequence and structure space, it may be informative to design new libraries of αββα miniproteins with even greater sequence and structural diversity. Then, we may learn additional determinants for folding stability if we test 43-residue αββα designs with a wider range of secondary structures (e.g. shorter and longer helices and β-strands than what we tested here). Furthermore, given recent advances in DNA synthesis like DropSynth[174,175] that enables the assembly of larger DNA oligo libraries, we could design and test the folding stability of αββα miniproteins that are larger than 43 residues.

To make the model more generalizable to proteins beyond the αββα fold, another direction to further this study is to build an interpretable model based on folding stability data across many different folds. Previous work[37] on which Chapter 2 is based, tested the folding stabilities of four folds (ααα, αββα, βαββ, and ββαββ) and could be used as an initial dataset for building a more generalizable model.

Folding stability under different environmental conditions

In vitro studies in the lab are commonly performed under standard laboratory conditions (e.g. cell-culture based experiments at 37°C and 5% CO_2; wet bench-based experiments at 25°C). However, varying environmental conditions, such as pH, pressure, temperature, and solvent [176–178] in laboratory experiments could model different aspects of human physiology (e.g. pH in gastrointestinal tract, pressure in arteries, solvent in blood vs. stomach). Testing miniprotein libraries under different environmental conditions by modifying the protease stability assay based on yeast-display[37] or cDNA-display[179] could reveal additional determinants for folding stability.

Integrating deep learning approaches

At the time of preparing Chapter 2 for publication and later writing this dissertation, the application of deep learning methods to protein structure prediction and design rapidly advanced. AlphaFold2 was able to predict the structures of proteins with high accuracy (median backbone RMSD 0.96 Å)[93] and has been used to predict the structures of over 98% of the human proteome.[180] Other methods like RoseTTAFold,[181] ESMFold,[182] and OmegaFold[183] have been shown to perform as well as AlphaFold 2. In addition, other deep learning algorithms like ProteinMPNN[184] can generate sequences given a backbone, and *de novo* backbones can now be made using RFDiffusion.[185]

Given these successes, what might be the role of physics-based energy functions in protein design now? After all, a workflow that combines RFDiffusion (generate backbone), ProteinMPNN (generate sequences onto backbone), and experimental characterization (compare design with ground truth) could bypass energy function calculations. Recent work has shown that this approach can lead to successful peptide binder designs,[186] and be computationally more efficient than using Rosetta.[187]

However, newer deep learning-based methods also encounter the same challenge as conventional physics-based methods in protein design because not all designs fold as designed. For example, 50/96 (52%) of α-β proteins designed using MPNN[184] and 70/608 (12%) designs of various topologies made by RFDiffusion[185] were soluble by *E. coli* expression and eluted at the target monomeric or oligomeric state by SEC. In addition, when we predicted the structures of our αββα designs using AlphaFold 2, 78% of the restricted-oriented designs and only 20% of the diversity-oriented designs had a predicted structure < 2 Å RMSD to the design structure. Moreover, AlphaFold 2 was unable to discriminate between stable, moderately stable, and unstable designs (Appendix, Fig. S17).

To address this limitation, high-throughput experimental data on protein folding stability could serve as useful training data to develop newer, more accurate deep learning models. Many natural proteins exist in a delicate equilibrium between folded and unfolded states[188] with a single amino acid substitution being sufficient enough to destabilize or further stabilize a protein.[86,179,189] Deep mutational scanning analysis of single amino acid mutations also revealed both destabilizing and stabilizing protein structures, many of which differ from the wildtype structure by only < 1.5 Å RMSD (Chapter 2, Appendix Fig. S13). If the difference between a stable and unstable protein could be a single amino acid change without significantly altering the structure, it may be that current deep learning models are missing something fundamental about how proteins remain folded. Large-scale protein folding stability data based on protease resistance using yeast display[37] or cDNA display[179] could provide useful training data for more improved deep learning methods that could be used to generate stable protein structures with even greater success.

4.2 Towards building a reporter cell to identify cell-penetrating miniproteins

Further tests on TMD-VidC-TF

Most of the observations testing mKate2 expression based on an intein-mediated release of a synthetic transcription factor were based on qualitatively analyzing fluorescence microscopy images. A semi-quantitative approach to validate our observations could involve Western blotting by separating cytosolic, membrane, and nuclear fractions and blotting for TMD-VidC-TF and EHEE-VidN. This could reveal the location of the constructs (or subpopulations of the constructs) in the cell, the size of the spliced components, and how much of the spliced products are made. Western blots of harvested cells at different time points could reveal splicing kinetics.[76] In addition, flow cytometry of transfected cells could provide a quantitative look on the effect of intein-splicing.

This analysis also investigated whether the promoter (CMV, EF1α, or SV40) or the location of the NLS in TMD-VidC-TF (at the C-terminus or in the middle of the construct) makes a difference in the gene expression of mKate2 (Chapter 3). However, we did not test a masked NLS construct driven by an SV40 promoter or another known weak promoter, ubiquitin-C (Ub).[159] Investigating whether TMD-VidC-TF expression under the Ub promoter, or whether a masked NLS construct driven under either SV40 or Ub can lead to extremely little background signal or a higher true-signal : background-signal ratio than what we have observed (Chapter 3, Fig. 11) could guide us in building a more robust reporter cell.

Test the S11 split-intein

There are only a few studies characterizing the split-intein, S11.[190–193] Unlike the fast kinetics of VidaL (~1x10^{-2} s^{-1}),[76] S11 splicing and reconstitution occurs more slowly by three orders of magnitude (6.9 x 10^{-5} s^{-1}).[190] However, what makes S11 a promising tool to test and potentially engineer is that its C-terminal intein fragment (S11C) consists of only 6 residues (compared to 147 residues in the N-terminal fragment, S11N). If a reporter cell could be built in which candidate CPMPs are fused to a 6-residue intein S11C fragment, this could be an excellent alternative to the 16-residue VidN.

Initial challenges to overcome are S11's temperature-dependent intein-splicing efficiency and the lack of data on its activity in mammalian cells. 84% of S11C has been shown to undergo trans-splicing at room temperature, but only 16% at 37℃.[190] In addition, all studies characterizing S11 used *E. coli* for experimentation.[190–193] Hence, initial experiments would need to verify whether S11 splicing can be observed in HEK

293T cells (or another mammalian cell line) under standard cell culture conditions. But, even if the splicing efficiency may be as low as 16%, this might be tolerable because the nature of constitutive gene transcription would still lead to the amplified expression of mKate2.

Analyze and improve split-intein kinetics and reporter gene expression

In order for an intein-based mKate2 expression system to be successful, an important goal is that the fluorescent protein should be detected as quickly as possible when EHEE-VidN and TMD-VidC-TF undergo intein-splicing. Improving this system could be achieved by engineering the intein (VidaL or S11) for increased kinetics, increased intein accessibility for each other, and/or selecting a brighter and more quickly maturing fluorescent protein.

First, improving the splicing kinetics of split-inteins could improve the reporter system. The rate-limiting step in intein-splicing has been shown to be influenced by the three residues immediately flanking either intein fragment.[154,194] Studies on the Npu DnaE intein, for example, revealed that mutations away from the wildtype phenylalanine at the second position flanking the C-terminal intein reduces splicing reaction efficiency.[154,195,196] Unlike DnaE, however, one intein fragment in VidaL and S11 is considerably smaller than the other fragment, and phenylalanine (or a bulky residue) is only found at the second residue position in S11N. As a result, it may be possible that a different chemistry governs the splicing reaction for VidaL and S11. Mutagenesis studies at this key position and analysis of the relative amount of spliced and unspliced mutant

TMD-VidC-TF across different time points by Western blot, fluorescent microscopy, and flow cytometry could be initial explorations.

Another approach to improve the intein splicing activity could be increasing the accessibility of the membrane-tethered intein by adding a glycine-serine or structured linker. Flexible glycine-serine linkers are common in biological research to join two proteins or a peptide tag, and testing such linkers of various lengths between the C-terminus of VidC and the N-terminus of the transcription factor in TMD-VidC-TF may allow increased space for the EHEE-VidN to access its complementary intein.

Flexible linkers have been shown to behave like a random coil but can be stiffened by replacing the glycines with serines.[197] Alternatively, a structured linker like EAAAK or PAPAP could reduce the degrees of freedom that a flexible linker would posses and create a stiff scaffold for the trans-splicing reaction to occur more efficiently.

Finally, we selected mKate2, which at the time of its generation was one of the brightest far-red fluorescent proteins,[198] as the reporter protein because it was used successfully with intein splicing and the transcription factor ZF1-VP64.[75] An alternative could be to test a recently engineered mScarlet-I3,[199] which has been shown to be brighter and mature faster (2 min vs. 34 min[200]) than mKate2.

Test cell uptake of CPPs and intein-mediated reporter gene expression

To validate that mKate2 can be activated because a protein crossed the cell plasma membrane, we could test the cytosolic uptake of known CPPs and subsequent downstream mKate2 expression. Fusing VidN to the CPPs, Tat and Penetratin, as well as the CPMP ZF5.3, and performing Western blot, fluorescent microscopy, and flow cytometry could demonstrate that this system can be used to successfully detect cell-penetration.

To test the sensitivity of TMD-VidC-TF, we could measure mKate2 activation when various concentrations of purified Tat-VidN, Penetratin-VidN, and ZF5.3-VidN (the three species referred to as CPP-VidN) are added onto HEK 293T cells. However, the relationship between the concentration of purified CPP-VidN and reporter gene expression may not be directly comparable because different amounts of each CPP-VidN may be trapped in endosomes and not enter the cytosol. Using fluorescence correlation spectroscopy[67] could be used to quantify the cytosolic entry of all CPP-VidN so that we can determine the efficiency at which CPP-VidN has entered the cytosol.

Generate stable reporter cell line

Once the optimized TMD-VidC-TF is determined, generating a stable cell line with the optimized construct and the reporter gene would facilitate testing CPMPs because we would bypass co-transfection of mammalian cells. Transducing cells using bicistronic lentivirus [201] could allow us to stably integrate an optimized TMD-VidC-TF driven under a constitutive promoter and a fluorescent reporter protein driven under the ZF1x6-C promoter.[78,75] An alternative method could be to utilize a Bxb1 recombinase system that can insert multiple genes at targeted locations on the chromosome in mammalian cells.[202] This approach can overcome the limitations of lentivirus, namely the inability to control the integration and copy number of a gene-of-interest.

4.3 Building the secretor cell

This dissertation did not focus on a “secretor” cell, but here I will offer brief thoughts on how we could build one. We envision a successful reporter cell to function in tandem with a secretor cell, which expresses and secretes a candidate CPMP.

Testing positive controls

A recent study demonstrated that inside a droplet a yeast cell secreting murine interleukin-3 (mIL-3) can be co-encapsulated with another murine cell that expresses GFP upon mIL-3 stimulation.[77] Thus, we could re-purpose yeast cells to secrete a candidate CPMP-VidN by fusing an app8 secretion signal peptide at the N-terminus of the CPMP-VidN gene in a yeast vector. The inducible GAL-1 promoter is commonly used to drive transgenes in yeast.[203] However, engineered yeast promoters, such as pGPD-15, have been shown to exceed natural promoters in driving gene expression and could be tested here.[204] An initial qualitative test could be separating transformed yeast from the media (where the secreted CPMP-VidN will be present) by centrifugation, and adding the supernatant to HEK 293T cells expressing the TMD-VidC-TF and reporter gene.

Optimize culture media for yeast and mammalian cell growth

Yeast and mammalian cells grow optimally under different conditions. *S. cerevisiae,* for example, is typically grown at 30°C and can be cultured in broth containing yeast extract, peptone, dextrose. However, mammalian cells like HEK 293T cells are cultured in DMEM, DMEM-F12, or RPMI-1640 that are supplemented with FBS and antibiotics at environmental conditions of 37°C and 5% CO_2. Successful co-culturing yeast cells with murine Ba/F3 cells in a mixture of RPMI and DMEM-F12 has been shown, although yeast cell growth was slightly reduced.[77] Nonetheless, this suggests that our planned use of co-encapsulating HEK 293T cells with yeast cells could work. But some optimization may be required to identify a mixture of media that both HEK 293T cells and yeast cells would tolerate.

4.4 Final thoughts

This dissertation investigated two main research ideas. First, I combined computational protein design with high-throughput experimentation to generate stable αββα miniproteins and quantify biophysical determinants of folding stability. Second, I tested the potential of coupling split-inteins with a synthetic transcription factor to create a reporter cell that could identify cell-penetrating miniproteins for a future high-throughput droplet microfluidic screen. Future research that can further integrate the work in this dissertation by designing and actually screening for stable CPMP binders against therapeutic targets could greatly contribute to improving human health. Advances in computational protein design, droplet microfluidics methods, and synthetic biology paradigms can make this a reality.

REFERENCES

1. Anson, M.L., and Mirsky, A.E. (1925). On Some General Properties of Proteins. J. Gen. Physiol. *9*, 169–179.

2. Mirsky, A.E., and Pauling, L. (1936). On the Structure of Native, Denatured, and Coagulated Proteins. Proc. Natl. Acad. Sci. U. S. A. *22*, 439–447. 10.1073/pnas.22.7.439.

3. Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry *29*, 7133–7155. 10.1021/bi00483a001.

4. Vogt, G., Woell, S., and Argos, P. (1997). Protein thermal stability, hydrogen bonds, and ion pairs. J. Mol. Biol. *269*, 631–643. 10.1006/jmbi.1997.1042.

5. Joh, N.H., Min, A., Faham, S., Whitelegge, J.P., Yang, D., Woods, V.L., and Bowie, J.U. (2008). Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. Nature *453*, 1266–1270. 10.1038/nature06977.

6. Pace, C.N., Fu, H., Fryar, K.L., Landua, J., Trevino, S.R., Shirley, B.A., Hendricks, M.M., Iimura, S., Gajiwala, K.S., Scholtz, J.M., et al. (2011). Contribution of Hydrophobic Interactions to Protein Stability. J. Mol. Biol. *408*, 514–528. 10.1016/j.jmb.2011.02.053.

7. Pace, C.N., Fu, H., Fryar, K.L., Landua, J., Trevino, S.R., Schell, D., Thurlkill, R.L., Imura, S., Scholtz, J.M., Gajiwala, K.S., et al. (2014). Contribution of hydrogen bonds to protein stability. Protein Sci. *23*, 652–661. 10.1002/pro.2449.

8. Nishio, M., Umezawa, Y., Fantini, J., S. Weiss, M., and Chakrabarti, P. (2014). CH–π hydrogen bonds in biological macromolecules. Phys. Chem. Chem. Phys. *16*, 12648–12683. 10.1039/C4CP00099D.

9. Serrano, L., and Fersht, A.R. (1989). Capping and α-helix stability. Nature *342*, 296–299. 10.1038/342296a0.

10. Chakrabartty, A., Doig, A.J., and Baldwin, R.L. (1993). Helix capping propensities in peptides parallel those in proteins. Proc. Natl. Acad. Sci. U. S. A. *90*, 11332–11336. 10.1073/pnas.90.23.11332.

11. George, R.A., and Heringa, J. (2002). An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng. Des. Sel. *15*, 871–879. 10.1093/protein/15.11.871.

12. Strickler, S.S., Gribenko, A.V., Gribenko, A.V., Keiffer, T.R., Tomlinson, J., Reihle, T., Loladze, V.V., and Makhatadze, G.I. (2006). Protein Stability and Surface Electrostatics: A Charged Relationship. Biochemistry *45*, 2761–2766. 10.1021/bi0600143.

13. Donald, J.E., Kulp, D.W., and DeGrado, W.F. (2011). Salt bridges: Geometrically specific, designable interactions. Proteins Struct. Funct. Bioinforma. *79*, 898–915. 10.1002/prot.22927.

14. Tsai, M.-Y., Zheng, W., Balamurugan, D., Schafer, N.P., Kim, B.L., Cheung, M.S., and Wolynes, P.G. (2016). Electrostatics, structure prediction, and the energy landscapes for protein folding and binding. Protein Sci. *25*, 255–269. 10.1002/pro.2751.

15. Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327. 10.1038/nature19946.

16. Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics. Protein Sci. *28*, 678–683. 10.1002/pro.3588.

17. Woolfson, D.N. (2021). A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. J. Mol. Biol. *433*, 167160. 10.1016/j.jmb.2021.167160.

18. Bhardwaj, G., O'Connor, J., Rettie, S., Huang, Y.-H., Ramelot, T.A., Mulligan, V.K., Alpkilic, G.G., Palmer, J., Bera, A.K., Bick, M.J., et al. (2022). Accurate de novo design of membrane-traversing macrocycles. Cell *185*, 3520-3532.e26. 10.1016/j.cell.2022.07.019.

19. Rhys, G.G., Cross, J.A., Dawson, W.M., Thompson, H.F., Shanmugaratnam, S., Savery, N.J., Dodding, M.P., Höcker, B., and Woolfson, D.N. (2022). De novo designed peptides for cellular delivery and subcellular localisation. Nat. Chem. Biol. *18*, 999–1004. 10.1038/s41589-022-01076-6.

20. Richardson, J.S., and Richardson, D.C. (1989). The de novo design of protein structures. Trends Biochem. Sci. *14*, 304–309. doi:10.1016/0968-0004(89)90070-4.

21. Korendovych, I.V., and DeGrado, W.F. (2020). De novo protein design, a retrospective. Q. Rev. Biophys. *53*. 10.1017/s0033583519000131.

22. Procko, E., Hedman, R., Hamilton, K., Seetharaman, J., Fleishman, S.J., Su, M., Aramini, J., Kornhaber, G., Hunt, J.F., Tong, L., et al. (2013). Computational Design of a Protein-Based Enzyme Inhibitor. J. Mol. Biol. *425*, 3563–3575. 10.1016/j.jmb.2013.06.035.

23. Chevalier, A., Silva, D.-A., Rocklin, G.J., Hicks, D.R., Vergara, R., Murapa, P., Bernard, S.M., Zhang, L., Lam, K.-H., Yao, G., et al. (2017). Massively parallel de novo protein design for targeted therapeutics. Nature *550*, 74–79. 10.1038/nature23912.

24. Bryan, C.M., Rocklin, G.J., Bick, M.J., Ford, A., Majri-Morrison, S., Kroll, A.V., Miller, C.J., Carter, L., Goreshnik, I., Kang, A., et al. (2021). Computational design of a synthetic PD-1 agonist. Proc. Natl. Acad. Sci. *118*, e2102164118. 10.1073/pnas.2102164118.

25. Quijano-Rubio, A., Yeh, H.-W., Park, J., Lee, H., Langan, R.A., Boyken, S.E., Lajoie, M.J., Cao, L., Chow, C.M., Miranda, M.C., et al. (2021). De novo design of modular and tunable protein biosensors. Nature *591*, 482–487. 10.1038/s41586-021-03258-z.

26. Yang, C., Sesterhenn, F., Bonet, J., van Aalen, E.A., Scheller, L., Abriata, L.A., Cramer, J.T., Wen, X., Rosset, S., Georgeon, S., et al. (2021). Bottom-up de novo design of functional proteins with complex structural features. Nat. Chem. Biol. *17*, 492–500. 10.1038/s41589-020-00699-x.

27. Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S., et al. (2019). De novo design of bioactive protein switches. Nature *572*, 205–210. 10.1038/s41586-019-1432-8.

28. Shui, S., Gainza, P., Scheller, L., Yang, C., Kurumida, Y., Rosset, S., Georgeon, S., Di Roberto, R.B., Castellanos-Rueda, R., Reddy, S.T., et al. (2021). A rational blueprint for the design of chemically-controlled protein switches. Nat. Commun. *12*, 5754. 10.1038/s41467-021-25735-9.

29. Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D.L., Prilusky, J., et al. (2018). Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. Mol. Cell *72*, 178-186.e5. 10.1016/j.molcel.2018.08.033.

30. Lau, Y.-T.K., Baytshtok, V., Howard, T.A., Fiala, B.M., Johnson, J.M., Carter, L.P., Baker, D., Lima, C.D., and Bahl, C.D. (2018). Discovery and engineering of enhanced SUMO protease enzymes. J. Biol. Chem. *293*, 13224–13233. 10.1074/jbc.RA118.004146.

31. Lapidoth, G., Khersonsky, O., Lipsh, R., Dym, O., Albeck, S., Rogotner, S., and Fleishman, S.J. (2018). Highly active enzymes by automated combinatorial backbone assembly and sequence design. Nat. Commun. *9*, 2780. 10.1038/s41467-018-05205-5.

32. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D., and André, I. (2011). Modeling Symmetric Macromolecular Structures in Rosetta3. PLOS ONE *6*, e20450. 10.1371/journal.pone.0020450.

33. Vulovic, I., Yao, Q., Park, Y.-J., Courbet, A., Norris, A., Busch, F., Sahasrabuddhe, A., Merten, H., Sahtoe, D.D., Ueda, G., et al. (2021). Generation of ordered protein assemblies using rigid three-body fusion. Proc. Natl. Acad. Sci. *118*, e2015037118. 10.1073/pnas.2015037118.

34. Courbet, A., Hansen, J., Hsia, Y., Bethel, N., Park, Y.-J., Xu, C., Moyer, A., Boyken, S.E., Ueda, G., Nattermann, U., et al. (2022). Computational design of mechanically coupled axle-rotor protein assemblies. Science *376*, 383–390. 10.1126/science.abm1183.

35. Marcos, E., and Silva, D. (2018). Essentials of *de novo* protein design: Methods and applications. WIREs Comput. Mol. Sci. *8*. 10.1002/wcms.1374.

36. Leman, J.K., Weitzner, B.D., Lewis, S.M., Adolf-Bryfogle, J., Alam, N., Alford, R.F., Aprahamian, M., Baker, D., Barlow, K.A., Barth, P., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat. Methods *17*, 665–680. 10.1038/s41592-020-0848-2.

37. Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V.K., Chevalier, A., et al. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. Science *357*, 168–175. 10.1126/science.aan0693.

38. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T., Montelione, G.T., and Baker, D. (2012). Principles for designing ideal protein structures. Nature *491*, 222–227. 10.1038/nature11600.

39. Marcos, E., Chidyausiku, T.M., McShan, A.C., Evangelidis, T., T. Evangelidis, Nerli, S., Carter, L., Nivón, L.G., Davis, A., Audrey Davis, et al. (2018). De novo design of a non-local beta-sheet protein with high stability and accuracy. Nat. Struct. Mol. Biol. *25*, 1028–1034. 10.2210/pdb6e5c/pdb.

40. Brunette, T.J., Fabio Parmeggiani, 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*, 580–584. 10.1038/nature16162.

41. Harteveld, Z., Bonet, J., Rosset, S., Yang, C., Sesterhenn, F., and Correia, B.E. (2022). A generic framework for hierarchical de novo protein design. Proc. Natl. Acad. Sci. *119*, e2206111119. 10.1073/pnas.2206111119.

42. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. *20*, 681–697. 10.1038/s41580-019-0163-x.

43. Yang, Y., Ding, X., Zhu, G., Niroula, A., Lv, Q., and Vihinen, M. (2019). ProTstab – predictor for cellular protein stability. BMC Genomics *20*, 804. 10.1186/s12864-019-6138-7.

44. Fang, J. (2020). A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief. Bioinform. *21*, 1285–1292. 10.1093/bib/bbz071.

45. Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput. Struct. Biotechnol. J. *18*, 1968–1979. 10.1016/j.csbj.2020.07.011.

46. Marabotti, A., Scafuri, B., and Facchiano, A. (2021). Predicting the stability of mutant proteins by computational approaches: an overview. Brief. Bioinform. *22*, bbaa074. 10.1093/bib/bbaa074.

47. Kortemme, T., and Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. Proc. Natl. Acad. Sci. *99*, 14116–14121. 10.1073/pnas.202485799.

48. Caldararu, O., Blundell, T.L., and Kepp, K.P. (2021). Three Simple Properties Explain Protein Stability Change upon Mutation. J. Chem. Inf. Model. *61*, 1981–1988. 10.1021/acs.jcim.1c00201.

49. Bazzoli, A., Kelow, S.P., and Karanicolas, J. (2015). Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions. PLOS ONE *10*, e0140359. 10.1371/journal.pone.0140359.

50. Pavlovicz, R.E., Park, H., and DiMaio, F. (2020). Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. PLOS Comput. Biol. *16*, e1008103. 10.1371/journal.pcbi.1008103.

51. Atsavapranee, B., Stark, C.D., Sunden, F., Thompson, S., and Fordyce, P.M. (2021). Fundamentals to function: Quantitative and scalable approaches for measuring protein stability. Cell Syst. *12*, 547–560. 10.1016/j.cels.2021.05.009.

52. Spradlin, J.N., Zhang, E., and Nomura, D.K. (2021). Reimagining Druggability Using Chemoproteomic Platforms. Acc. Chem. Res. *54*, 1801–1813. 10.1021/acs.accounts.1c00065.

53. Crook, Z.R., Nairn, N.W., and Olson, J.M. (2020). Miniproteins as a Powerful Modality in Drug Development. Trends Biochem. Sci. *45*, 332–346. 10.1016/j.tibs.2019.12.008.

54. Carter, P.J., and Rajpal, A. (2022). Designing antibodies as therapeutics. Cell *185*, 2789–2805. 10.1016/j.cell.2022.05.029.

55. Boutajangout, A., Lindberg, H., Awwad, A., Paul, A., Baitalmal, R., Almokyad, I., Höidén-Guthenberg, I., Gunneriusson, E., Frejd, F.Y., Härd, T., et al. (2019). Affibody-Mediated Sequestration of Amyloid β Demonstrates Preventive Efficacy in a Transgenic Alzheimer's Disease Mouse Model. Front. Aging Neurosci. *11*.

56. Grossmann, T.N., Yeh, J.T.-H., Bowman, B.R., Chu, Q., Moellering, R.E., and Verdine, G.L. (2012). Inhibition of oncogenic Wnt signaling through direct targeting of β-catenin. Proc. Natl. Acad. Sci. *109*, 17942–17947. 10.1073/pnas.1208396109.

57. de la Torre, B.G., and Albericio, F. (2023). The Pharmaceutical Industry in 2022: An Analysis of FDA Drug Approvals from the Perspective of Molecules. Molecules *28*, 1038. 10.3390/molecules28031038.

58. Hariton-Gazal, E., Rosenbluh, J., Graessmann, A., Gilon, C., and Loyter, A. (2003). Direct translocation of histone molecules across cell membranes. J. Cell Sci. *116*, 4577–4586. 10.1242/jcs.00757.

59. Hopkins, B.D., Fine, B., Steinbach, N., Dendy, M., Rapp, Z., Shaw, J., Pappas, K., Yu, J.S., Hodakoski, C., Mense, S., et al. (2013). A Secreted PTEN Phosphatase That Enters Cells to Alter Signaling and Survival. Science *341*, 399–402. 10.1126/science.1234907.

60. Appelbaum, J.S., LaRochelle, J.R., Smith, B.A., Balkin, D.M., Holub, J.M., and Schepartz, A. (2012). Arginine Topology Controls Escape of Minimally Cationic Proteins from Early Endosomes to the Cytoplasm. Chem. Biol. *19*, 819–830. 10.1016/j.chembiol.2012.05.022.

61. Desale, K., Kuche, K., and Jain, S. (2021). Cell-penetrating peptides (CPPs): an overview of applications for improving the potential of nanotherapeutics. Biomater. Sci. *9*, 1153–1188. 10.1039/D0BM01755H.

62. Guidotti, G., Brambilla, L., and Rossi, D. (2017). Cell-Penetrating Peptides: From Basic Research to Clinics. Trends Pharmacol. Sci. *38*, 406–424. 10.1016/j.tips.2017.01.003.

63. Zorko, M., Jones, S., and Langel, Ü. (2022). Cell-penetrating peptides in protein mimicry and cancer therapeutics. Adv. Drug Deliv. Rev. *180*, 114044. 10.1016/j.addr.2021.114044.

64. Steinauer, A., LaRochelle, J.R., Knox, S.L., Wissner, R.F., Berry, S., and Schepartz, A. (2019). HOPS-dependent endosomal fusion required for efficient cytosolic delivery of

therapeutic peptides and small proteins. Proc. Natl. Acad. Sci. *116*, 512–521. 10.1073/pnas.1812044116.

65. Chen, K., and Pei, D. (2020). Engineering Cell-Permeable Proteins through Insertion of Cell-Penetrating Motifs into Surface Loops. ACS Chem. Biol. *15*, 2568–2576. 10.1021/acschembio.0c00593.

66. Pei, D., and Dalbey, R.E. (2022). Membrane translocation of folded proteins. J. Biol. Chem. *298*, 102107. 10.1016/j.jbc.2022.102107.

67. Wissner, R.F., Steinauer, A., Knox, S.L., Thompson, A.D., and Schepartz, A. (2018). Fluorescence Correlation Spectroscopy Reveals Efficient Cytosolic Delivery of Protein Cargo by Cell-Permeant Miniature Proteins. ACS Cent. Sci. *4*, 1379–1393. 10.1021/acscentsci.8b00446.

68. Mientkiewicz, K.M., Peraro, L., and Kritzer, J.A. (2021). Parallel Screening Using the Chloroalkane Penetration Assay Reveals Structure-Penetration Relationships. ACS Chem. Biol. *16*, 1184–1190. 10.1021/acschembio.1c00434.

69. Liu, J., and Afshar, S. (2020). In Vitro Assays: Friends or Foes of Cell-Penetrating Peptides. Int. J. Mol. Sci. *21*, 4719. 10.3390/ijms21134719.

70. Milech, N., Longville, B.A., Cunningham, P.T., Scobie, M.N., Bogdawa, H.M., Winslow, S., Anastasas, M., Connor, T., Ong, F., Stone, S.R., et al. (2015). GFP-complementation assay to detect functional CPP and protein delivery into living cells. Sci. Rep. *5*, 18329. 10.1038/srep18329.

71. Kauffman, W.B., Guha, S., and Wimley, W.C. (2018). Synthetic molecular evolution of hybrid cell penetrating peptides. Nat. Commun. *9*, 2568. 10.1038/s41467-018-04874-6.

72. Peraro, L., Deprey, K.L., Moser, M.K., Zou, Z., Ball, H.L., Levine, B., and Kritzer, J.A. (2018). Cell Penetration Profiling Using the Chloroalkane Penetration Assay. J. Am. Chem. Soc. *140*, 11360–11369. 10.1021/jacs.8b06144.

73. Peier, A., Ge, L., Boyer, N., Frost, J., Duggal, R., Biswas, K., Edmondson, S., Hermes, J.D., Yan, L., Zimprich, C., et al. (2021). NanoClick: A High Throughput, Target-Agnostic Peptide Cell Permeability Assay. ACS Chem. Biol. *16*, 293–309. 10.1021/acschembio.0c00804.

74. Shah, N.H., and Muir, T.W. (2011). Split Inteins: Nature’s Protein Ligases. Isr. J. Chem. *51*, 854–861. 10.1002/ijch.201100094.

75. Muldoon, J.J., Kandula, V., Hong, M., Donahue, P.S., Boucher, J., Bagheri, N., and Leonard, J.N. (2021). Model-guided design of mammalian genetic programs. Sci. Adv. *7*, 13. 10.1126/sciadv.abe937.

76. Burton, A.J., Haugbro, M., Parisi, E., and Muir, T.W. (2020). Live-cell protein engineering with an ultra-short split intein. Proc. Natl. Acad. Sci. *117*, 12041–12049. 10.1073/pnas.2003613117.

77. Yanakieva, D., Elter, A., Bratsch, J., Friedrich, K., Becker, S., and Kolmar, H. (2020). FACS-Based Functional Protein Screening via Microfluidic Co-encapsulation of Yeast

Secretor and Mammalian Reporter Cells. Sci. Rep. *10*, 10182. 10.1038/s41598-020-66927-5.

78. Donahue, P.S., Draut, J.W., Muldoon, J.J., Edelstein, H.I., Bagheri, N., and Leonard, J.N. (2020). The COMET toolkit for composing customizable genetic programs in mammalian cells. Nat. Commun. *11*, 779. 10.1038/s41467-019-14147-5.

79. Goldenzweig, A., and Fleishman, S.J. (2018). Principles of Protein Stability and Their Application in Computational Design. Annu. Rev. Biochem. *87*, 105–129. 10.1146/annurev-biochem-062917-012102.

80. Arai, M. (2018). Unified Understanding of Folding and Binding Mechanisms of Globular and Intrinsically Disordered Proteins. Biophys. Rev. *10*, 163–181. 10.1007/s12551-017-0346-7.

81. Boyken, S.E., Chen, Z., Groves, B., Langan, R.A., Oberdorfer, G., Ford, A., Gilmore, J.M., Xu, C., DiMaio, F., Pereira, J.H., et al. (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. Science *352*, 680–687. 10.1126/science.aad8865.

82. Pan, X., Thompson, M.C., Zhang, Y., Liu, L., Fraser, J.S., Kelly, M.J.S., and Kortemme, T. (2020). Expanding the space of protein geometries by computational design of de novo fold families. Science *369*, 1132–1136.

83. Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., et al. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. Mol. Cell *63*, 337–346. 10.1016/j.molcel.2016.06.012.

84. Wiese, J.G., Shanmugaratnam, S., and Höcker, B. (2021). Extension of a de novo TIM barrel with a rationally designed secondary structure element. Protein Sci. *30*, 982–989. 10.1002/pro.4064.

85. Lalaurie, C.J., Dufour, V., Meletiou, A., Ratcliffe, S., Harland, A., Wilson, O., Vamasiri, C., Shoemark, D.K., Williams, C., Arthur, C.J., et al. (2018). The de novo design of a biocompatible and functional integral membrane protein using minimal sequence complexity. Sci. Rep. *8*, 14564. 10.1038/s41598-018-31964-8.

86. Nisthal, A., Wang, C.Y., Ary, M.L., and Mayo, S.L. (2019). Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc. Natl. Acad. Sci. *116*, 16367–16377. 10.1073/pnas.1903888116.

87. Broom, A., Trainor, K., Jacobi, Z., and Meiering, E.M. (2020). Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. Structure *28*, 717-726.e3. 10.1016/j.str.2020.04.003.

88. Brini, E., Simmerling, C., and Dill, K. (2020). Protein storytelling through physics. Science *370*, eaaz3041. 10.1126/science.aaz3041.

89. Basanta, B., Bick, M.J., Bera, A.K., Norn, C., Chow, C.M., Carter, L.P., Goreshnik, I., Dimaio, F., and Baker, D. (2020). An enumerative algorithm for de novo design of proteins

with diverse pocket structures. Proc. Natl. Acad. Sci. *117*, 22135–22145. 10.1073/pnas.2005412117.

90. Dou, J., Vorobieva, A.A., Sheffler, W., Doyle, L., Park, H., Bick, M.J., Mao, B., Binchen Mao, Foight, G.W., G.W. Foight, et al. (2018). De novo design of a fluorescence-activating β-barrel. Nature *561*, 485–491. 10.1038/s41586-018-0509-0.

91. Linsky, T.W., Noble, K., Tobin, A.R., Crow, R., Carter, L., Urbauer, J.L., Baker, D., and Strauch, E.-M. (2022). Sampling of structure and sequence space of small protein folds. Nat. Commun. *13*, 7151. 10.1038/s41467-022-34937-8.

92. Maxwell, K.L., Yee, A.A., Booth, V., Arrowsmith, C.H., Gold, M., and Davidson, A.R. (2001). The solution structure of bacteriophage λ protein W, a small morphogenetic protein possessing a novel fold11Edited by P. E. Wright. J. Mol. Biol. *308*, 9–14. 10.1006/jmbi.2001.4582.

93. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. 10.1038/s41586-021-03819-2.

94. Huang, P.-S., Ban, Y.-E.A., Richter, F., André, I., Vernon, R.B., Schief, W.R., and Baker, D. (2011). RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. PLOS ONE *6*. 10.1371/journal.pone.0024109.

95. Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D., and DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J. Chem. Theory Comput. *12*, 6201–6212. 10.1021/acs.jctc.6b00819.

96. Maguire, J.B., Haddox, H.K., Strickland, D., Halabiya, S.F., Coventry, B., Griffin, J.R., Pulavarti, S.V.S.R.K., Cummins, M., Thieker, D.F., Klavins, E., et al. (2021). Perturbing the energy landscape for improved packing during computational protein design. Proteins Struct. Funct. Bioinforma. *89*, 436–449. 10.1002/prot.26030.

97. Monera, O.D., Sereda, T.J., Zhou, N.E., Kay, C.M., and Hodges, R.S. (1995). Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. J. Pept. Sci. *1*, 319–329. 10.1002/psc.310010507.

98. Huyghues-Despointes, B.M.P., Scholtz, J.M., and Pace, C.N. (1999). Protein conformational stabilities can be determined from hydrogen exchange rates. Nat. Struct. Biol. *6*, 3.

99. Negin, R.S., and Carbeck, J.D. (2002). Measurement of Electrostatic Interactions in Protein Folding with the Use of Protein Charge Ladders. J. Am. Chem. Soc. *124*, 2911–2916. 10.1021/ja0169567.

100. Wy, W., Ej, M.-W., and Milner-White, E.J. (1999). A recurring two-hydrogen-bond motif incorporating a serine or threonine residue is found both at alpha-helical N termini and in other situations. J. Mol. Biol. *286*, 1651–1662. 10.1006/jmbi.1999.2551.

101. Pace, C.N., Scholtz, J.M., and Grimsley, G.R. (2014). Forces stabilizing proteins. FEBS Lett. *588*, 2177–2184. 10.1016/j.febslet.2014.05.006.

102. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. 10.1002/bip.360221211.

103. Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend, G. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. *43*, D364–D368. 10.1093/nar/gku1028.

104. Lin, Y.-R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A.F., Montelione, G.T., and Baker, D. (2015). Control over overall shape and size in de novo designed proteins. Proc. Natl. Acad. Sci. *112*. 10.1073/pnas.1509508112.

105. Singer, J.M., Novotney, S., Strickland, D., Haddox, H.K., Leiby, N., Rocklin, G.J., Chow, C.M., Roy, A., Bera, A.K., Motta, F.C., et al. (2022). Large-scale design and refinement of stable proteins using sequence-only models. PLOS ONE *17*, e0265020. 10.1371/journal.pone.0265020.

106. Baker, E.G., Williams, C., Hudson, K.L., Bartlett, G.J., Heal, J.W., Porter Goff, K.L., Sessions, R.B., Crump, M.P., and Woolfson, D.N. (2017). Engineering protein stability with atomic precision in a monomeric miniprotein. Nat. Chem. Biol. *13*, 764–770. 10.1038/nchembio.2380.

107. Trotter, D., and Wallin, S. (2020). Effects of Topology and Sequence in Protein Folding Linked via Conformational Fluctuations. Biophys. J. *118*, 1370–1380. 10.1016/j.bpj.2020.01.020.

108. Marin, F.I., Johansson, K.E., O’Shea, C., Lindorff-Larsen, K., and Winther, J.R. (2021). Computational and Experimental Assessment of Backbone Templates for Computational Redesign of the Thioredoxin Fold. J. Phys. Chem. B *125*, 11141–11149. 10.1021/acs.jpcb.1c05528.

109. Bellesia, G., Jewett, A.I., and Shea, J.-E. (2011). Relative stability of *de novo* four-helix bundle proteins: Insights from coarse grained molecular simulations: Relative Stability of *de novo* Four Helix Bundle Proteins. Protein Sci. *20*, 818–826. 10.1002/pro.605.

110. Ha-Duong, T. (2014). Coarse-Grained Models of the Proteins Backbone Conformational Dynamics. In Protein Conformational Dynamics Advances in Experimental Medicine and Biology., K. Han, X. Zhang, and M. Yang, eds. (Springer International Publishing), pp. 157–169. 10.1007/978-3-319-02970-2_7.

111. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O’Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. *13*, 3031–3048. 10.1021/acs.jctc.7b00125.

112. Cao, L., Goreshnik, I., Coventry, B., Case, J.B., Miller, L., Kozodoy, L., Chen, R.E., Chen, R.E., Carter, L., Walls, A.C., et al. (2020). De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science *370*, 426–431. 10.1126/science.abd9909.

113. Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J.S., Jude, K.M., Marković, I., Kadam, R.U., Verschueren, K.H.G., et al. (2022). Design of protein-binding proteins from the target structure alone. Nature *605*, 551–560. 10.1038/s41586-022-04654-9.

114. Nguyen, D., Mayne, L., Phillips, M.C., and Walter Englander, S. (2018). Reference Parameters for Protein Hydrogen Exchange Rates. J. Am. Soc. Mass Spectrom. *29*, 1936–1939. 10.1007/s13361-018-2021-z.

115. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

116. Hoover, D.M. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res. *30*, 43e–443. 10.1093/nar/30.10.e43.

117. Satwik Kamtekar, Satwik Kamtekar, Satwik Kamtekar, Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., and Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. Science *262*, 1680–1685. 10.1126/science.8259512.

118. Vrana, J., de Lange, O., Yang, Y., Newman, G., Saleem, A., Miller, A., Cordray, C., Halabiya, S., Parks, M., Lopez, E., et al. (2021). Aquarium: open-source laboratory software for design, execution and data management. Synth. Biol. *6*, ysab006. 10.1093/synbio/ysab006.

119. Lemak, A., Steren, C.A., Arrowsmith, C.H., and Llinás, M. (2008). Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach. J. Biomol. NMR *41*, 29–41. 10.1007/s10858-008-9238-2.

120. Lemak, A., Gutmanas, A., Chitayat, S., Karra, M., Farès, C., Sunnerhagen, M., and Arrowsmith, C.H. (2011). A novel strategy for NMR resonance assignment and protein structure determination. J. Biomol. NMR *49*, 27–38. 10.1007/s10858-010-9458-0.

121. Kazimierczuk, K., and Orekhov, V.Yu. (2011). Accelerated NMR Spectroscopy by Using Compressed Sensing. Angew. Chem. Int. Ed. *50*, 5556–5559. 10.1002/anie.201100370.

122. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293. 10.1007/BF00197809.

123. Lee, W., Tonelli, M., and Markley, J.L. (2015). NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics *31*, 1325–1327. 10.1093/bioinformatics/btu830.

124. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR *44*, 213–223. 10.1007/s10858-009-9333-z.

125. Güntert, P. (2004). Automated NMR Structure Calculation With CYANA. In Protein NMR Techniques Methods in Molecular Biology^{TM}., A. K. Downing, ed. (Humana Press), pp. 353–378. 10.1385/1-59259-809-9:353.

126. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921. 10.1107/S0907444998003254.

127. Linge, J.P., Williams, M.A., Spronk, C.A.E.M., Bonvin, A.M.J.J., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins Struct. Funct. Bioinforma. *50*, 496–506. 10.1002/prot.10299.

128. Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291. 10.1107/S0021889892009944.

129. Bhattacharya, A., Tejero, R., and Montelione, G.T. (2007). Evaluating protein structures determined by structural genomics consortia. Proteins Struct. Funct. Bioinforma. *66*, 778–795. 10.1002/prot.21165.

130. Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D., and Kay, L.E. (1994). Backbone Dynamics of a Free and a Phosphopeptide-Complexed Src Homology 2 Domain Studied by 15N NMR Relaxation. Biochemistry *33*, 5984–6003. 10.1021/bi00185a040.

131. Kay, L.E., Torchia, D.A., and Bax, A. (1989). Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry *28*, 8972–8979. 10.1021/bi00449a003.

132. Connelly, G.P., Bai, Y., Jeng, M.-F., and Englander, S.W. (1993). Isotope effects in peptide group hydrogen exchange. Proteins Struct. Funct. Bioinforma. *17*, 87–92. 10.1002/prot.340170111.

133. Bai, Y., Milne, J.S., Mayne, L., and Englander, S.W. (1993). Primary structure effects on peptide group hydrogen exchange. Proteins Struct. Funct. Bioinforma. *17*, 75–86. 10.1002/prot.340170110.

134. Mabonga, L., and Kappo, A.P. (2019). Protein-protein interaction modulators: advances, successes and remaining challenges. Biophys. Rev. *11*, 559–581. 10.1007/s12551-019-00570-x.

135. Samatar, A.A., and Poulikakos, P.I. (2014). Targeting RAS–ERK signalling in cancer: promises and challenges. Nat. Rev. Drug Discov. *13*, 928–942. 10.1038/nrd4281.

136. Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W.R., Wang, R., Huang, J., Hao, T., et al. (2021). Comprehensive characterization of protein–protein interactions perturbed by disease mutations. Nat. Genet. *53*, 342–353. 10.1038/s41588-020-00774-y.

137. Sinsky, J., Pichlerova, K., and Hanes, J. (2021). Tau Protein Interaction Partners and Their Roles in Alzheimer's Disease and Other Tauopathies. Int. J. Mol. Sci. *22*, 9207. 10.3390/ijms22179207.

138. Contini, A., Ferri, N., Bucci, R., Lupo, M.G., Erba, E., Gelmi, M.L., and Pellegrino, S. (2018). Peptide modulators of Rac1/Tiam1 protein-protein interaction: An alternative approach for cardiovascular diseases. Pept. Sci. *110*, e23089. 10.1002/bip.23089.

139. Stevers, L.M., Sijbesma, E., Botta, M., MacKintosh, C., Obsil, T., Landrieu, I., Cau, Y., Wilson, A.J., Karawajczyk, A., Eickhoff, J., et al. (2018). Modulators of 14-3-3 Protein–Protein Interactions. J. Med. Chem. *61*, 3755–3778. 10.1021/acs.jmedchem.7b00574.

140. Azzarito, V., Long, K., Murphy, N.S., and Wilson, A.J. (2013). Inhibition of α-helix-mediated protein–protein interactions using designed molecules. Nat. Chem. *5*, 161–173. 10.1038/nchem.1568.

141. Khalil, A.S., Lu, T.K., Bashor, C.J., Ramirez, C.L., Pyenson, N.C., Joung, J.K., and Collins, J.J. (2012). A Synthetic Biology Framework for Programming Eukaryotic Transcription Functions. Cell *150*, 647–658. 10.1016/j.cell.2012.05.045.

142. Lebar, T., Bezeljak, U., Golob, A., Jerala, M., Kadunc, L., Pirš, B., Stražar, M., Vučko, D., Zupančič, U., Benčina, M., et al. (2014). A bistable genetic switch based on designable DNA-binding domains. Nat. Commun. *5*, 5007. 10.1038/ncomms6007.

143. Purcell, O., Peccoud, J., and Lu, T.K. (2014). Rule-Based Design of Synthetic Transcription Factors in Eukaryotes. ACS Synth. Biol. *3*, 737–744. 10.1021/sb400134k.

144. Brödel, A.K., Jaramillo, A., and Isalan, M. (2016). Engineering orthogonal dual transcription factors for multi-input synthetic promoters. Nat. Commun. *7*, 13858. 10.1038/ncomms13858.

145. Chen, Y., Zhang, S., Young, E.M., Jones, T.S., Densmore, D., and Voigt, C.A. (2020). Genetic circuit design automation for yeast. Nat. Microbiol. *5*, 1349–1360. 10.1038/s41564-020-0757-2.

146. Chen, W.C.W., Gaidukov, L., Lai, Y., Wu, M.-R., Cao, J., Gutbrod, M.J., Choi, G.C.G., Utomo, R.P., Chen, Y.-C., Wroblewska, L., et al. (2022). A synthetic transcription platform for programmable gene expression in mammalian cells. Nat. Commun. *13*, 6167. 10.1038/s41467-022-33287-9.

147. Barnea, G., Strapps, W., Herrada, G., Berman, Y., Ong, J., Kloss, B., Axel, R., and Lee, K.J. (2008). The genetic design of signaling cascades to record receptor activation. Proc. Natl. Acad. Sci. *105*, 64–69. 10.1073/pnas.0710487105.

148. Daringer, N.M., Dudek, R.M., Schwarz, K.A., and Leonard, J.N. (2014). Modular extracellular sensor architecture for engineering mammalian cell-based devices. ACS Synth. Biol. *3*, 892–902. 10.1021/sb400128g.

149. Morsut, L., Roybal, K.T., Xiong, X., Gordley, R.M., Coyle, S.M., Thomson, M., and Lim, W.A. (2016). Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. Cell *164*, 780–791. 10.1016/j.cell.2016.01.012.

150. Schwarz, K.A., Daringer, N.M., Dolberg, T.B., and Leonard, J.N. (2017). Rewiring human cellular input–output using modular extracellular sensors. Nat. Chem. Biol. *13*, 202–209. 10.1038/nchembio.2253.

151. Chung, H.K., Zou, X., Bajar, B.T., Brand, V.R., Huo, Y., Alcudia, J.F., Ferrell, J.E., and Lin, M.Z. (2019). A compact synthetic pathway rewires cancer signaling to therapeutic effector release. Science *364*, eaat6982. 10.1126/science.aat6982.

152. Manhas, J., Edelstein, H.I., Leonard, J.N., and Morsut, L. (2022). The evolution of synthetic receptor systems. Nat. Chem. Biol. *18*, 244–255. 10.1038/s41589-021-00926-z.

153. Edelstein, H.I., Donahue, P.S., Muldoon, J.J., Kang, A.K., Dolberg, T.B., Battaglia, L.M., Allchin, E.R., Hong, M., and Leonard, J.N. (2020). Elucidation and refinement of synthetic receptor mechanisms. Synth. Biol. *5*, ysaa017. 10.1093/synbio/ysaa017.

154. Shah, N.H., Eryilmaz, E., Cowburn, D., and Muir, T.W. (2013). Extein Residues Play an Intimate Role in the Rate-Limiting Step of Protein Trans-Splicing. J. Am. Chem. Soc. *135*, 5839–5847. 10.1021/ja401015p.

155. Vermeire, K., Bell, T.W., Puyenbroeck, V.V., Giraut, A., Noppen, S., Liekens, S., Schols, D., Hartmann, E., Kalies, K.-U., and Marsh, M. (2014). Signal Peptide-Binding Drug as a Selective Inhibitor of Co-Translational Protein Translocation. PLOS Biol. *12*, e1002011. 10.1371/journal.pbio.1002011.

156. Hlavaty, J., Schittmayer, M., Stracke, A., Jandl, G., Knapp, E., Felber, B.K., Salmons, B., Günzburg, W.H., and Renner, M. (2005). Effect of posttranscriptional regulatory elements on transgene expression and virus production in the context of retrovirus vectors. Virology *341*, 1–11. 10.1016/j.virol.2005.06.037.

157. Chatterjee, S., Min, L., Karuturi, R.K.M., and Lufkin, T. (2010). The role of post-transcriptional RNA processing and plasmid vector sequences on transient transgene expression in zebrafish. Transgenic Res. *19*, 299–304. 10.1007/s11248-009-9312-x.

158. Lange, A., Mills, R.E., Lange, C.J., Stewart, M., Devine, S.E., and Corbett, A.H. (2007). Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin α*. J. Biol. Chem. *282*, 5101–5105. 10.1074/jbc.R600026200.

159. Qin, J.Y., Zhang, L., Clift, K.L., Hulur, I., Xiang, A.P., Ren, B.-Z., and Lahn, B.T. (2010). Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. PLOS ONE *5*, e10611. 10.1371/journal.pone.0010611.

160. Wagstaff, K.M., and Jans, D.A. (2006). Intramolecular masking of nuclear localization signals: Analysis of importin binding using a novel AlphaScreen-based method. Anal. Biochem. *348*, 49–56. 10.1016/j.ab.2005.10.029.

161. Ganchi, P.A., Sun, S.C., Greene, W.C., and Ballard, D.W. (1992). I kappa B/MAD-3 masks the nuclear localization signal of NF-kappa B p65 and requires the transactivation domain to inhibit NF-kappa B p65 DNA binding. Mol. Biol. Cell *3*, 1339–1352. 10.1091/mbc.3.12.1339.

162. Latimer, M., Ernst, M.K., Dunn, L.L., Drutskaya, M., and Rice, N.R. (1998). The N-Terminal Domain of IκBα Masks the Nuclear Localization Signal(s) of p50 and c-Rel Homodimers. Mol. Cell. Biol. *18*, 2640–2649. 10.1128/MCB.18.5.2640.

163. Silhan, J., Vacha, P., Strnadova, P., Vecer, J., Herman, P., Sulc, M., Teisinger, J., Obsilova, V., and Obsil, T. (2009). 14-3-3 protein masks the DNA binding interface of forkhead transcription factor FOXO4. J. Biol. Chem. *284*, 19349–19360. 10.1074/jbc.M109.002725.

164. Chen, B.B., and Mallampalli, R.K. (2009). Masking of a nuclear signal motif by monoubiquitination leads to mislocalization and degradation of the regulatory enzyme cytidylyltransferase. Mol. Cell. Biol. *29*, 3062–3075. 10.1128/MCB.01824-08.

165. Meng, W., Swenson, L.L., Fitzgibbon, M.J., Hayakawa, K., ter Haar, E., Behrens, A.E., Fulghum, J.R., and Lippke, J.A. (2002). Structure of Mitogen-activated Protein Kinase-activated Protein (MAPKAP) Kinase 2 Suggests a Bifunctional Switch That Couples Kinase Activation with Nuclear Export*. J. Biol. Chem. *277*, 37401–37405. 10.1074/jbc.C200418200.

166. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. 10.1038/s41592-022-01488-1.

167. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. (2021). Protein complex prediction with AlphaFold-Multimer (Bioinformatics) 10.1101/2021.10.04.463034.

168. Pieters, B.J.G.E., Eldijk, M.B. van, Nolte, R.J.M., and Mecinović, J. (2015). Natural supramolecular protein assemblies. Chem. Soc. Rev. *45*, 24–39. 10.1039/C5CS00157A.

169. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci. Rep. *7*, 10480. 10.1038/s41598-017-09654-8.

170. Gwyther, R.E.A., Jones, D.D., and Worthy, H.L. (2019). Better together: building protein oligomers naturally and by design. Biochem. Soc. Trans. *47*, 1773–1780. 10.1042/BST20190283.

171. Bergendahl, L.T., Gerasimavicius, L., Miles, J., Macdonald, L., Wells, J.N., Welburn, J.P.I., and Marsh, J.A. (2019). The role of protein complexes in human genetic disease. Protein Sci. *28*, 1400–1411. 10.1002/pro.3667.

172. Kayed, R., Dettmer, U., and Lesné, S.E. (2020). Soluble endogenous oligomeric α-synuclein species in neurodegenerative diseases: Expression, spreading, and cross-talk. J. Park. Dis. *10*, 791–818. 10.3233/JPD-201965.

173. Hernández, F., Ferrer, I., Pérez, M., Zabala, J.C., del Rio, J.A., and Avila, J. (2022). Tau Aggregation. Neuroscience. 10.1016/j.neuroscience.2022.04.024.

174. Plesa, C., Sidore, A.M., Lubock, N.B., Zhang, D., and Kosuri, S. (2018). Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. Science *359*, 343–347. 10.1126/science.aao5167.

175. Sidore, A.M., Plesa, C., Samson, J.A., Lubock, N.B., and Kosuri, S. (2020). DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. Nucleic Acids Res. *48*, e95. 10.1093/nar/gkaa600.

176. Scharnagl, C., Reif, M., and Friedrich, J. (2005). Stability of proteins: Temperature, pressure and the role of the solvent. Biochim. Biophys. Acta BBA - Proteins Proteomics *1749*, 187–213. 10.1016/j.bbapap.2005.03.002.

177. Roche, J., and Royer, C.A. (2018). Lessons from pressure denaturation of proteins. J. R. Soc. Interface *15*, 20180244. https://doi.org/10.1098/rsif.2018.0244.

178. Abe, F. (2021). Molecular Responses to High Hydrostatic Pressure in Eukaryotes: Genetic Insights from Studies on Saccharomyces cerevisiae. Biology *10*, 1305. 10.3390/biology10121305.

179. Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J.J., Mangan, N.M., Ovchinnikov, S., and Rocklin, G.J. (2022). Mega-scale experimental analysis of protein folding stability in biology and protein design (Biophysics) 10.1101/2022.12.06.519132.

180. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. 10.1038/s41586-021-03828-1.

181. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network.

182. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. 2022.07.20.500902. 10.1101/2022.07.20.500902.

183. Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. (2022). High-resolution de novo structure prediction from primary sequence. 2022.07.21.500999. 10.1101/2022.07.21.500999.

184. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N., et al. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science *378*, 49–56. 10.1126/science.add2187.

185. Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., et al. (2022). Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. 2022.12.09.519842. 10.1101/2022.12.09.519842.

186. Torres, S.V., Leung, P.J.Y., Lutz, I.D., Venkatesh, P., Watson, J.L., Hink, F., Huynh, H.-H., Yeh, A.H.-W., Juergens, D., Bennett, N.R., et al. (2022). De novo design of high-affinity protein binders to bioactive helical peptides. 2022.12.10.519862. 10.1101/2022.12.10.519862.

187. Bennett, N., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y.P., Dauparas, J., Baek, M., Stewart, L., et al. (2022). Improving de novo Protein Binder Design with Deep Learning. 2022.06.15.495993. 10.1101/2022.06.15.495993.

188. Pastore, A., Martin, S.R., and Temussi, P.A. (2019). Generalized View of Protein Folding: In Medio Stat Virtus. J. Am. Chem. Soc. *141*, 2194–2200. 10.1021/jacs.8b10779.

189. Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604. 10.1016/j.sbi.2009.08.003.

190. Appleby, J.H., Zhou, K., Volkmann, G., and Liu, X.-Q. (2009). Novel Split Intein for trans-Splicing Synthetic Peptide onto C Terminus of Protein. J. Biol. Chem. *284*, 6194–6199. 10.1074/jbc.M805474200.

191. Volkmann, G., and Liu, X.-Q. (2011). Intein lacking conserved C-terminal motif G retains controllable N-cleavage activity. FEBS J. *278*, 3431–3446. 10.1111/j.1742-4658.2011.08266.x.

192. Lin, Y., Li, M., Song, H., Xu, L., Meng, Q., and Liu, X.-Q. (2013). Protein Trans-Splicing of Multiple Atypical Split Inteins Engineered from Natural Inteins. PLoS ONE *8*, e59516. 10.1371/journal.pone.0059516.

193. Li, X., Zhang, L., Wang, S., Liu, X., and Lin, Y. (2021). Site-specific internal protein labeling through trans-splicing. Int. J. Biol. Macromol. *186*, 40–46. 10.1016/j.ijbiomac.2021.07.009.

194. Cheriyan, M., Pedamallu, C.S., Tori, K., and Perler, F. (2013). Faster Protein Splicing with the Nostoc punctiforme DnaE Intein Using Non-native Extein Residues. J. Biol. Chem. *288*, 6202–6211. 10.1074/jbc.M112.433094.

195. Lockless, S.W., and Muir, T.W. (2009). Traceless protein splicing utilizing evolved split inteins. Proc. Natl. Acad. Sci. *106*, 10999–11004. 10.1073/pnas.0902964106.

196. Shah, N.H., Dann, G.P., Vila-Perelló, M., Liu, Z., and Muir, T.W. (2012). Ultrafast Protein Splicing is Common among Cyanobacterial Split Inteins: Implications for Protein Engineering. J. Am. Chem. Soc. *134*, 11338–11341. 10.1021/ja303226x.

197. Ceballos-Alcantarilla, E., and Merkx, M. (2021). Chapter One - Understanding and applications of Ser/Gly linkers in protein engineering. In Methods in Enzymology Linkers in Biomacromolecules., M. Merkx, ed. (Academic Press), pp. 1–22. 10.1016/bs.mie.2020.12.001.

198. Shcherbo, D., Murphy, C.S., Ermakova, G.V., Solovieva, E.A., Chepurnykh, T.V., Shcheglov, A.S., Verkhusha, V.V., Pletnev, V.Z., Hazelwood, K.L., Roche, P.M., et al. (2009). Far-red fluorescent tags for protein imaging in living tissues. Biochem. J. *418*, 567–574. 10.1042/BJ20081949.

199. Gadella, T.W.J., van Weeren, L., Stouthamer, J., Hink, M.A., Wolters, A.H.G., Giepmans, B.N.G., Aumonier, S., Dupuy, J., and Royant, A. (2023). mScarlet3: a brilliant and fast-maturing red fluorescent protein. Nat. Methods *20*, 541–545. 10.1038/s41592-023-01809-y.

200. Balleza, E., Kim, J.M., and Cluzel, P. (2018). Systematic characterization of maturation time of fluorescent proteins in living cells. Nat. Methods *15*, 47–51. 10.1038/nmeth.4509.

201. Bogert, N.V., Furkel, J., Din, S., Braren, I., Eckstein, V., Müller, J.A., Uhlmann, L., Katus, H.A., and Konstandin, M.H. (2020). A novel approach to genetic engineering of T-cell subsets by hematopoietic stem cell infection with a bicistronic lentivirus. Sci. Rep. *10*, 13740. 10.1038/s41598-020-70793-6.

202. Duportet, X., Wroblewska, L., Guye, P., Li, Y., Eyquem, J., Rieders, J., Rimchala, T., Batt, G., and Weiss, R. (2014). A platform for rapid prototyping of synthetic gene networks in mammalian cells. Nucleic Acids Res. *42*, 13440–13451. 10.1093/nar/gku1082.

203. Maya, D., Quintero, M.J., de la Cruz Muñoz-Centeno, M., and Chávez, S. (2008). Systems for applied gene control in Saccharomyces cerevisiae. Biotechnol. Lett. *30*, 979–987. 10.1007/s10529-008-9647-z.

204. Kotopka, B.J., and Smolke, C.D. (2020). Model-driven generation of artificial yeast promoters. Nat. Commun. *11*, 2113. 10.1038/s41467-020-15977-4.

205. Buist, G., Steen, A., Kok, J., and Kuipers, O.P. (2008). LysM, a widely distributed protein motif for binding to (peptido)glycans. Mol. Microbiol. *68*, 838–847. 10.1111/j.1365-2958.2008.06211.x.

206. Yeats, C., Finn, R.D., and Bateman, A. (2002). The PASTA domain: a β-lactam-binding domain. Trends Biochem. Sci. *27*, 438–440. 10.1016/S0968-0004(02)02164-3.

207. Graumann, P.L., and Marahiel, M.A. (1998). A superfamily of proteins that contain the cold-shock domain. Trends Biochem. Sci. *23*, 286–290. 10.1016/S0968-0004(98)01255-9.

208. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In, pp. 92–96. 10.25080/Majora-92bf1922-011.

209. Jones, D.T. (1999). Protein Secondary Structure Prediction Based on Position-specific scoring matrices. J. Mol. Biol. *292*, 195–202. 10.1006/jmbi.1999.3091.

APPENDIX 1

The figures in Appendix is a reprint of the Supplementary Information in:

> Kim, T.E. et al. Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. *Proceedings of the National Academy of Sciences*. **119** (41), e2122676119. 2022

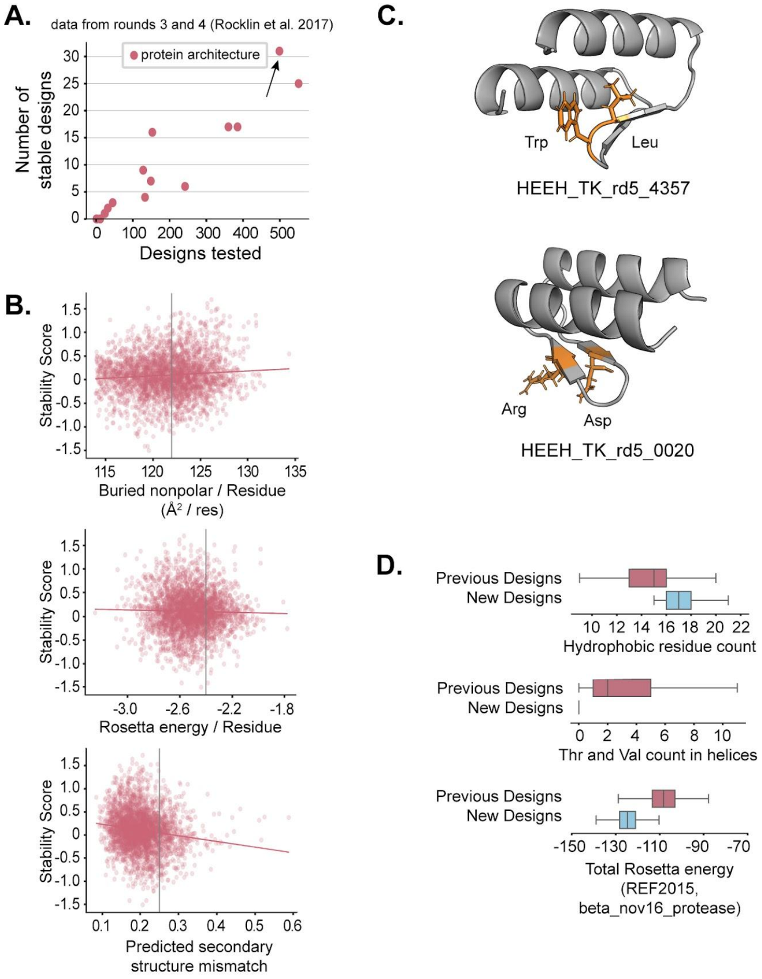


Fig. S1. Restricted αββα miniprotein design strategy. (A) We selected the protein architecture that led to generating the greatest number of stable designs (black arrow) from our previous study.[37] Here, we set the threshold of stability as a design having a stability score ≥ 0.8 (B) Based on the Round 3 and 4 αββα design and stability data from our previous study,[37] we set thresholds for specific features that a design can have: buried nonpolar surface area / residue > 122 (top), Rosetta energy / residue < -2.4 (middle), and predicted secondary structure mismatch < 0.25 (bottom); vertical gray line denotes the minimum or maximum threshold, and red line denotes best fit line. (C) We also forced all designs to have one hydrophobic residue in the middle loop (top) and polar/charged residues at solvent-facing β-strand positions (bottom); example residues are highlighted (orange) on a cartoon model. (D) We also set additional constraints to the αββα designs (blue) in comparison to our previous study:[37] 15-21 nonpolar residues allowed (top) and no Thr/Val in helices (middle); using these restrictions, we were able to generate αββα designs whose Rosetta energies were lower than previously designed (bottom).

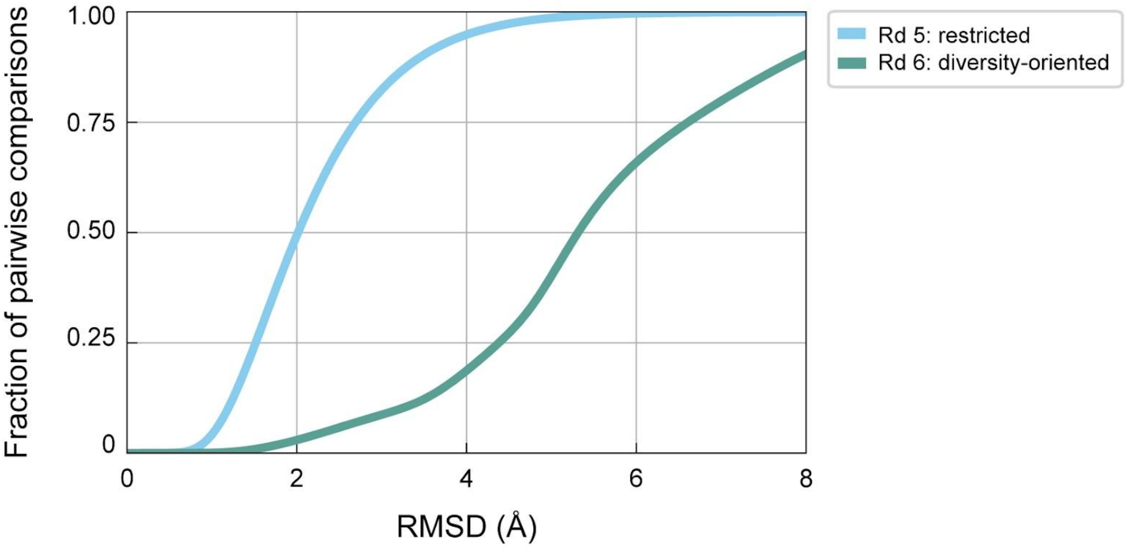


Fig. S2. Structural diversity of restricted and diversity-oriented αββα miniprotein designs. Each miniprotein design was compared to every design within the same library (Round 5 or Round 6) by the distance between alpha-carbons.

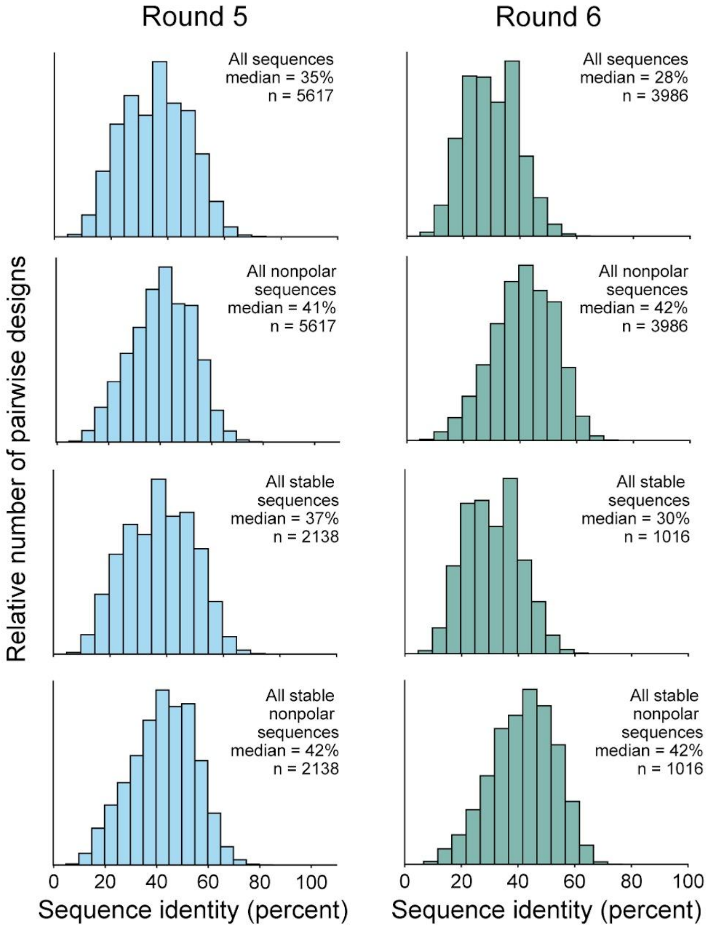


Fig S3. Sequence identity of designs. To evaluate the sequence diversity among Round 5 (restricted) and Round 6 (diversity-oriented) designs, we calculated the sequence identity (all sequences or all nonpolar sequences only) between all possible pairs of designs (all or stable). The distributions of the sequence identities are shown as histograms, with bins being 5 percentage points wide.

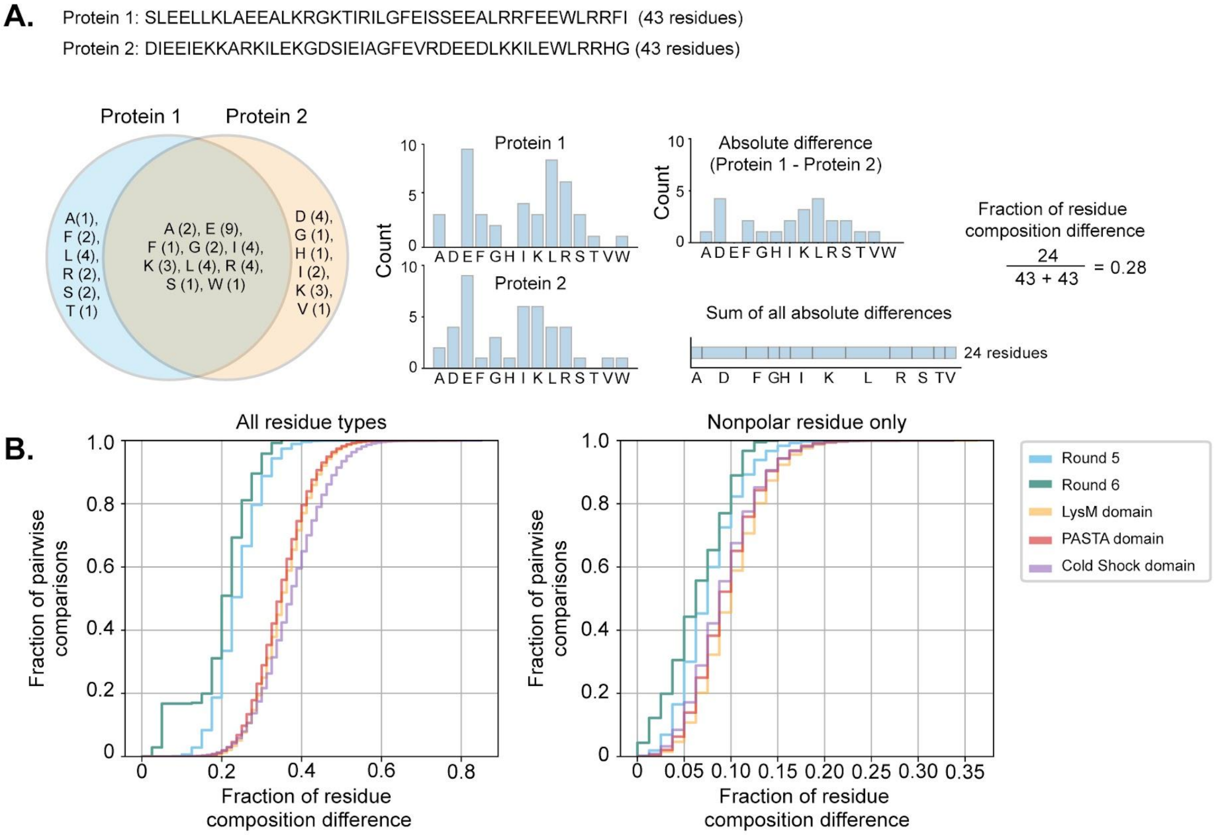


Fig S4. Residue composition of αββα miniprotein designs in comparison to similarly-sized protein domains. To compare the compositional diversity of αββα miniproteins with known protein domains of similar sizes (LysM: 44-65 residues,[205] PASTA: ~70 residues,[206] and Cold Shock: 65-70 residues [207]), we calculated (A) how different a pair of protein sequences are to each other by identifying the total number of unique residues for each protein sequence, taking the absolute difference for each number of unique residue, summing the absolute difference, and calculating the fraction of the sum over the total number of possible residues. (B) The fraction of pairwise comparisons for all possible protein pairs in Round 5, Round 6, and three natural protein domains are shown as a function of the fraction of residue difference (all residue types, left; nonpolar residues only, right).

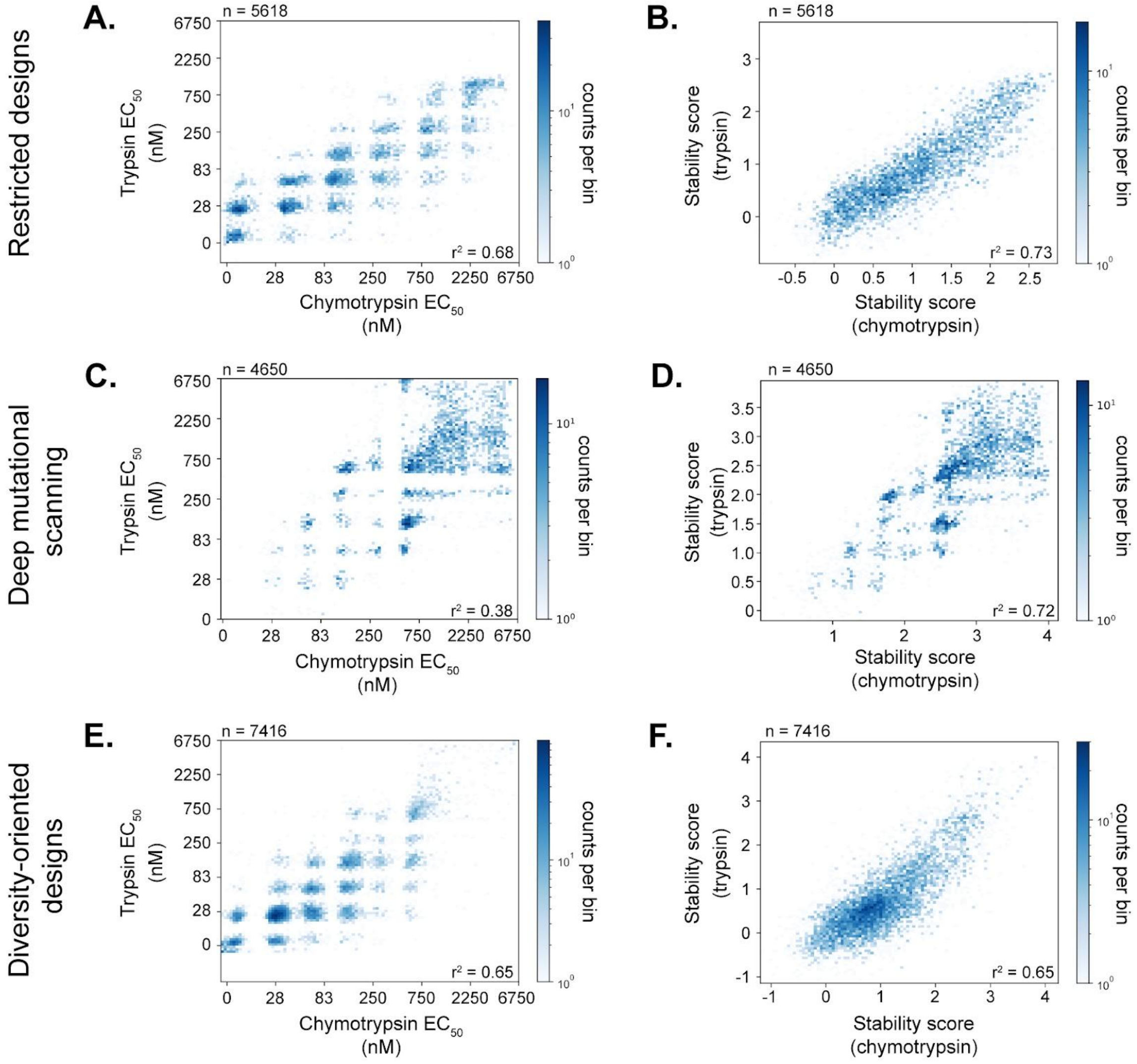


Fig. S5. Protease stability assay using trypsin and chymotrypsin. αββα miniproteins generated from (A-B) a restricted design strategy, (C-D) deep mutational scanning, and (E-F) a diversity-oriented strategy were displayed on the surface of yeast cells and subject to varying concentrations of either trypsin of chymotrypsin (from 0 to 6750 nM). EC$_{50}$ values (left) and calculated stability scores (right) at each protease concentration are depicted as 2D-histograms.

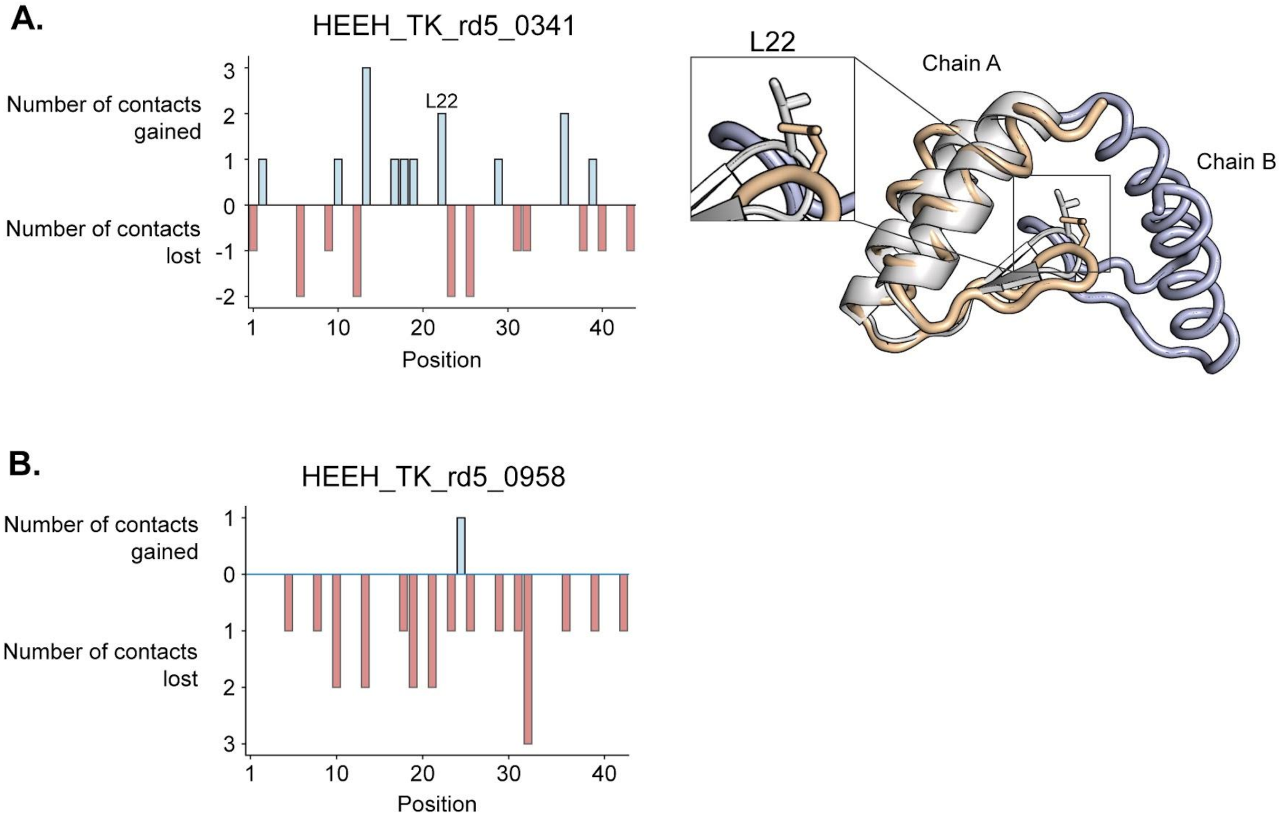


Fig. S6. Structural agreement between Rosetta design model and NMR ensemble for HEEH_TK_rd5_0341 and HEEH_TK_rd5_0958. (A-B) We determined the Euclidean distance between all possible residue-residue pairs (using the beta-carbon positions for each residue, excluding Gly) in the design model and an NMR model (whose beta-carbon positions after superposition were the average of an NMR ensemble consisting of twenty structures for each chain in HEEH_TK_rd5_0341 and twenty structures for HEEH_TK_rd5_0958). From the distance calculations, we quantified the number of contacts (< 8 Å) a residue in each position gained or lost from the design model to the NMR model. A cartoon of HEEH_TK_rd5_0341 design model (gray) is overlaid on Chain A (orange) of the NMR model.

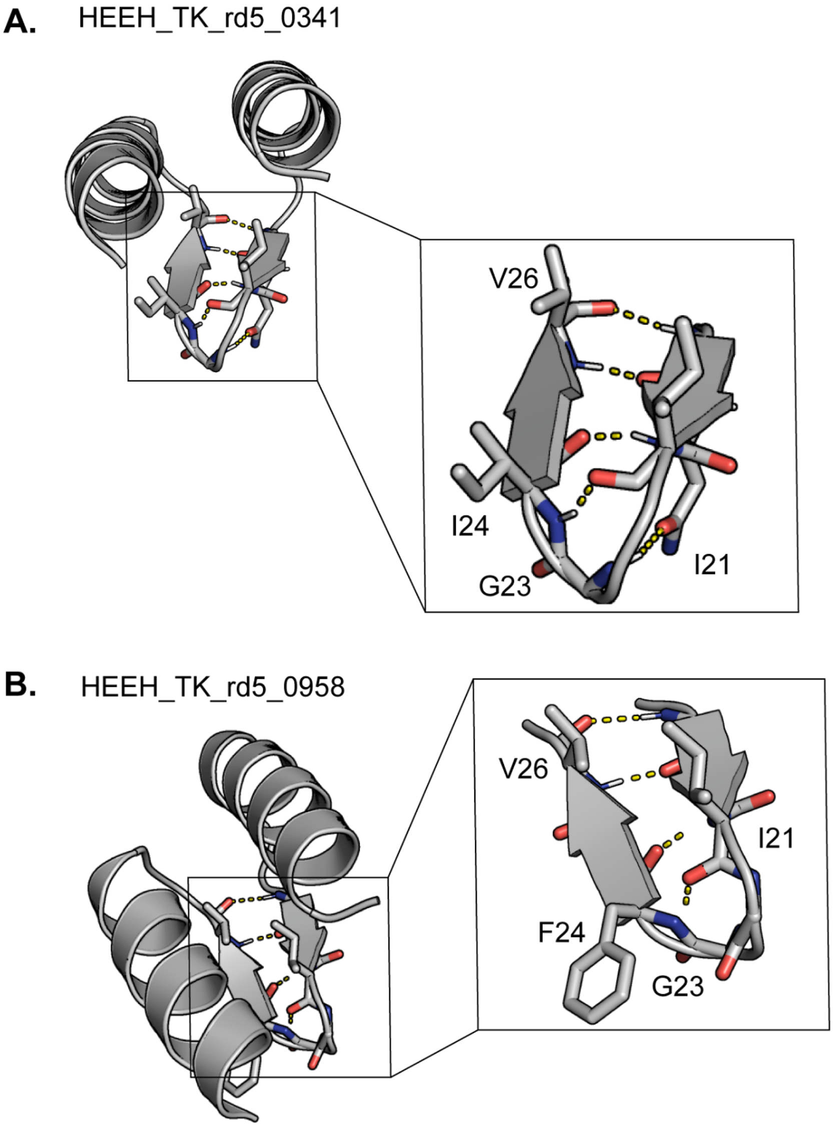
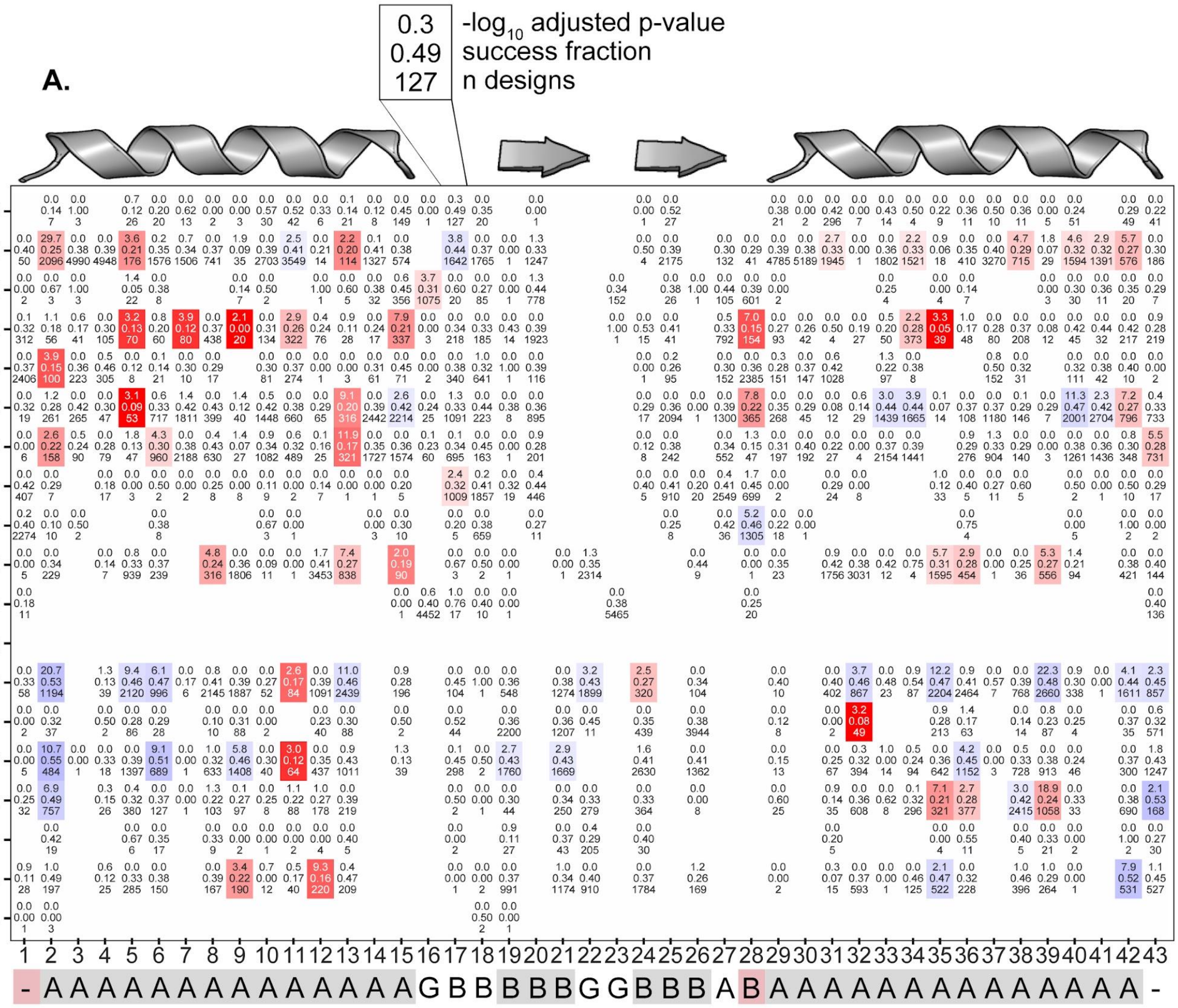
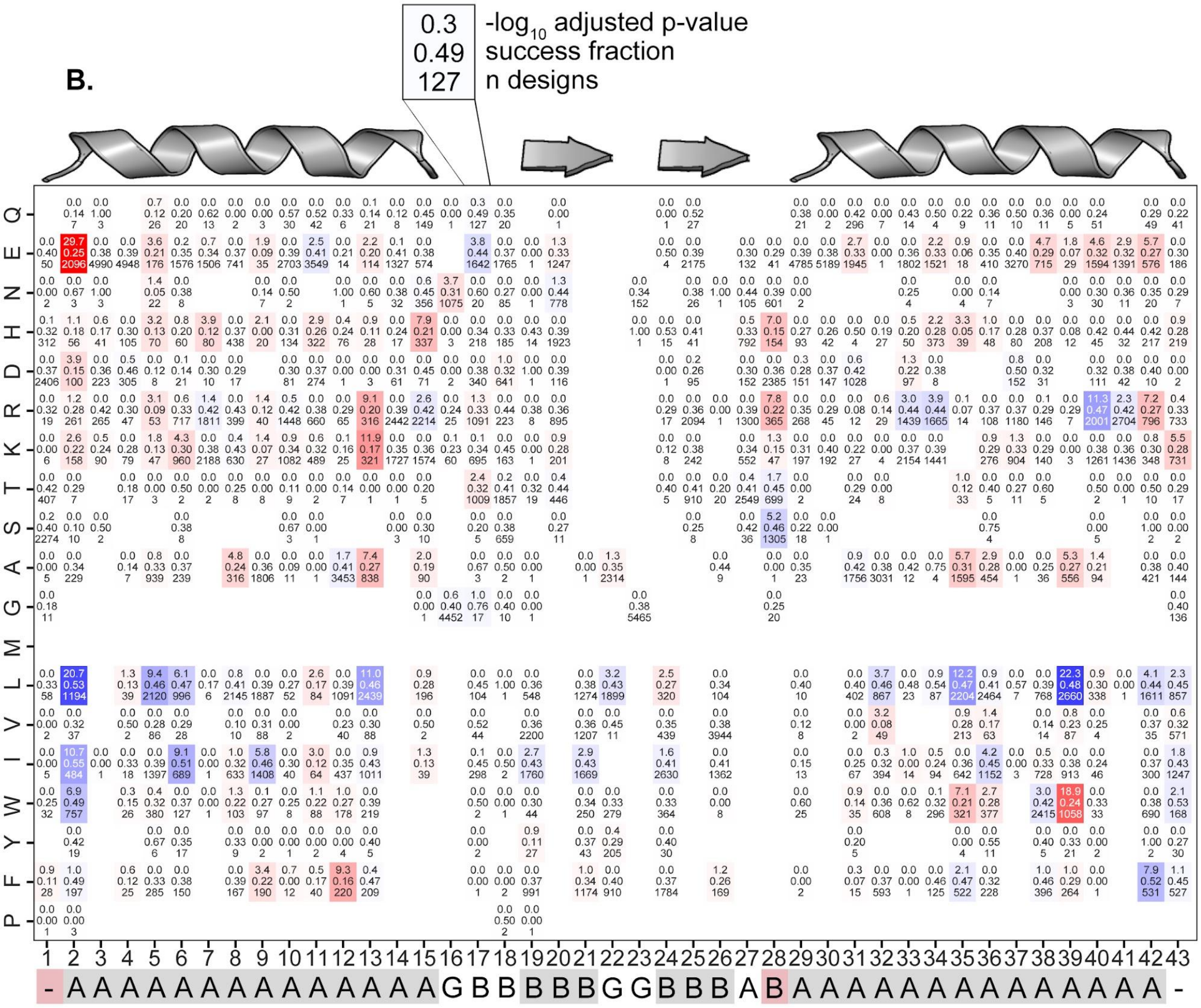


Fig. S7. Intramolecular hydrogen bonds in the β-hairpin of HEEH_TK_rd5_0341 and HEEH_TK_rd5_0958. Cartoons of (A) HEEH_TK_rd5_0341 and (B) HEEH_TK_rd5_0958 highlighting the hydrogen bonds (yellow) that are formed within the β-hairpin.


A.
0.3 -log10 adjusted p-value
0.49 success fraction
127 n designs
Q
E
N
H
D
R
K
T
S
A
G
M
L
V
I
W
Y
F
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
- A A A A A A A A A A A A A A G B B B B B G G B B B A B A A A A A A A A A A A A A A -


B.
0.3
0.49
127
-log$_{10}$ adjusted p-value
success fraction
n designs
Q
E
N
H
D
R
K
T
S
A
G
M
L
V
I
W
Y
F
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
- A A A A A A A A A A A A A A G B B B B B G G B B B A B A A A A A A A A A A A A A A -

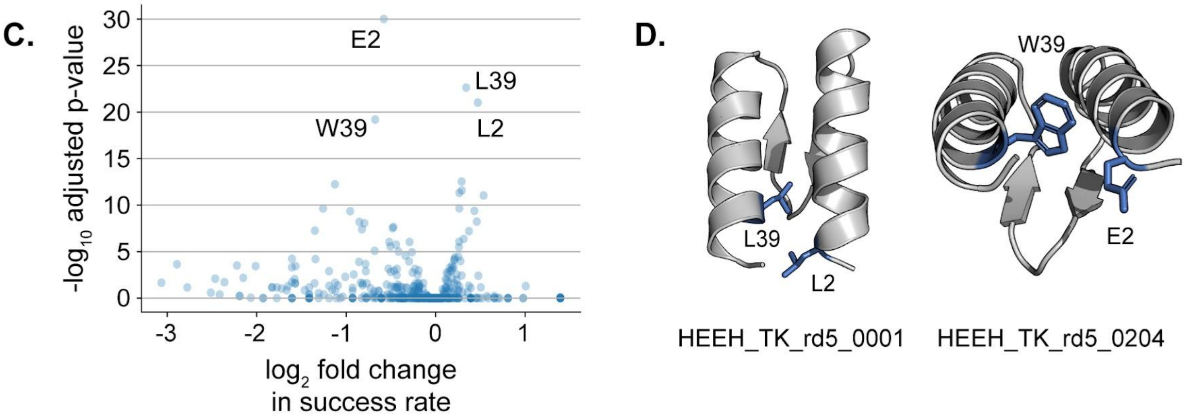
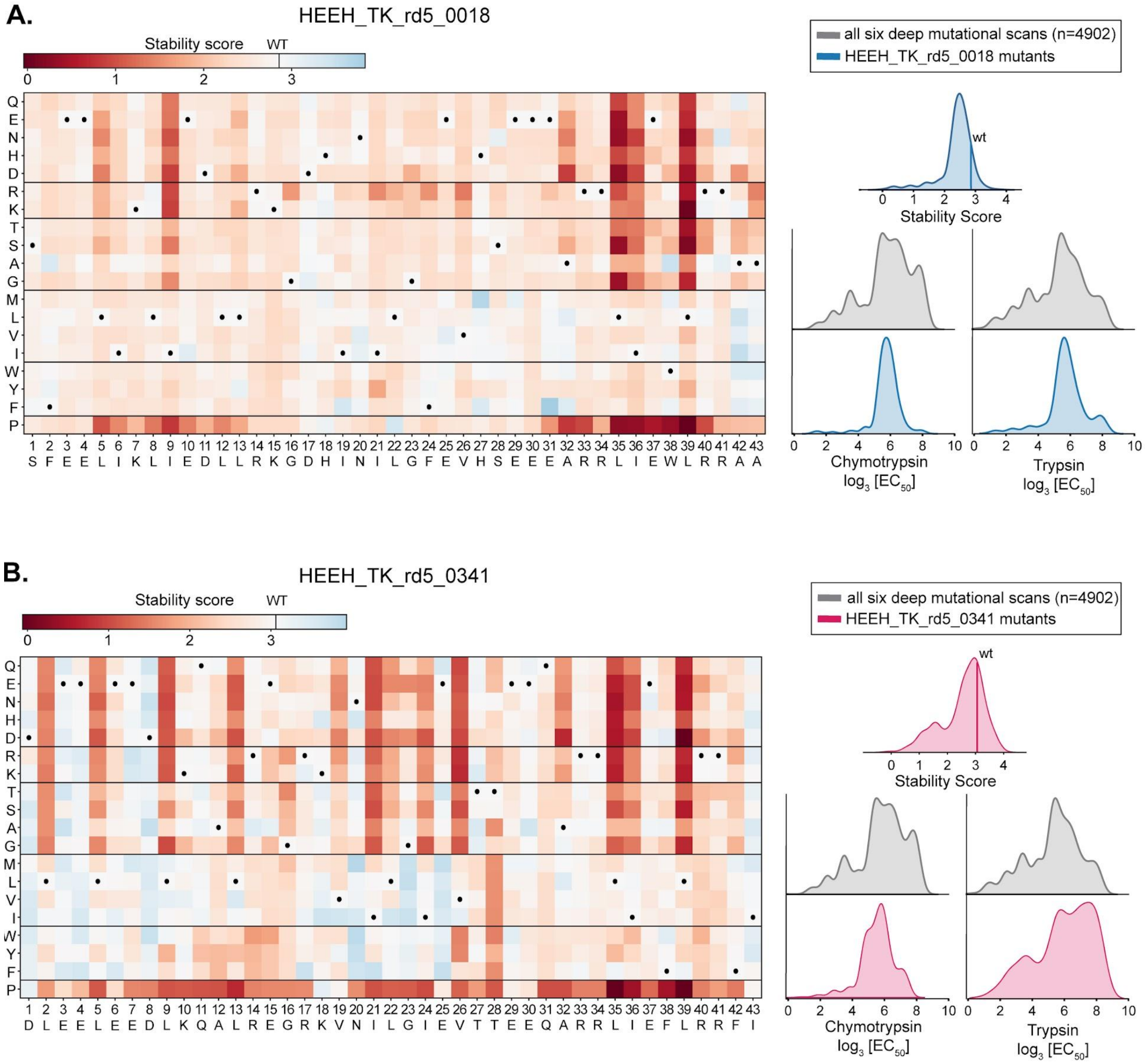
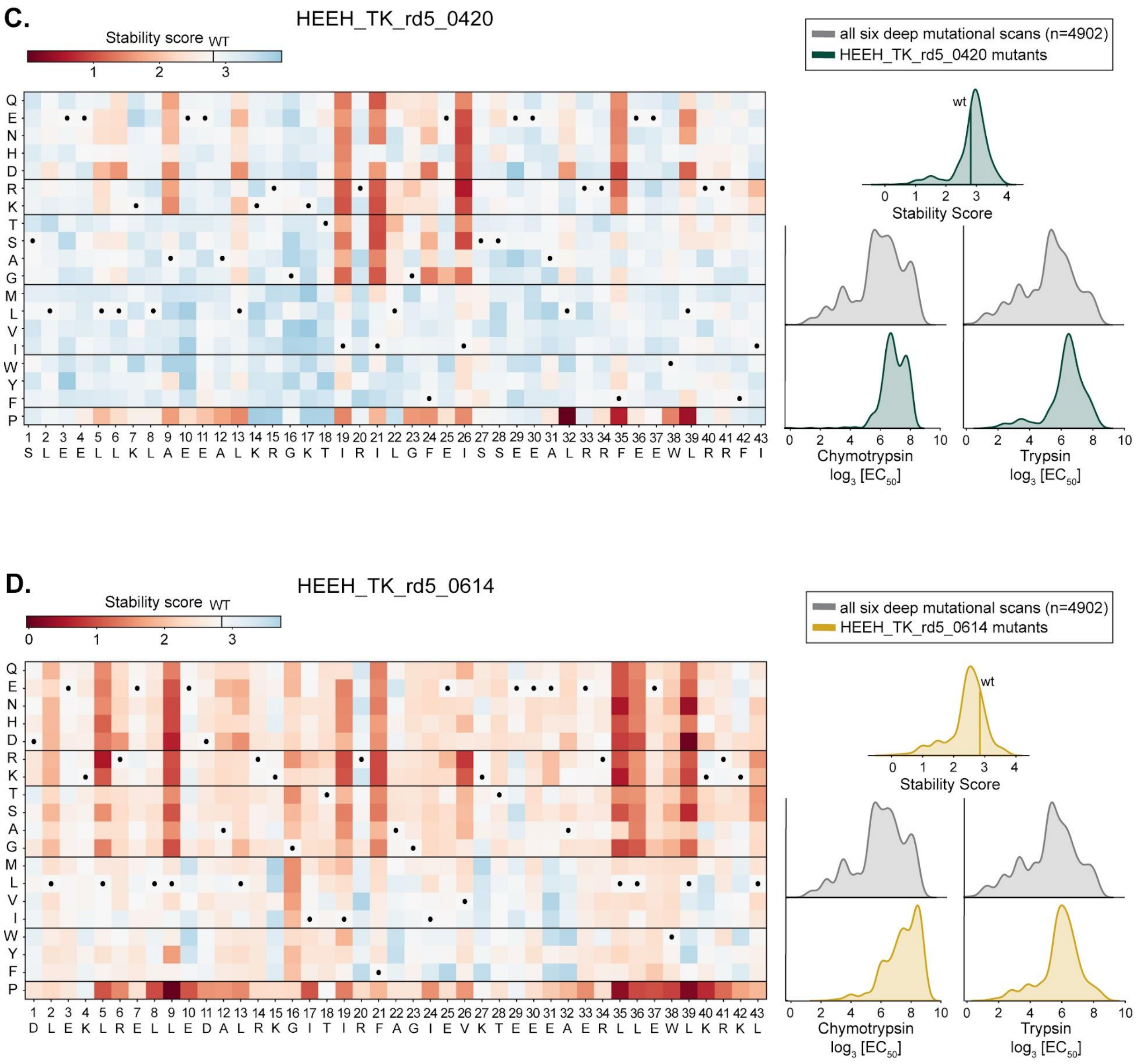
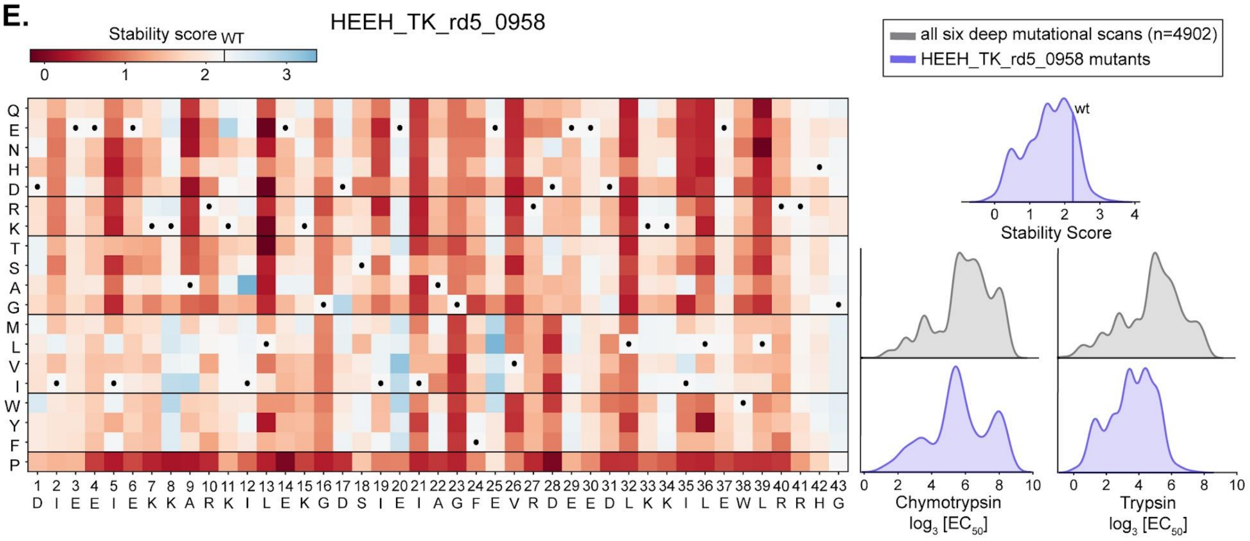


Fig. S8. Contribution of specific residues on folding stability. For αββα miniproteins made using a restricted design strategy (Round 5), we performed binomial tests for all possible residues in all 43 positions to examine whether the success rate of designs containing that residue differed from the overall success rate (0.38). All p-values were adjusted for multiple testing using the Benjamini–Yekutieli procedure as implemented in [208] . (A-B) The p-value, success rate, and number of designs containing each amino acid at each position. In A, significant residues (adjusted p-value < 0.01) are colored according to the fold-change in success rate, with favorable amino acids in blue and unfavorable amino acids in red. In B, residues are colored by statistical significance. The ABEGO pattern of the Round 5 architecture is shown below; secondary structure is colored in gray and helix caps are shown in red. (C) Volcano plot indicating the p-value and change in success rate associated with each amino acid at each position. The most significant residue-positions are labeled on the plot (W39, E2, L39, L2). (D) Two representative αββα cartoons visualizing L2 and L39 (left) and E2 and W39 (right).


A.
HEEH_TK_rd5_0018
Stability score
WT
0
1
2
3
Q
E
N
H
D
R
K
T
S
A
G
M
L
V
I
W
Y
F
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
S F E E L I K L I E D L L R K G D H I N I L G F E V H S E E E A R R L I E W L R R A A
all six deep mutational scans (n=4902)
HEEH_TK_rd5_0018 mutants
wt
0 1 2 3 4
Stability Score
0 2 4 6 8 10
Chymotrypsin
log3 [EC50]
0 2 4 6 8 10
Trypsin
log3 [EC50]
B.
HEEH_TK_rd5_0341
Stability score
WT
0
1
2
3
Q
E
N
H
D
R
K
T
S
A
G
M
L
V
I
W
Y
F
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
D L E E L E E D L K Q A L R E G R K V N I L G I E V T T E E Q A R R L I E F L R R F I
all six deep mutational scans (n=4902)
HEEH_TK_rd5_0341 mutants
wt
0 1 2 3 4
Stability Score
0 2 4 6 8 10
Chymotrypsin
log3 [EC50]
0 2 4 6 8 10
Trypsin
log3 [EC50]


C.
HEEH_TK_rd5_0420
Stability score WT
1 2 3
Q E N H D R K T S A G M L V I W Y F P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
S L E E L L K L A E E A L K R G K T I R I L G F E I S S E E A L R R F E E W L R R F I
all six deep mutational scans (n=4902)
HEEH_TK_rd5_0420 mutants
wt
0 1 2 3 4
Stability Score
0 2 4 6 8 10
Chymotrypsin
log3 [EC50]
0 2 4 6 8 10
Trypsin
log3 [EC50]
D.
HEEH_TK_rd5_0614
Stability score WT
0 1 2 3
Q E N H D R K T S A G M L V I W Y F P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
D L E K L R E L L E D A L R K G I T I R F A G I E V K T E E E A E R L L E W L K R K L
all six deep mutational scans (n=4902)
HEEH_TK_rd5_0614 mutants
wt
0 1 2 3 4
Stability Score
0 2 4 6 8 10
Chymotrypsin
log3 [EC50]
0 2 4 6 8 10
Trypsin
log3 [EC50]

E.



F.

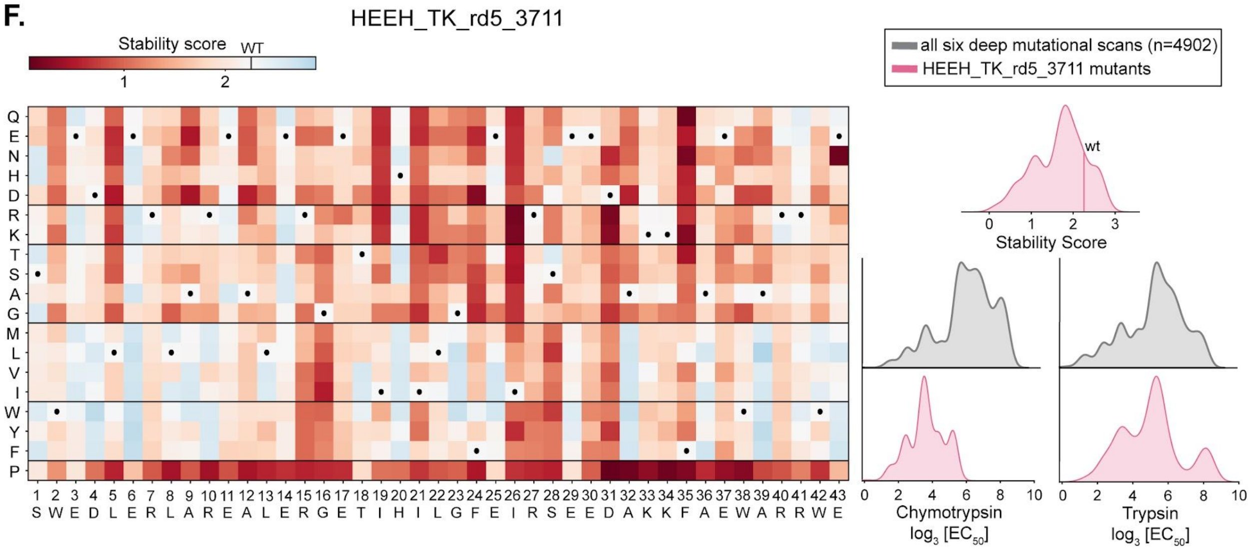


Fig. S9. Stability scores of six αββα miniproteins by deep mutational scanning. For six designs (Table 1) we created a library consisting of all possible single mutants and tested for their folding stability by a yeast display-based protease sensitivity assay.[37] (Left) Results are shown as heatmaps. The wildtype residue is represented as a black dot, and the stability scores of each mutant relative to the wildtype is shown as a range from red (less stable than the wild type) to blue (more stable than wildtype). (Right, top) The distribution of stability scores for each design, (right, middle), EC_{50} concentrations of protease for all six deep mutational scans, and (right, bottom) for each design are all shown as histograms. Mutants for five of the six designs were more likely to destabilize the protein when compared to their corresponding wild type (A-B, D-F). However, this is the opposite for HEEH_TK_rd5_0420 (C). Because the typical mutant in HEEH_TK_rd5_0420 does not seem to have particularly unique properties when compared to the other five designs, it may be that the high EC_{50} values (e.g. for

chymotrypsin) prevents us from discriminating the true differences between the wild type and mutant stabilities.

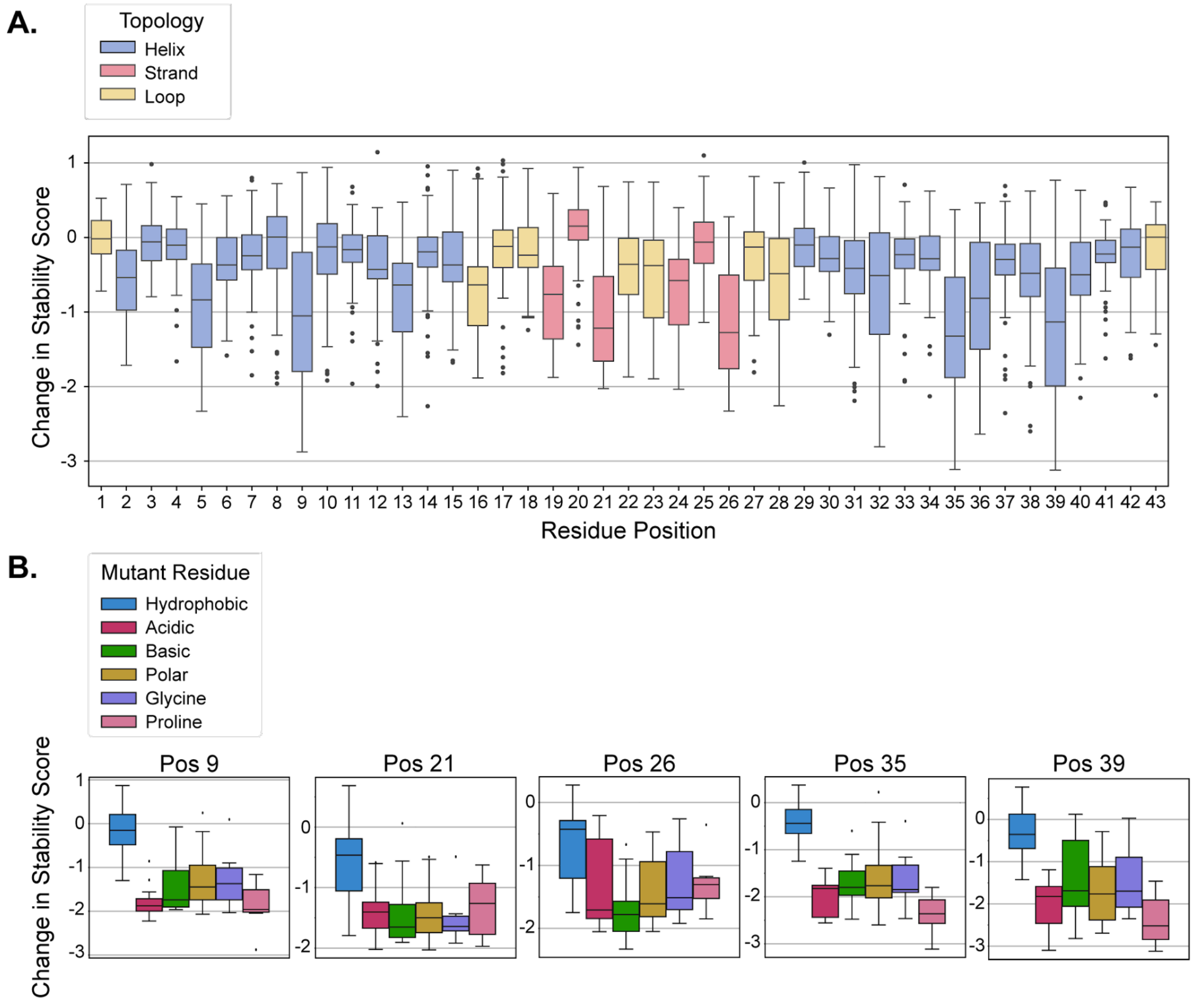


Fig. S10. Impact of mutant residue type and the position of the mutant on stability. To analyze whether certain mutations at specific positions on the αββα miniprotein are more likely to destabilize the structure, we (A) showed the distributions of the change in stability score (mutant stability - wild type stability) at each position of the six miniproteins of which we have deep mutational scanning data (Fig. S9). The top five destabilizing positions (9, 21, 26, 35, and 39) have an average change in stability score < 1 and are located in the buried core. For these positions (B) the change in stability score is broken down and shown by the type of mutant residue.

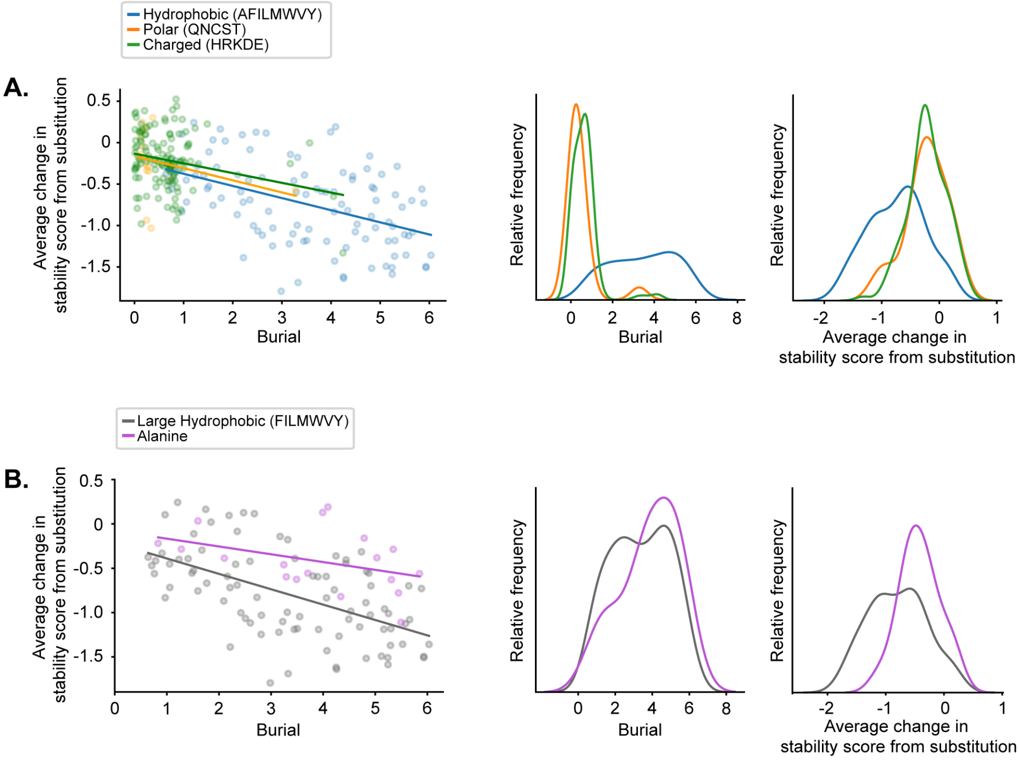


Fig S11. Sensitivity to mutation among buried and exposed residues. To evaluate the distribution of mutational sensitivity among buried and solvent-exposed residues, we computed the extent to which a residue in the wildtype miniprotein (that was tested in the deep mutational scanning, Fig. S9) is buried in the miniprotein (higher values indicate buried in the core, lower values indicate exposed to solvent). We compared each residue's burial value to its mutational sensitivity (average change in stability score from substitution). More negative values of the average change in stability indicate greater destabilization of the miniprotein. (A) Hydrophobic, polar, and charged residues are colored in blue, orange, and green, respectively; (B) Large hydrophobic and Alanine (small) are colored in gray and purple, respectively. The burial for a given residue was computed from the backbone coordinates of the designed structure by summing the number of CA atoms in a cone projecting out from the residue's CA-CB vector. The script is provided in https://github.com/kimte1/abba_protein_stability_manuscript.

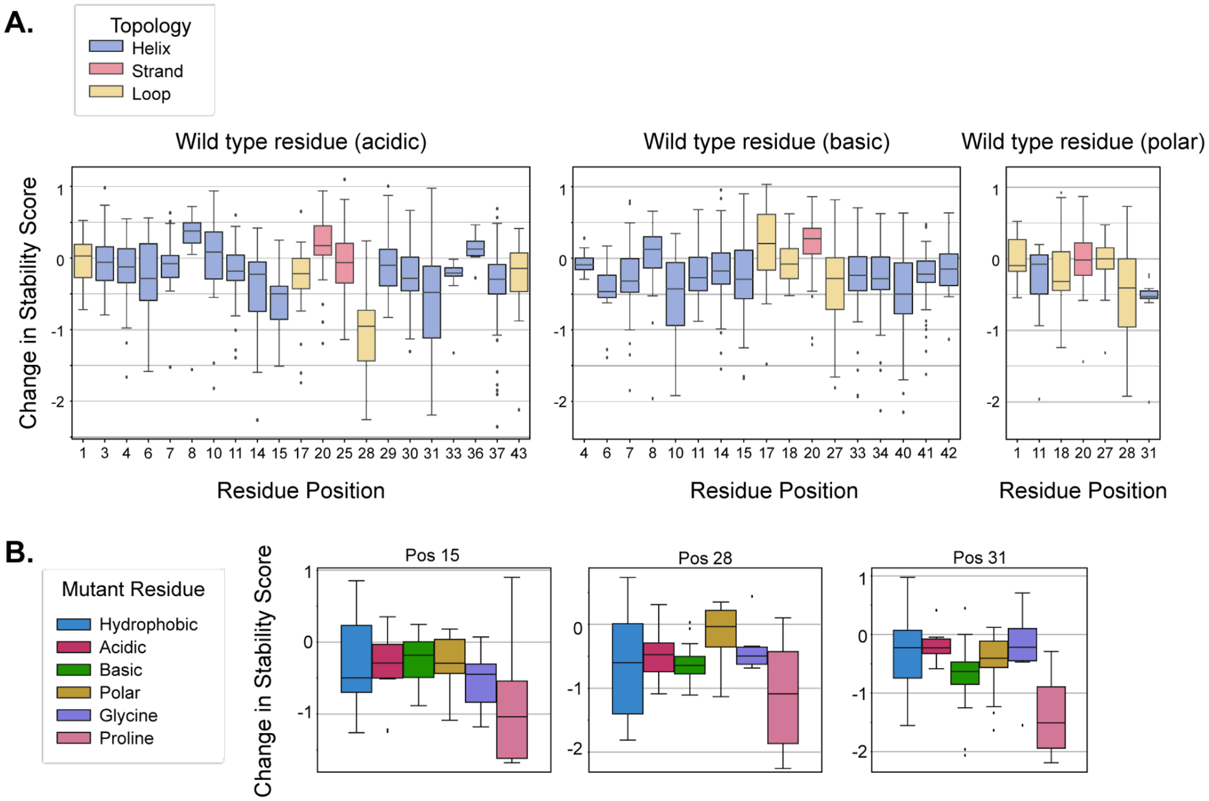
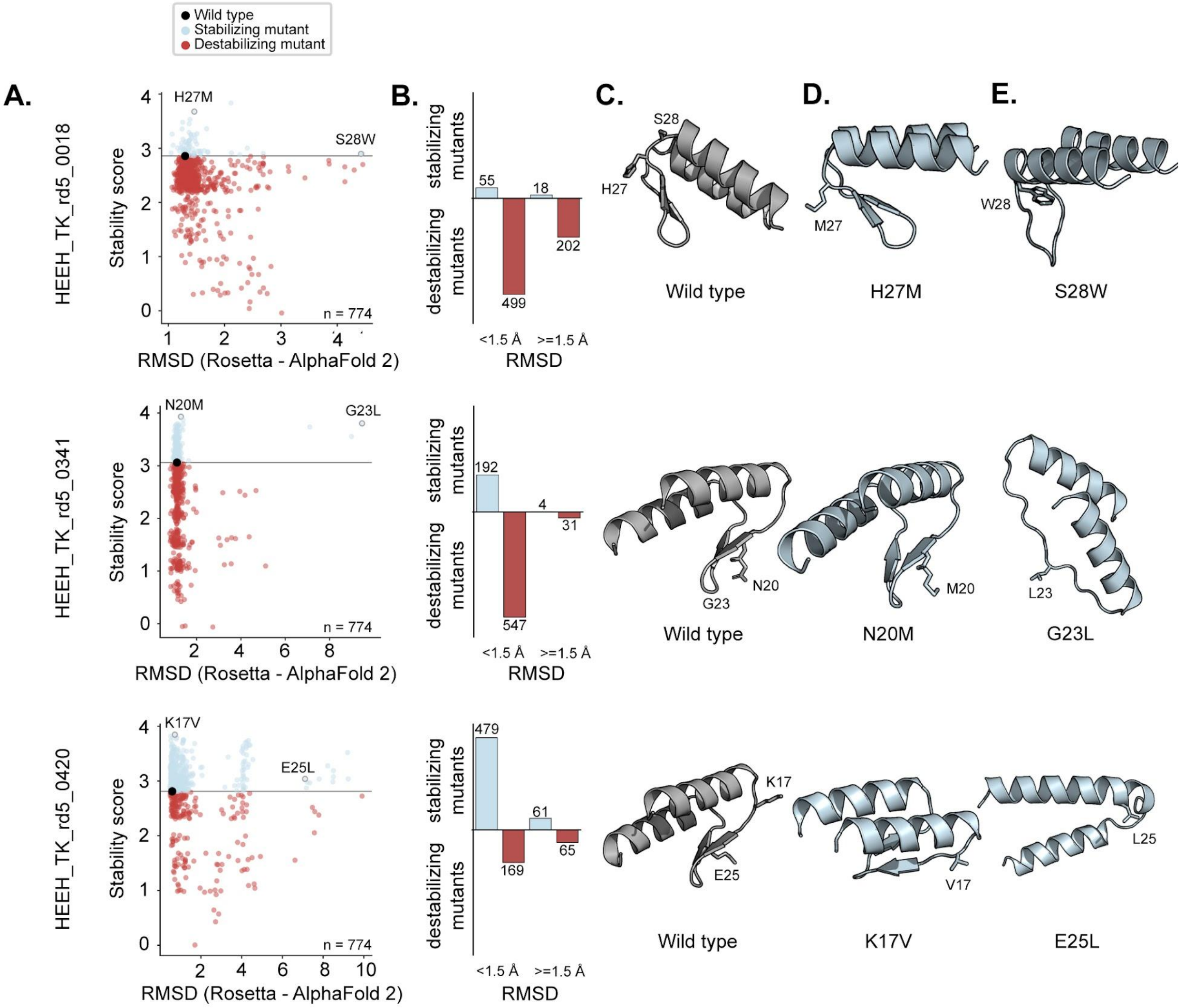


Fig. S12. Impact of mutant residue type and the position of the mutant on stability. To analyze whether certain mutations at specific positions on the backbone are more likely to destabilize the miniprotein, we (A) showed the distributions of the change in stability score (mutant stability - wild type stability) at each position of the six miniproteins of which we have deep mutational scanning data. The top five destabilizing positions (9, 21, 26, 35, and 36) had an average change in stability score < 1. (B) For positions 9, 21, 26, 35, and 39, the change in stability score are shown by the type of mutant residue.


● Wild type
● Stabilizing mutant
● Destabilizing mutant
A.
B.
C.
D.
E.
HEEH_TK_rd5_0018
Stability score
RMSD (Rosetta - AlphaFold 2)
H27M
S28W
n = 774
stabilizing mutants
destabilizing mutants
55
499
18
202
<1.5 Å
>=1.5 Å
RMSD
S28
H27
Wild type
M27
H27M
W28
S28W
HEEH_TK_rd5_0341
N20M
G23L
n = 774
192
547
4
31
G23
N20
Wild type
M20
N20M
L23
G23L
HEEH_TK_rd5_0420
K17V
E25L
n = 774
479
169
61
65
K17
E25
Wild type
V17
K17V
L25
E25L

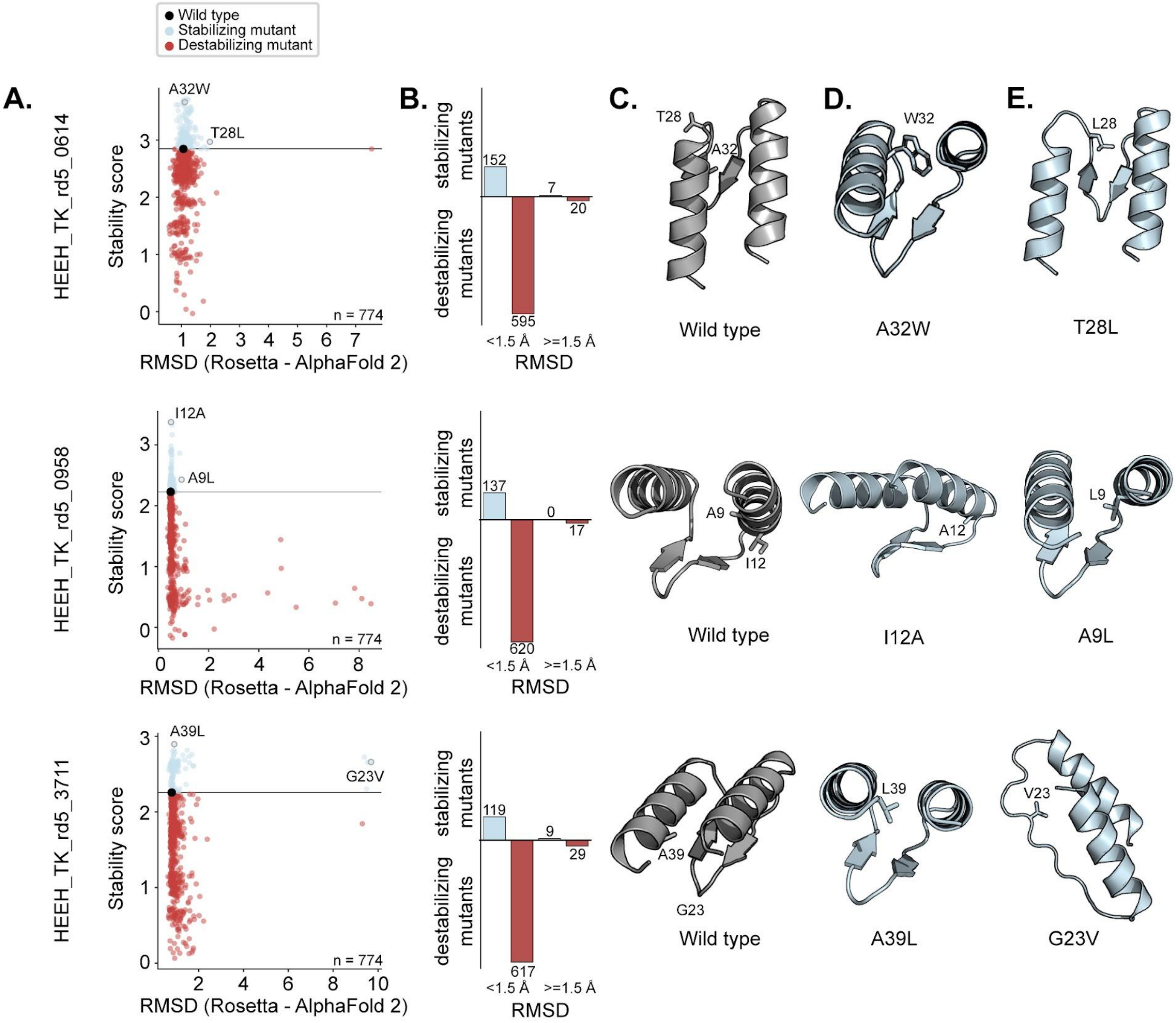


Fig. S13. Agreement between wildtype Rosetta design models and mutant AlphaFold 2-predicted structures. We compared how well the structure of a wildtype design (Rosetta design model) agrees with all possible mutants (whose structures are predicted by AlphaFold 2). (A) The stability score of all mutants and its corresponding wildtype are plotted by their structural agreement (RMSD); stabilizing mutants are colored in blue whereas destabilizing mutants are colored in red. (B) Stabilizing and destabilizing mutants separated by RMSD < 1.5 Å or > 1.5 Å. Cartoons of a (C) wildtype design (Rosetta design model), (D) a stabilizing mutant (AlphaFold 2-predicted structure) with close structural agreement to its corresponding wildtype, and (E) a

stabilizing mutant (AlphaFold 2-predicted structure) with structural agreement that is less than (C).

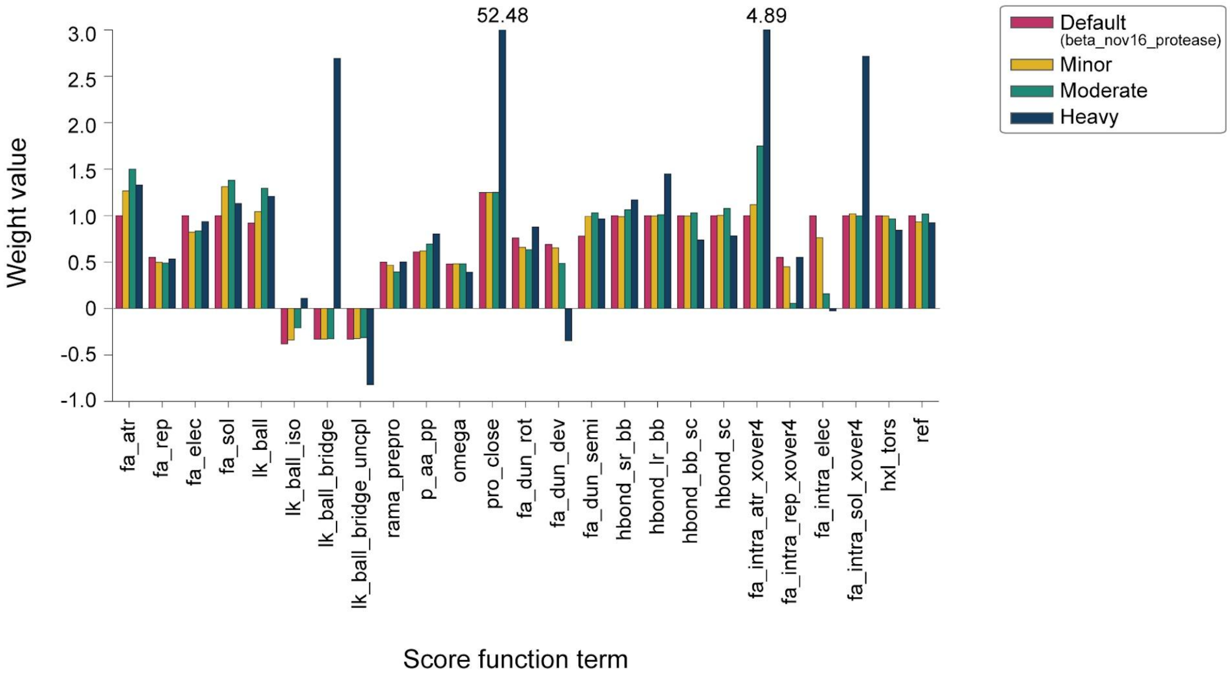


Fig. S14. Re-weighted values of Rosetta energy score terms. We used three differently weighted values (Minor, Moderate, and Heavy) in addition to the default values of the Rosetta score terms when designing diversity-oriented αββα miniproteins. The values of each score term are presented as a barplot. In the reweighting ridge regression, the Minor set used regularization strength alpha=200,000, Moderate used alpha=20,000, and “Heavy” used alpha=0.1.

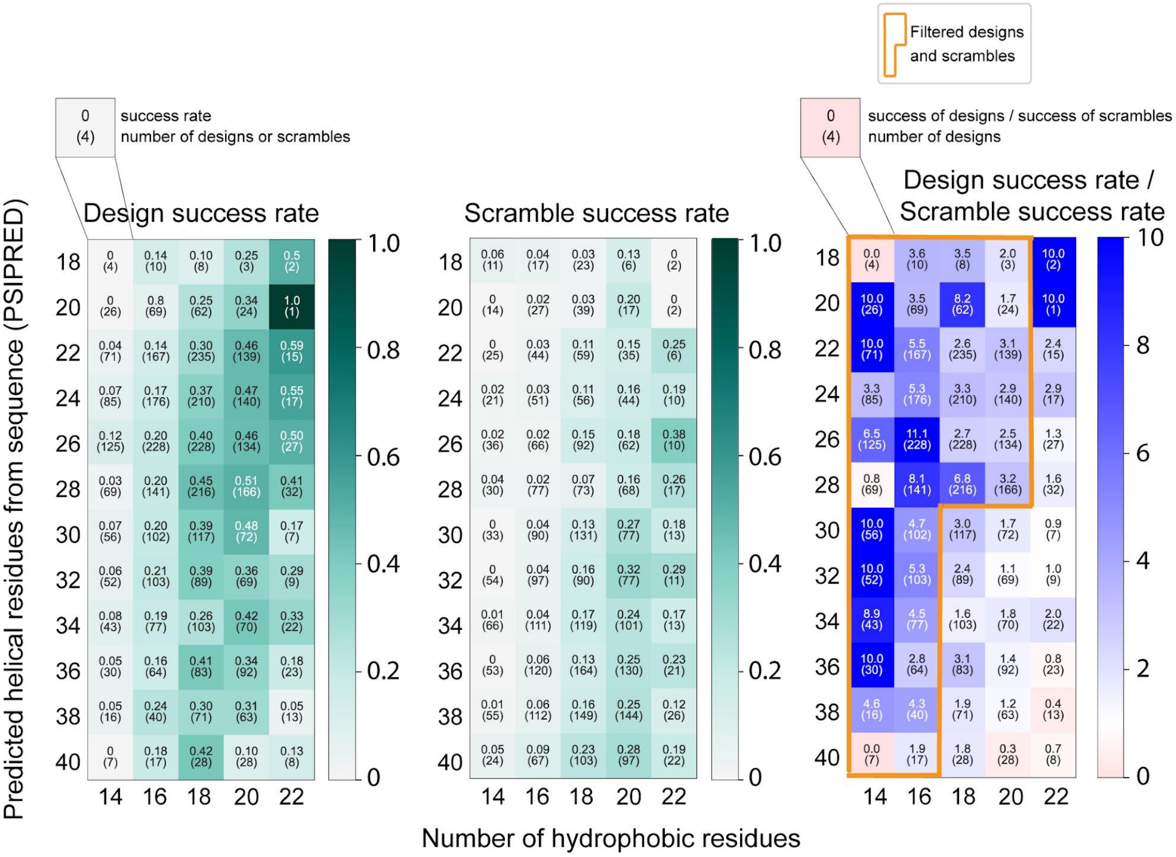
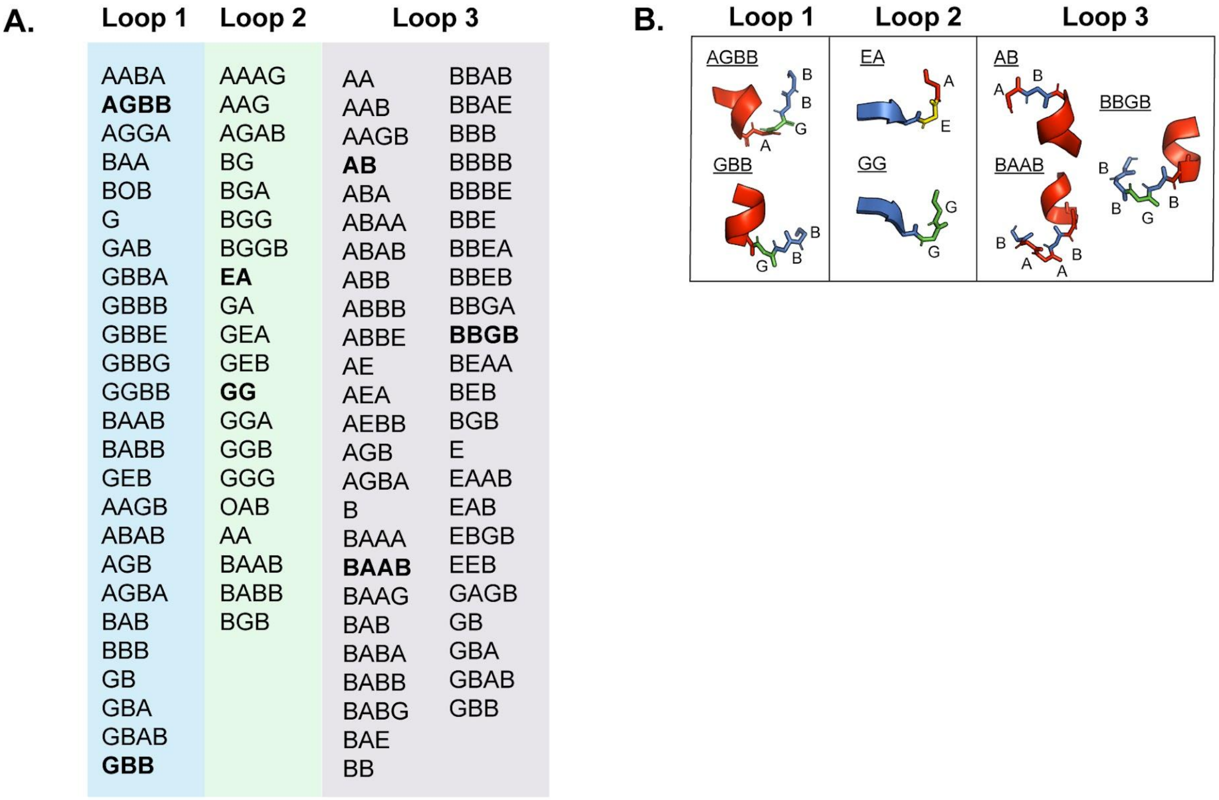


Fig. S15. Filtering strategy for diversity-oriented designs. In order to rule out sequences that might be "false positives" (i.e. sequences that are stable even without a specific designed structure), we examined the success rates of designs (left) and scrambled sequences (middle) after dividing them into bins according to their number of hydrophobic residues and number of residues predicted to be helical by PSIPRED.[209] We minimized false positives by filtering out regions where the scrambled success rate was high (outside the orange box).



A.

Loop 1	Loop 2	Loop 3	
AABA	AAAG	AA	BBAB
AGBB	AAG	AAB	BBAE
AGGA	AGAB	AAGB	BBB
BAA	BG	**AB**	BBBB
BOB	BGA	ABA	BBBE
G	BGG	ABAA	BBE
GAB	BGGB	ABAB	BBEA
GBBA	**EA**	ABB	BBEB
GBBB	GA	ABBB	BBGA
GBBE	GEA	ABBE	**BBGB**
GBBG	GEB	AE	BEAA
GGBB	**GG**	AEA	BEB
BAAB	GGA	AEBB	BGB
BABB	GGB	AGB	E
GEB	GGG	AGBA	EAAB
AAGB	OAB	B	EAB
ABAB	AA	BAAA	EBGB
AGB	BAAB	**BAAB**	EEB
AGBA	BABB	BAAG	GAGB
BAB	BGB	BAB	GB
BBB		BABA	GBA
GB		BABB	GBAB
GBA		BABG	GBB
GBAB		BAE	
GBB		BB	

Fig. S16. Loop patterning in diversity-oriented designs. (A) A list of all unique loop structures (written in ABEGO notation) observed in the diversity-oriented designs. Loop structures in bold indicate that > 300 designs with that particular structure were identified in our dataset. As a result, these loop structures were selected as inputs to a topology-focused regression model (see Fig. 4F-G). (B) Cartoons visualizing the seven most common loop patterns in our data.

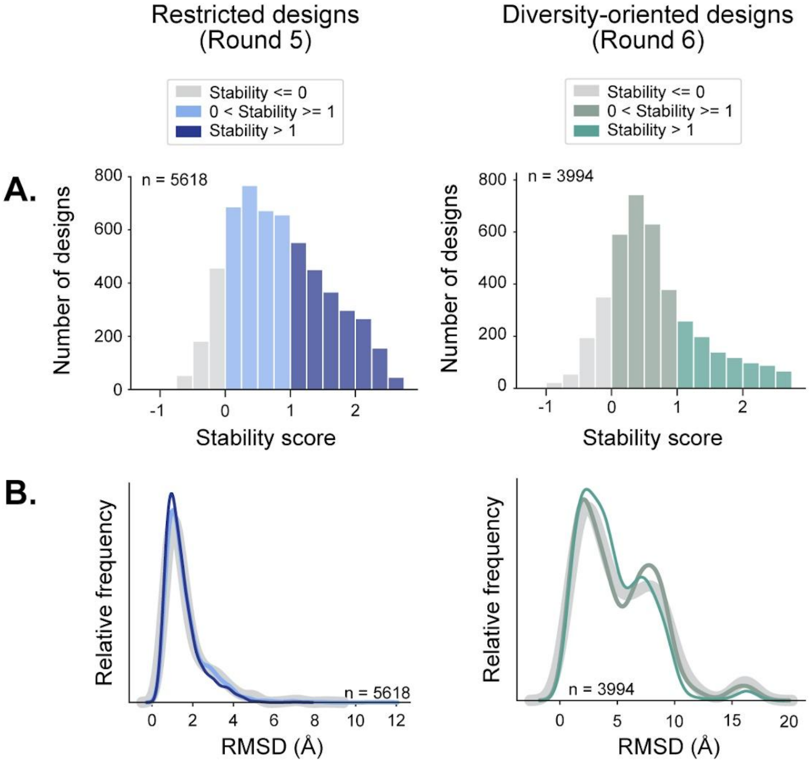


Fig. S17. Structure agreement between AlphaFold 2 and Rosetta design models of varying stability. (A) The distribution of αββα miniprotein stability scores, divided into three overall stability levels. (B) For each stability level, we show the distribution of structural agreement (RMSD) between Rosetta designs models and AlphaFold 2-predicted structures. Each level of stability has a similar distribution of RMSDs.

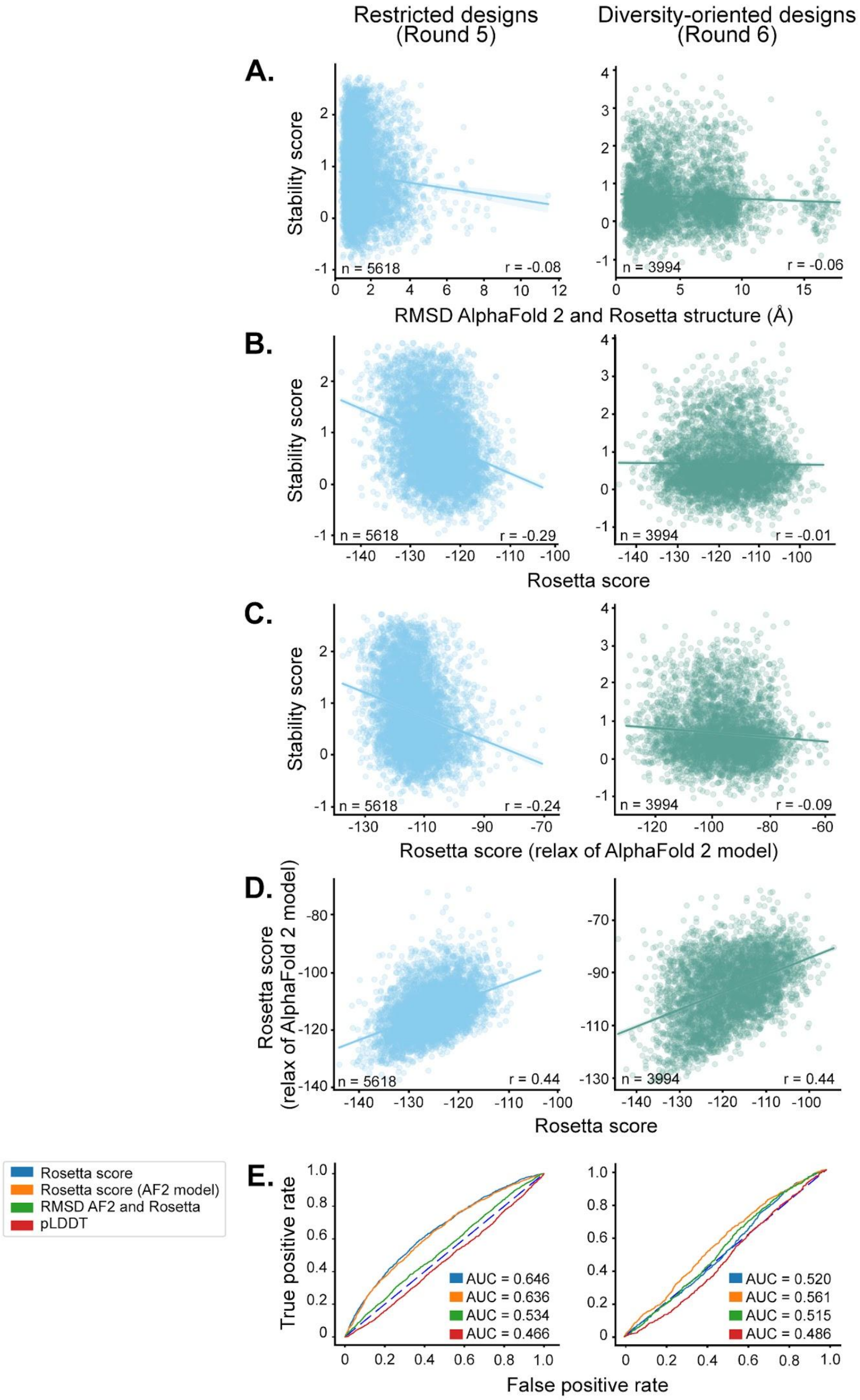

Restricted designs (Round 5)
Diversity-oriented designs (Round 6)
A.
Stability score
n = 5618
r = -0.08
n = 3994
r = -0.06
RMSD AlphaFold 2 and Rosetta structure (Å)
B.
Stability score
n = 5618
r = -0.29
n = 3994
r = -0.01
Rosetta score
C.
Stability score
n = 5618
r = -0.24
n = 3994
r = -0.09
Rosetta score (relax of AlphaFold 2 model)
D.
Rosetta score (relax of AlphaFold 2 model)
n = 5618
r = 0.44
n = 3994
r = 0.44
Rosetta score
E.
True positive rate
False positive rate
AUC = 0.646
AUC = 0.636
AUC = 0.534
AUC = 0.466
AUC = 0.520
AUC = 0.561
AUC = 0.515
AUC = 0.486
Rosetta score
Rosetta score (AF2 model)
RMSD AF2 and Rosetta
pLDDT

Fig. S18. Comparison of Rosetta design models and AlphaFold 2-predicted structures in predicting stability. Structures of all αββα miniproteins were predicted by AlphaFold 2, and the predicted model with the lowest RMSD to the designed structure was used for further analysis. The experimental stability scores are compared to the (A) structural agreement (RMSD) between AlphaFold 2-predicted models and Rosetta design models, (B) Rosetta energy determined from the Rosetta design models, (C) and Rosetta energy determined from the AlphaFold 2-predicted structures. (D) Rosetta energy scores determined from the Rosetta design models and AlphaFold 2-predicted models are compared to each other. (E) ROC curves to classify αββα miniproteins as stable (stability score ≥ 1) or unstable based on scores determined from the Rosetta model (blue), AlphaFold 2-predicted structure (orange), the agreement between the Rosetta model and AlphaFold 2 structure (green), and the AlphaFold 2 confidence score (red).

Table S1. Selected αββα designs for circular dichroism, thermal denaturation, and deep mutational scanning

Design	Sequence (Helix, Strand)	Hydrophobicity
HEEH_TK_rd5_0958	DIEEIEKKARKILEKGDSIEIAGFEVRDEEDLKKILEWLRRHG	347
HEEH_TK_rd5_3711	SWEDLERLAREALERGETIHILGFEIRSEEDAKKFAEWARRWE	842
HEEH_TK_rd5_0341	DLEELEEDLKQALREGRKVNILGIEVTTEEQARRLIEFLRRFI	968
HEEH_TK_rd5_0614	DLEKLRELLEDALRKGITIRFAGIEVKTEEEAERLLEWLKRKL	1018
HEEH_TK_rd5_0018	SFEELIKLIEDLLRKGDHINILGFEVHSEEEARRLIEWLRRAA	1174
HEEH_TK_rd5_0420	SLEELLKLAEEALKRGKTIRILGFEISSEEALRRFEEWLRRFI	1258

* gray and blue boxes indicate residues that are in the helices and β-strands, respectively; hydrophobic residues are colored in orange. Hydrophobicity values quantified based on.[97]

Table S2. NMR restraints, structural statistics, quality scores, rotational correlation times and hydrodynamic radii for HEEH_TK_rd5_0341 and HEEH_TK_rd5_0958

Design ID[a]	HEEH_TK_rd5_0341	HEEH_TK_rd5_0958
PDBID	7T2F	8DOA
BMRBID	30974	31033
Relaxation-derived Dynamics		
Rotational correlation time, τc (ns)	11.9	6.1
Hydrodynamic radius, rH (nm)	2.15	1.72
NMR restraints:		
Total NOEs	1564	546
Intra-residual	437	198
Sequential (i – j = 1)	465	151
Medium-range (1 < i – j < 5)	333	98
Long-range (i – j ≥ 5)	275	99
Inter-molecular	54	N/A
Hydrogen Bonds	44	N/A
Dihedral Angles:		
φ	78	42
ψ	78	42
Structural Statistics:		
r.m.s.d. from experimental restraints:		
Distance restraints (Å)	0.0197 ± 0.002	0.0072 ± 0.0016
Dihedral angle restraints (°)	0.280 ± 0.14	0.463 ± 0.19
Violations in the NMR ensemblea:		
Max. distance restraint violation (Å)	0.503	< 0.3
Max. dihedral restraint violation (°)	6.2	5.2
r.m.s.d. from idealized geometry:		
Bond lengths (Å)	0.0147 ± 0.0002	0.0140 ± 0.0004
Bond angles (°)	0.99 ± 0.017	0.90 ± 0.027
Impropers (°)	1.65 ± 0.09	1.59 ± 0.15
Average pair-wise r.m.s.d. (Å)[b]:		
Heavy	1.2	1.5
Backbone	0.6	0.7
Structure quality scores:		
Ramachandran plot (%)[b,c]		
Most favored	92.7	92.7
Additionally allowed	7.3	7.3
Generously allowed	0.0	0.0
Disallowed	0.0	0.0
Structural Quality Factors (raw/Z-scores)[d]		
Procheck (phi/psi)	0.00/0.31	0.13/0.83
Procheck (all)	-0.18/-1.06	-0.04/-0.24
Molprobity clash	15.72/-1.17	8.71/0.03

[a]The NMR ensemble consists of the 20 lowest energy structures out of 100 calculated; [b]Calculated for residues 22 to 62 inclusive for HEEH_TK_rd5_0958, and residues 23-36, 40-42, 45-47 and 50-62 inclusive, for HEEH_TK_rd5_0341; [c]Based on Procheck analysis;[128] [d]Calculated using the PSVS server.[129] (https://montelionelab.chem.rpi.edu/PSVS/).

Table S3. Coefficients of a ten-feature linear regression model

Feature	**Coefficient Δ Stability Score (per 1 σ)**	**Coefficient Δ Stability Score (per contact)**
Large nonpolar count (FILMWY)	0.256	0.161
Nonpolar residue-residue contacts	0.189	0.048
Local sequence-structure propensity	0.129	0.129
Ser or Thr in helix caps	0.086	0.125
Glu-Arg residue-residue contacts	0.082	0.038
Nonpolar residue at design ends	0.070	0.078
Favorable net charge at helix ends	0.042	0.020
Glu-Glu residue-residue contacts	-0.023	0.086
Buried unsatisfied polar atoms	-0.058	-0.134
Increased net charge^2	-0.116	-0.085

Table S4. Mean correlation coefficients and 95% confidence intervals of linear regression models

Model	Mean r	95% CI
10 features	0.642	(0.628, 0.657)
10 features + 25 score terms	0.673	(0.659, 0.687)
10 features + 25 shuffled score terms	0.646	(0.630, 0.662)
10 features + shuffled stability scores	0.059	(0.038, 0.081
10 features + 25 shuffled score terms with shuffled stability scores	0.107	(0.083, 0.130)

Table S5. Reweighted Rosetta score functions applied to miniprotein designs of varying topologies

Spearman's Correlation					
Rounds 1-4	**ααα**	**αββα**	**βαββ**	**ββαββ**	**Score function**
	-0.505	-0.089	-0.390	-0.628	beta_nov16_protease
	-0.610	-0.172	-0.479	-0.631	minor
	-0.605	-0.165	-0.482	-0.624	moderate
ROC					
Rounds 1-4	0.767	0.636	0.710	0.901	beta_nov16_protease
	0.824	0.672	0.767	0.911	minor
	0.821	0.681	0.769	0.908	moderate
Spearman's Correlation					
Round 4 only	-0.058	-0.070	-0.189	-0.375	beta_nov16_protease
	-0.185	-0.153	-0.262	-0.473	minor
	-0.213	-0.147	-0.268	-0.431	moderate
ROC					
	0.515	0.602	0.596	0.714	beta_nov16_protease
	0.592	0.654	0.644	0.758	minor
	0.609	0.647	0.656	0.734	moderate

Re-weighted energy functions (minor and moderate) were used to calculate the predicted scores of previously published protein designs (four different topologies). The Spearman's correlation coefficient was determined based on comparing the experimental stability score vs. the predicted stability score. beta_nov16_protease is the score function used previously.[37] Negative Spearman's correlations are expected because more negative energy scores and more positive stability scores both imply greater stability.

APPENDIX 2

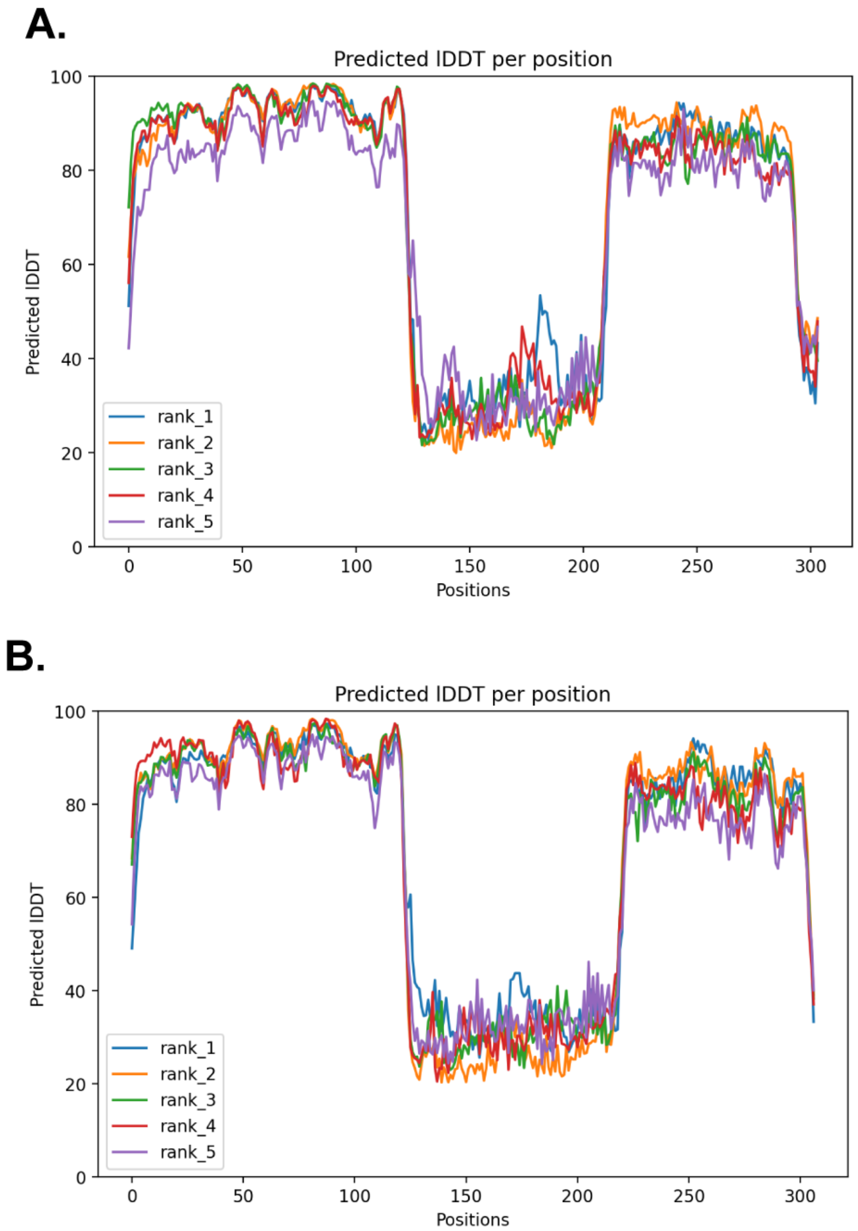


Fig S19. pLDDT plots for TMD-VidC-TF. pLDDT plots for TMD-VidC-TF with the NLS at the (A) C-terminus or (B) between VidC and VP64. AlphaFold 2 outputs five predicted structures (rank_1 - rank_5). Rank_1 is the best predicted structure and chosen for analysis in Chapter 3. A pLDDT score > 80 indicates high confidence in the accuracy of the predicted structure at that residue.

Table S6. Design sequences of intein constructs

Construct	Sequence
EHEE_rd2_0005-FLAG-VidN	GSSTTRYRFTDEEEARRAAKEWARRGYQVHVTQNGTYWEVEVRDYKDDDDKESGCLPKEAVVQIRLTKKG
CD4-FLAG-FRB-TMD-VidC-VP64-ZF1-NLS	MCRAISLRRLLLLLLQLSQLLAVTQGDYKDHDGDYKDHDIDYKDDDDKMLEMWHEGLEEASRLYFGERNVKGMFEVLEPLHAMMERGPQTLKETSFNQAYGRDLMEAQEWCRKYMKSGNVKDLLQAWDLYYHVFRRISKTRGGSGGSGGSGGSGGSGGSTGIILYASGSLALAVLLLLAGLAGSGMIEEKKVTVQELRELYLSGEYTIEIDTPDGYQTIGKWFDKGVLSMVRVATATYETVCAFNHMIQLADNTWVQACELDVGVDIQTAAGIQPVMLVEDTSDAECYDFEVMHPNHRYYGDGIVSHNSGKASGGSGGTGDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSRSSGSSGSGSGSGSGTGGARPGERPFQCRICMRNFSRQDRLDRHTRTHTGEKPFQCRICMRNFSQKEHLAGHLRTHTGEKPFQCRICMRNFSRRDNLNRHLKTHLRGSGPKKKRKV
CD4-FLAG-FRB-TMD-VidC-NLS-VP64-ZF1	MCRAISLRRLLLLLLQLSQLLAVTQGDYKDHDGDYKDHDIDYKDDDDKMLEMWHEGLEEASRLYFGERNVKGMFEVLEPLHAMMERGPQTLKETSFNQAYGRDLMEAQEWCRKYMKSGNVKDLLQAWDLYYHVFRRISKTRGGSGGSGGSGGSGGSGGSTGIILYASGSLALAVLLLLAGLAGSGMIEEKKVTVQELRELYLSGEYTIEIDTPDGYQTIGKWFDKGVLSMVRVATATYETVCAFNHMIQLADNTWVQACELDVGVDIQTAAGIQPVMLVEDTSDAECYDFEVMHPNHRYYGDGIVSHNSGKASGGSGGTGPKKKRKVGSGDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLGSDALDDFDLDMLINSRSSGSSGSGSGSGSGTGGARPGERPFQCRICMRNFSRQDRLDRHTRTHTGEKPFQCRICMRNFSQKEHLAGHLRTHTGEKPFQCRICMRNFSRRDNLNRHLKTHLRGSG

Table S7. Primers used for cloning TMD-VidC-TF into pRL-plasmid

Primer	**Sequence**
FW primer to linearize pRL-SV40p	TTCTAGAGCGGCCGCT
REV primer to linearize pRL-SV40p	AACTTCTGCAGCTTAAGTTCGAGACT
FW primer to linearize TMD-VidC-TF	TAAGCTGCAGAAGTTATGTGTAGAGCAATTTCCCTGC
REV primer to linearize TMD-VidC-TF	GCGGCCGCTCTAGAATCACACCTTTCTCTTCTTCTTTGGT