UNIVERSITÄT BAYREUTH

Functionalization of *de novo* TIM barrels

Dissertation

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
an der Fakultät für Biologie, Chemie und Geowissenschaften
der Universität Bayreuth

Vorgelegt von

Julian Beck

aus Darmstadt

Bayreuth, 2025

Die vorliegende Arbeit wurde in der Zeit von Januar 2021 bis Juli 2025 in Bayreuth am Lehrstuhl Biochemie III unter Betreuung von Frau Prof. Dr. Birte Höcker angefertigt.

Vollständiger Abdruck der von Fakultät für Biologie, Chemie und Geowissenschaften der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat).

Art der Dissertation: Kumulative Dissertation

Dissertation eingereicht am: 08.07.2025

Zulassung durch die Promotionskommission: 16.07.2025

Wissenschaftliches Kolloquium: 26.02.2026

Amtierender Dekan: Prof. Dr. Janosch Hennig

Prüfungsausschuss:

Prof. Dr. Birte Höcker (Gutachterin)

Prof. Dr. Janosch Hennig (Gutachter)

Prof. Dr. Johannes Margraf (Vorsitz)

PD Dr. Alfons Weig

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
ATP	Adenosine Triphosphate
CANVAS	Customizing Amino-acid Networks for Virtual Active-site Scaffolding
CASP	Critical Assessment of protein Structure Prediction
CD	Circular-Dichroism
DNA	Deoxyribonucleic Acid
DHAP	Dihydroxyacetone Phosphate
EC	Enzyme Commission
EMBL-EBI	European Molecular Biology Laboratory's European Bioinformatics Institute
GAP	Glyceraldehyde 3-Phosphate
HIV	Human Immunodeficiency Virus
KI	Künstliche Intelligenz
MALS	Multi Angle Light Scattering
MD	Molecular Dynamics
MPPN	Message-Passing Neural Network
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PET	Polyethylene Terephthalate
pLDDT	predicted Local Distance Difference Test
RNA	Ribonucleic Acid
SAXS	Small Angle X-ray Scattering
SEC	Size Exclusion Chromatography
TIM	Triosephosphate Isomerase
XAI	explainable Artificial Intelligence

Summary

Nature generated a vast array of proteins with diverse functions over the course of evolution, yet it explored only a small fraction of the theoretically possible sequence space. *De novo* protein design seeks to explore the entire space, expanding the repertoire of natural proteins with entirely new proteins from scratch – potentially, with tailor-made properties. Despite being inherently challenging, the field achieved several milestones, with progress accelerating due to the increasing integration of machine learning. A notable milestone was the design of the first *de novo* TIM barrel, sTIM11, which resembled nature's most versatile protein fold.

The TIM- or $(\beta\alpha)_8$-barrel fold consists of eight parallel β-strands forming a central barrel, surrounded by eight α-helices on the outside. This architecture is present in six of the seven enzyme commission classes, making it a highly interesting design target. Given the enzymatic versatility of natural TIM barrels, the design of sTIM11 raised hope that tailor-made enzymes were soon within reach. However, progress toward functional *de novo* TIM barrels remained limited. A key reason for this is their highly idealized topology with minimal loops. In contrast, natural TIM barrels have additional structural features – such as elongated βα-loops including further secondary structural elements – that form binding pockets and position catalytic residues thereby enabling the versatility of this fold. To achieve functionalization of *de novo* TIM barrels, deviations from their idealized architecture are essential. A possible strategy involves two steps: first, the introduction of structural extensions to form suitable pockets, and second, incorporation of functional residues to enable a specific activity.

To advance the functionalization, we designed and characterized two sets of *de novo* TIM barrels with structural extensions generated to form pockets for downstream functionalization. For the first set, we utilized a physics-based modular design approach with the program Rosetta and generated so-called αTIMs extended with one or two helix-loop-helix motifs. Circular dichroism (CD) spectroscopy revealed a stepwise increase in α-helical content in comparison to a non-extended *de novo* TIM barrel. AlphaFold2 predictions of the αTIMs indicated the formation of the designed motifs with high confidence. Additionally, due to the inserted motifs possible binding sites were predicted enabling downstream functionalization. Building on the obtained insights, we designed the second set of *de novo* TIM barrels – so-called HalluTIMs. This set was diversified with two or three extensions generated by constrained hallucination. The hallucinated extensions were mainly α-helical and exhibited features similar to those of the αTIMs. The resulting proteins revealed an increase in α-helical content in CD spectroscopy compared to a non-extended *de novo* TIM barrel. Two solved crystal structures confirmed the formation of the extensions. Within these crystal structures potential binding sites were predicted once again, showcasing the potential for downstream functionalization. However, small angle X-ray scattering (SAXS) measurements revealed high flexibility of the extensions, complicating downstream functionalization and underscoring the challenges of stepwise functionalization.

To overcome these challenges, we introduced both structural extensions and enzymatic activity simultaneously in a third set of *de novo* TIM barrels. As proof of principle, we selected the Kemp elimination, a common benchmark reaction in enzyme design, and named this set of *de novo* TIM barrels KempTIMs. The design workflow combined AI-based methods like RFdiffusion and ProteinMPNN, for generating structural extensions, and the physics-based software Triad for enzyme design. A key aspect of the extension design was the positioning of one catalytic residue above the barrel. This placement enabled the extensions to anchor the catalytic residue and form active sites simultaneously. Experimental characterization revealed KempTIM4, the first *de novo* TIM barrel with an enzymatic function based on a tailor-made extension. Sequence optimization of KempTIM4 resulted in an improved variant facilitating crystallization. The crystal structure revealed that the entire lid, including multiple elongated loops, was resolved and in close agreement to its prediction. To understand why some designs exhibited activity while others did not, we performed molecular dynamics (MD) simulations and identified a narrower entrance to the active site for inactive designs as a key difference. The developed workflows and obtained insights in this thesis provide the framework for a future family of enzymatically active *de novo* TIM barrels and the exploration of the full potential of this fold.

Zusammenfassung

Die Natur hat im Verlauf der Evolution eine enorme Vielfalt an Proteinen mit unterschiedlichsten Funktionen hervorgebracht, dabei aber nur einen geringen Teil des theoretisch möglichen Sequenzbereiches benutzt. Das *de novo* Proteindesign versucht den ganzen Bereich zu verwenden, um das Repertoire natürlicher Proteine durch völlig neue Proteine mit maßgeschneiderten Eigenschaften zu erweitern. Trotz der immensen Komplexität dieses Vorhabens wurden bereits zahlreiche Meilensteine erreicht und die Häufigkeit neuer Errungenschaften erhöht sich durch den intensivierten Einsatz von künstlicher Intelligenz (KI). Ein bedeutender Durchbruch war das Design des ersten *de novo* TIM barrels, sTIM11, welches der vielseitigsten Proteinfaltung der Natur nachempfunden wurde.

Das TIM oder $(\beta\alpha)_8$-barrel besteht aus acht parallelen β-Strängen, welche ein zentrales Fass (barrel) bilden, und von acht α-Helices umgeben sind. Diese Topologie findet sich in sechs der sieben Enzymklassen, was das TIM barrel zu einem sehr attraktiven Ziel im Proteindesign macht. Aufgrund dieser enormen enzymatischen Vielseitigkeit bei natürlichen TIM barrels weckte das Design von sTIM11 die Hoffnung, dass maßgeschneiderte Enzyme bald realisierbar sein könnten. Jedoch blieb der Fortschritt in Richtung funktionaler *de novo* TIM barrels bislang begrenzt. Ein wesentlicher Grund dafür liegt in der idealisierten Faltung von *de novo* TIM barrels, die nur über minimale βα-Schleifen verfügt. Im Vergleich zeigen natürliche TIM barrels zusätzliche strukturelle Erweiterungen auf – etwa längere βα-Schleifen oder sogar Sekundärstrukturelemente – die Bindetaschen formen und katalytisch relevante Reste platzieren und somit die funktionelle Vielfalt ermöglichen. Um funktionelle *de novo* TIM barrels zu ermöglichen, sind daher Abweichungen von ihrer idealisierten Topologie notwendig. Eine mögliche Strategie besteht aus zwei Schritten: zunächst werden Strukturelemente eingefügt, die Bindetaschen formen, und anschließend werden in diesen Taschen funktionelle Reste eingebracht, um die gewünschte Aktivität auszuführen.

Um eine Funktionalisierung zu ermöglichen, generierten und charakterisierten wir zwei Arten von *de novo* TIM barrels mit verschiedenen strukturellen Erweiterungen, die zur Ausbildung von Bindetaschen für eine nachfolgende Funktionalisierung beitragen. Für die erste Art kam eine physikbasierte, modulare Designstrategie mit dem Programm Rosetta zum Einsatz, und es wurden sogenannte αTIMs hergestellt, die um ein oder zwei Helix-Schleife-Helix-Motive erweitert wurden. Circular-Dichroismus (CD) Spektroskopie zeigte eine stufenweise Zunahme des α-helikalen Anteils im Vergleich zu einem idealisierten *de novo* TIM barrel. Strukturvorhersagen der αTIMs mit AlphaFold2 zeigten mit hoher Konfidenz die korrekte Ausbildung der eingebrachten Motive. Zusätzlich wurden durch die eingebrachten Motive potenzielle Bindetaschen vorhergesagt, die eine nachfolgende Funktionalisierung ermöglichen. Aufbauend auf den gewonnenen Erkenntnissen, generierten wir die zweite Art von *de novo* TIM barrels – so genannte HalluTIMs. Diese Art wurde mit zwei oder drei Erweiterungen mittels *constrained hallucination* diversifiziert. Die halluzinierten Erweiterungen waren vorwiegend α-helikal und wiesen große Ähnlichkeit zu den αTIMs auf. In der experimentellen Charakterisierung zeigten die

Proteine erneut eine Zunahme des α-helikalen Anteils in einem CD-Spektrum auf. Zwei aufgeklärte Kristallstrukturen bestätigten die korrekte Ausbildung der halluzinierten Erweiterungen. Erneut wurden innerhalb dieser Kristallstrukturen mögliche Bindetaschen vorhergesagt, welche das Potenzial für eine nachfolgende Funktionalisierung aufzeigen. Jedoch deuteten *small angle X-ray scattering* (SAXS) Messungen darauf hin, dass die Erweiterungen eine hohe inhärente Flexibilität aufweisen, was eine nachfolgende Funktionalisierung erschwert und die Herausforderungen einer schrittweisen Funktionalisierung verdeutlicht.

Um diese Herausforderungen zu meistern, generierten wir eine dritte Art von *de novo* TIM barrels mit strukturellen Erweiterungen, die direkt maßgeschneidert für eine spezifische enzymatische Reaktion waren. Als Modellreaktion wurde die Kemp Eliminierung ausgewählt, eine etablierte Enzymreaktion im Proteindesign, weshalb diese Art von *de novo* TIM barrels KempTIMs genannt wurde. Der Designprozess involvierte KI-basierte Methoden wie RFdiffusion und ProteinMPNN, um die strukturellen Erweiterungen zu entwerfen, und die physik-basierte Software Triad für Enzymdesign. Ein zentraler Aspekt der Designstrategie für die Erweiterungen war die Positionierung eines katalytisch relevanten Restes oberhalb des Barrels, wodurch Erweiterungen resultierten, die diesen Rest präzise positionierten und gleichzeitig ein aktives Zentrum ausbildeten. Die experimentelle Charakterisierung identifizierte KempTIM4 als das erste *de novo* TIM barrel mit einer Aktivität, die auf einer maßgeschneiderten Erweiterung basiert. Eine anschließende Sequenzoptimierung resultierte in einer verbesserten Varianten, welche die Kristallisation ermöglichte. Die aufgeklärte Kristallstruktur zeigte, dass die gesamten Erweiterungen, inklusive mehrerer verlängerter Schleifen, aufgelöst und in Übereinstimmung mit den Vorhersagen sind. Um zu verstehen, warum manche Designs Aktivität aufwiesen und andere nicht, führten wir Molekulardynamik Simulationen durch und identifizierten als entscheidenden Unterschied einen wesentlich engeren Zugang zum aktiven Zentrum im Falle des inaktiven Designs. Unsere entwickelten Designstrategien und die gewonnenen Einsichten in dieser Arbeit bilden zukünftig die Grundlage für weitere enzymatisch aktive *de novo* TIM barrels und die Erschließung des vollen Potenzials dieser Faltung.

Introduction

1 From sequence to structure to function

Proteins control and enable a myriad of diverse biological functions, including cell division, signaling, immunity, and metabolism. To perform these tasks, proteins adopt specific three-dimensional structures tailored to their functions. Protein structures range from small monomers, such as myoglobin involved in oxygen transport, to large complexes, such as chaperones assisting other proteins during their folding (*1–3*). To date, over 100,000 unique protein structures have been identified, each encoded by a distinct amino acid sequence (*4*). This sequence determines the three-dimensional structure of a protein and, consequently, its function – a principle known as the sequence-structure relationship (*5*).

To understand how an amino acid sequence gives rise to structure, it is essential to consider the chemical nature of amino acids and the way they assemble into protein chains. Each amino acid consists of an amine group, a carboxyl group, and a chemically diverse side chain. They are linked by their amine and carboxyl groups forming peptide bonds and resulting in a long polypeptide chain. Each residue adopts specific angles between the N-Cα (phi angle) and Cα-CO (psi angle), with certain combinations highly preferred depending on the amino acid identity and the eventually formed secondary structure element (*6*). Furthermore, each rotatable side chain increases the conformational space exponentially with each additional residue in the protein. Assuming each residue has only three degrees of freedom, even a relatively short protein of 101 amino acids has an astronomical number of 3^{100} ($\approx 5x10^{47}$) possible conformations highlighting the vast structural complexity encoded within amino acid sequences (*7*). Yet in reality, proteins fold reliably and rapidly into their well-defined three-dimensional structures. If we further assume that a protein samples new conformations at a rate of 10^{13} per second, exhaustively exploring all possible conformations would take approximately 10^{27} years. This thought experiment, known as Levinthal's paradox, underscores the impossibility of a purely random search through all conformations, emphasizing the central question of how an amino acid sequence folds into its final structure – commonly referred to as the protein folding problem.

Since proteins are known to fold within seconds or less, an alternative mechanism must guide the protein folding process. A shift from a purely kinetic perspective to a thermodynamic one introduces the concept of a guided folding pathway or energy landscape, where proteins adopt the structure with the lowest free energy – known as the native state (*8*). This hypothesis – referred to as Anfinsen's hypothesis – states that the amino acid sequence directly encodes its native state, enabling spontaneously folding without external guidance along the energy landscape. This energy landscape resembles a funnel that guides the folding process toward the global minimum. Hereby, protein folding is driven by the formation of short- or long-range interactions, hydrophobic packing, and the stepwise assembly of structural elements, progressing from secondary structures to the final native state fulfilling its biological function (Figure 1) (*9*).

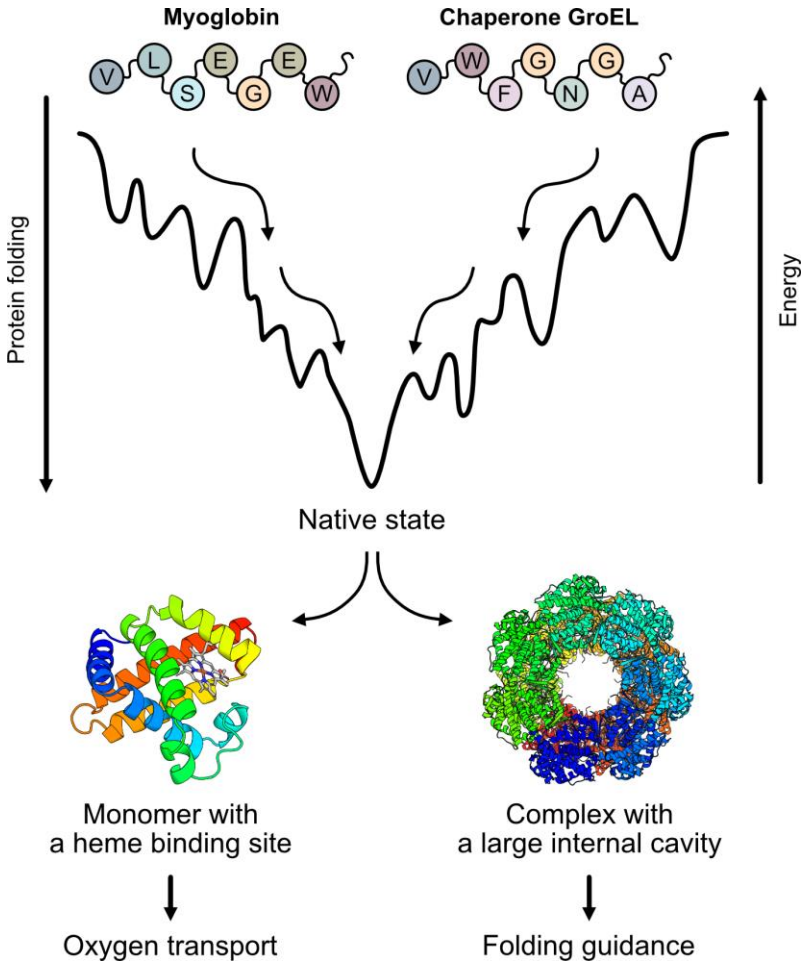


Figure 1: Schematic overview of the sequence-structure relationship and protein folding process. Amino acid sequences, for example myoglobin or the chaperone GroEL, fold along their folding pathway into their native states at the global minimum. This structure has features to fulfill the relevant biological function. In the case of myoglobin (PDB: 1MBN, cartoon representation, rainbow coloring from N- to C-terminus) the monomeric structure can bind heme to facilitate oxygen transport. The complex of the chaperone GroEL (PDB: 2EU1, cartoon representation, the biological assembly is displayed and colored over all chains in rainbow coloring) has a large internal cavity to assist other proteins during folding.

Anfinsen's hypothesis provides valuable insights into the sequence-structure relationship and the protein folding problem. However, with today's knowledge minor refinements are necessary to enhance its generalizability and account for additional observations in nature. The energy landscape is more complex and rougher than a simple funnel leading to the global minimum; it contains multiple local minima where proteins may become kinetically trapped, explaining the observation of misfolding and aggregation (*10*). Furthermore, the assumption of a single native state is an oversimplification as some proteins adopt multiple stable conformations and transition between states in response to external factors such as small molecules or temperature (*11*, *12*), suggesting the existence of multiple native states. Additionally, while proteins can fold spontaneously *in vitro*, folding *in vivo* is often assisted by chaperones, post-translational modifications, and other cellular factors that significantly influence the process (*13*). For some proteins, correct folding may not be possible without this external guidance. Although all these refinements are subtle, they once again underscore the complexity of the protein folding problem.

Understanding the complex relationship between sequence, structure, and function remains essential. If this relationship can be fully deciphered, it could be applied in a broad range of fields (*5*). The three-dimensional structure of proteins could be predicted directly from the sequence without experimental effort, saving tremendous time and resources (*4*). Furthermore, the entire protein folding problem could be inverted: given a desired function, the corresponding structure and encoding sequence could be designed. Such capabilities would pave the way for addressing some of the major challenges of our lifetime (*14*, *15*).

2 Protein design - The inverse folding problem

Today's diversity in protein structures and functions is the result of billions of years of evolution (*16*). Throughout this time, proteins have continuously adapted to environmental changes, either by altering their existing function or acquiring entirely new ones. A striking example of recent evolutionary adaptation is the emergence of PETases – enzymes capable of degrading polyethylene terephthalate (PET) – as a reaction to the abundance of microplastic in the environment (*17*, *18*). The first efficient PETase was discovered in the marine organism *Ideonella sakaiensis*, which utilizes products of PET degradation as a carbon source. Since its discovery, numerous studies have aimed to improve catalytic efficiency or find more potent PETases directly in nature to address environmental pollution caused by microplastics (*19–23*). However, in today's world we are confronted with many more challenges like climate change, global pandemics or the rising prevalence of neurodegenerative diseases linked to longer lifespans, requiring immediate solutions (*24–26*). While evolution could eventually sample new proteins to address these issues, the process is not directed and is inherently slow, unfolding over an uncertain amount of time. Based on urgency, we cannot afford to rely purely on natural selection and its slow pace. Instead, protein design offers a powerful alternative: the generation of novel proteins with tailored properties and functions, without relying on the slow process of natural evolution (*27*, *28*).

To fully grasp the potential and difficulties of protein design, it is essential to examine the sequence-structure relationship. As discussed previously, the sequence dictates the structure, which in turn determines function. To design proteins with specific functions inverting the protein folding problem is required, hence protein design is often called the inverse folding problem (Figure 2). Depending on the design objective, deep insight into different aspects can be required. Enzyme design, for instance, requires an understanding of the chemical mechanism underlying the desired reaction and the necessary active site in the protein (*29*). Similarly, designing an inhibitory binding protein against the SARS-Cov-2 receptor requires exquisite knowledge about possible binding sites to ensure specificity (*30*). Regardless of the individual design objective, the fundamental challenge remains the same: identifying a three-dimensional structure capable of the desired function and determining the encoding sequence. Given the vast combinatorial sequence space; already for relatively small proteins of 100 amino acids the possibilities are astronomically high ($20^{100} \approx 10^{130}$); this task is far from trivial (*15*). Nonetheless, the

field has achieved numerous milestones over the past decades, exemplified by the Nobel Prize in Chemistry awarded to David Baker in recognition of his pioneering contributions in computational protein design (*31*).

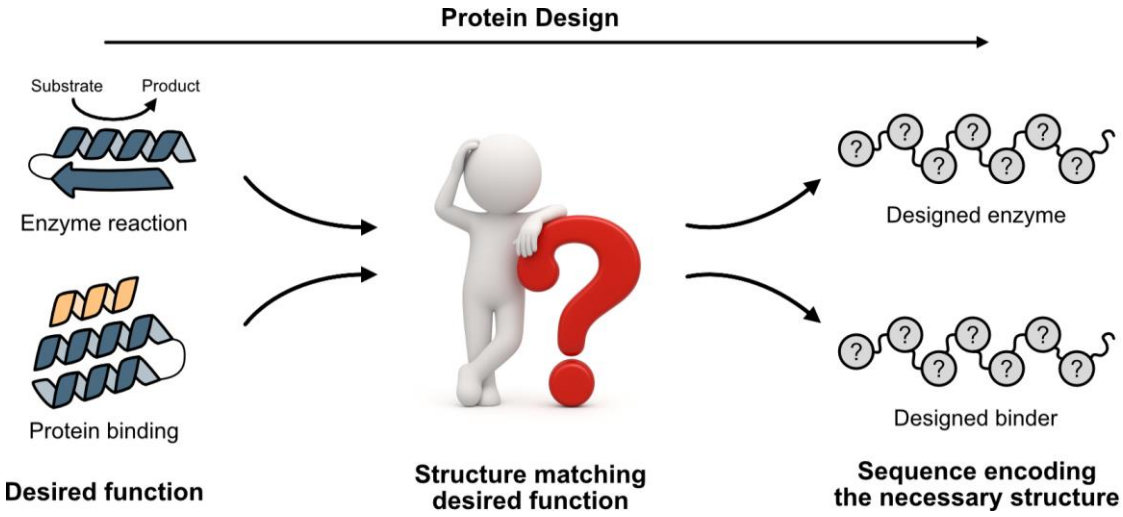


Figure 2: Schematic overview for many protein design tasks. For the design of a protein with a certain function, such as catalyzing an enzymatic reaction or binding a specific target, a protein structure needs to be found that is able to fulfill the specific function. Additionally, an amino acid sequence encoding the necessary protein structure needs to be found. Comic figure in the middle panel was generated with ChatGPT.

2.1 Computational and *de novo* protein design

Designing proteins with specific functions and properties is a significant challenge due to the vast combinatorial sequence space and required structural precision. To overcome this, modern protein design commonly employs computational modeling, including energy calculations, sequence optimization, and structure prediction algorithms (*32*). Although the term “computational protein design” is classically used to highlight these approaches, the use of such tools is now so widespread that is it frequently implied by “protein design” alone. Accordingly, this thesis uses “protein design” as a shorthand for computational protein design throughout.

To further manage the complexity of protein design, many strategies build on insights from natural proteins. Given that nature has already sampled a rich diversity of functional proteins, many design efforts begin with existing scaffolds, narrowing the design space and enabling targeted improvements – such as enhancing the catalytic efficiency or substrate specificity of PETases. However, not every desired function or property is represented in natural proteins. In such cases, entirely new proteins must be designed from first principles. *De novo* protein design addresses this need by attempting to create novel proteins from scratch, expanding the structural and functional repertoire beyond what evolution has produced (*15*). Beyond its possible practical applications, *de novo* design also serves as a test of our fundamental understanding of protein folding and function – an idea captured by Richard Feynman’s famous quote: *"What I cannot create, I do not understand"*, making a successful designed protein a proof of deeper understanding (*32*, *33*).

2.2 Eras of *de novo* protein design

Given the challenges involved in designing entire proteins from scratch, one might assume that *de novo* protein design is a relatively recent subset of protein design. However, the first attempts date back approximately 50 years. In retrospect *de novo* protein design can be divided into three historical eras (*34*) and an ongoing fourth one (Figure 3). Each era must be understood not only in the context of prevailing knowledge about proteins but also in terms of available methodologies. The first era, starting in the late 1970s, was dominated by manual protein design using physical models. In this first era, modern gene synthesis was not yet available, and protein constructs were achieved by solid-phase peptide synthesis restricting the length to roughly 30-50 amino acids. Nevertheless, pioneering achievements, such as a betabellin protein that mimics β-sandwich proteins, were achieved and provided valuable insights paving the way for future milestones in protein design (*35*, *36*).

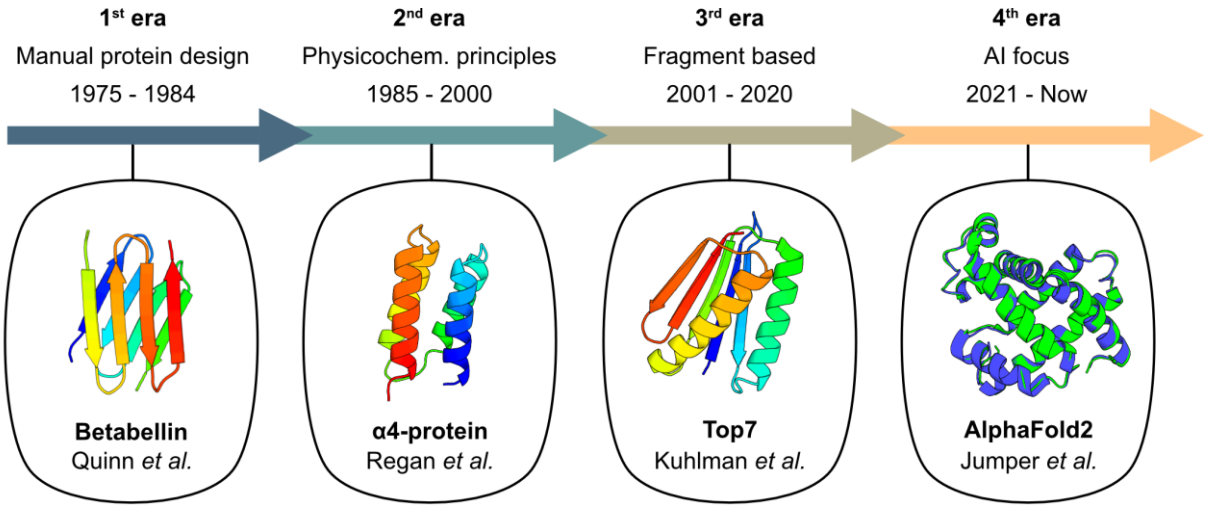


Figure 3: Timeline and milestones of *de novo* protein design. For each era the approximate timeframe, a characteristic and a milestone from the highlighted publications are displayed (*4*, *35*, *37*, *38*). For the displayed structure of betabellin an optimized sequence from Quinn *et al*. was used for structure prediction. For the displayed structure of the α4-protein the reported sequence from Regan *et al*. was used for structure prediction. Betabellin, α4-protein and Top7 (PDB: 1QYS) are shown as cartoon representations and colored in rainbow coloring N- to C-terminus. For AlphaFold2 a structural comparison between the structure prediction (blue) of Myoglobin and its crystal structure (PDB: 1MBN, green) is displayed. AlphaFold2 is not to be described as *de novo* protein design but based on its high impact on the field, it is displayed as the starting point of the fourth era. For all generated structure prediction ColabFold v1.5.5 was used (*39*).

The second era of *de novo* protein design, spanning the mid-1980s to the early 2000s, was driven by computational approaches guided by fundamental physicochemical principles. Advances in computational power with Molecular Dynamics (MD) simulations and structure determination based on X-ray crystallography and Nuclear Magnetic Resonance (NMR) facilitated progress in the field. This era focused on the generation of protein backbones by mathematical equations and sequence design by sidechain repacking algorithms. Although constrained by the existing methodologies, remarkable breakthroughs were achieved, such as the *α*4-protein, the first *de novo* designed protein with a globular conformation in aqueous solution, or the successful repacking of small domains of natural proteins via automated sequence design (*37*, *40–42*).

The third era emerged in the early 2000s, driven by fragment-based and bioinformatically informed computational methods. The expanding Protein Data Bank (PDB) (*43*) provided a growing repository of structural data enabling the deconstruction of proteins into smaller fragments with defined sequences and interaction patterns. Additionally, advances in computing power, along with improvements in the synthetic manufacture of deoxyribonucleic acid (DNA), opened up new possibilities for protein design (*15*). During this era an outstanding milestone was achieved with the design of Top7, a protein with an entire novel fold not observed in nature demonstrating the potential of *de novo* protein design (*38*).

The ongoing fourth era of *de novo* protein design is characterized by the integration of machine learning techniques (*44*). While AlphaFold2 – a groundbreaking algorithm for protein structure prediction – is not a tool for protein design, its development marked a turning point by showcasing the potential of machine learning in protein science (*4*). This breakthrough has driven the creation of machine learning algorithms tailored for protein design, outperforming the tools developed in previous eras and enhancing the complexity of achievable designs. A detailed discussion of AlphaFold2 and the role of machine learning in protein design is provided in section 3 of the introduction.

2.3 The protein design software suite Rosetta

During the third era of *de novo* protein design, numerous computer programs and algorithms have been developed to harness the full potential of protein design (*45–49*). Among them, Rosetta stands out as one of the most used software packages for modeling macromolecular structures. Rosetta was originally developed in the mid-1990s to tackle protein structure prediction, but has since expanded to include diverse modelling tasks, such as protein-protein or small molecule docking (*50*). Despite the increasing number of packages, they all rely on a fundamental component – Rosetta's energy function (*51*). While the energy function has been continuously refined, its core principle remains unchanged. It scores the energy of a macromolecular system using a linear combination of weighted terms that balance physics-based and statistically derived potentials. These terms incorporate various factors including van der Waals energies, hydrogen bonds, electrostatics, disulfide bonds, residue solvation, backbone torsion angles, sidechain rotamer energies, and an average unfolded state reference energy.

To address the protein folding problem, Rosetta *ab initio* structure prediction was developed. This method constructs the tertiary structure of the query sequence by assembling small residue fragments of known structures with similar local sequences (*52*). To navigate the energy landscape, escape local minima and convert to the global minimum, a Monte Carlo simulated annealing procedure is employed (*53*). The iterative prediction, driven by Monte Carlo sampling, begins with a fully extended conformation. At each step, a randomly selected fragment window is sampled and scored using an adapted energy function. Normally, conformations with increased energy are rejected, but they can be accepted with a certain probability based on the Metropolis criterion. The probability of acceptance is based on the Boltzmann factor including the energy difference of the sampled conformations, the

Boltzmann constant and a temperature factor to regulate the exploration of higher energy states. During the simulated annealing protocol, additional energy terms are added to enhance the accuracy in structure sampling, while the temperature factor is gradually decreased to lower the acceptance rate of higher-energy conformations. This approach allows for the generation of many predictions for a given sequence in a short time, effectively sampling a broad range of possible conformations. To evaluate the predictions, the energies are scored and compared. A prediction is typically considered promising if there are structures in a global energy minimum with a substantial energy gap to alternative conformations resulting in a so-called funnel shaped landscape. Comparable Monte Carlo samplings are employed in other Rosetta applications, such as sequence optimization, as they provide an efficient and computationally feasible method for exploring conformational space (*54*). While Rosetta *ab initio* structure prediction demonstrated encouraging results compared to techniques available at that time (*55*), it is outperformed in accuracy, consistency and sequence length by nowadays established deep learning methods.

Although one of Rosetta's original purposes was to solve the protein folding problem, it also provides a range of tools for other design tasks such as protein-protein docking, binder design or enzyme design (*56–61*). Among these tools within the Rosetta software suite, RosettaRemodel stands out as a versatile framework capable of including diverse challenges in the design objective, such as loop insertion or deletion, symmetrical design or even full *de novo* structure modeling (*62*). RosettaRemodel relies on several key features of Rosetta: (1) its energy function, (2) fragment-based structural building derived from the PDB, and (3) iterative sequence design for optimization. A key component of RosettaRemodel are so-called blueprint files, which define each residue and their corresponding secondary structure within a protein. For design purposes these blueprints can be modified by deleting or adding residues with specific constraints allowing full customizability to achieve the desired design objective. The potential of RosettaRemodel is evident in multiple successful design applications, including the circular permutation of an epitope scaffold for binding an anti-HIV (Human Immunodeficiency Virus) antibody, and the design of a chimeric protein combining an HIV glycoprotein with a human granulocyte-macrophage colony-stimulating factor to enhance immune responses (*63*, *64*).

3 AI revolution in protein science

As discussed previously, advances in fields not directly related to protein design – such as synthetic gene manufacturing – had significantly influenced its development throughout its history. One of the most rapidly advancing fields today is artificial intelligence (AI), particularly deep learning neural networks, which have revolutionized both everyday applications and advanced scientific research (*65*). These networks, inspired by the interconnection of neurons in the human brain, can learn to identify patterns from large datasets and perform complex tasks such as image recognition and natural language processing (*66*). Typically, these networks are composed of three connected layers: (1) an input layer that receives the raw data, (2) hidden layers that process the received data, and (3) an output layer that generates the final output. The learning process involves the adjustment of parameters in the hidden layers to minimize a loss function, which quantifies the difference between the generated output and the actual target, thereby guiding the model toward improved accuracy (*44*). Another key factor influencing the performance of machine learning algorithms is the availability of large amounts of high-quality data. In this regard, protein science is particularly well-suited for AI usage, as it provides a wide variety of diverse and publicly accessible well-curated databases. Consequently, numerous machine learning approaches were developed, supporting applications such as structure prediction, sequence design or *de novo* protein design (Figure 4).

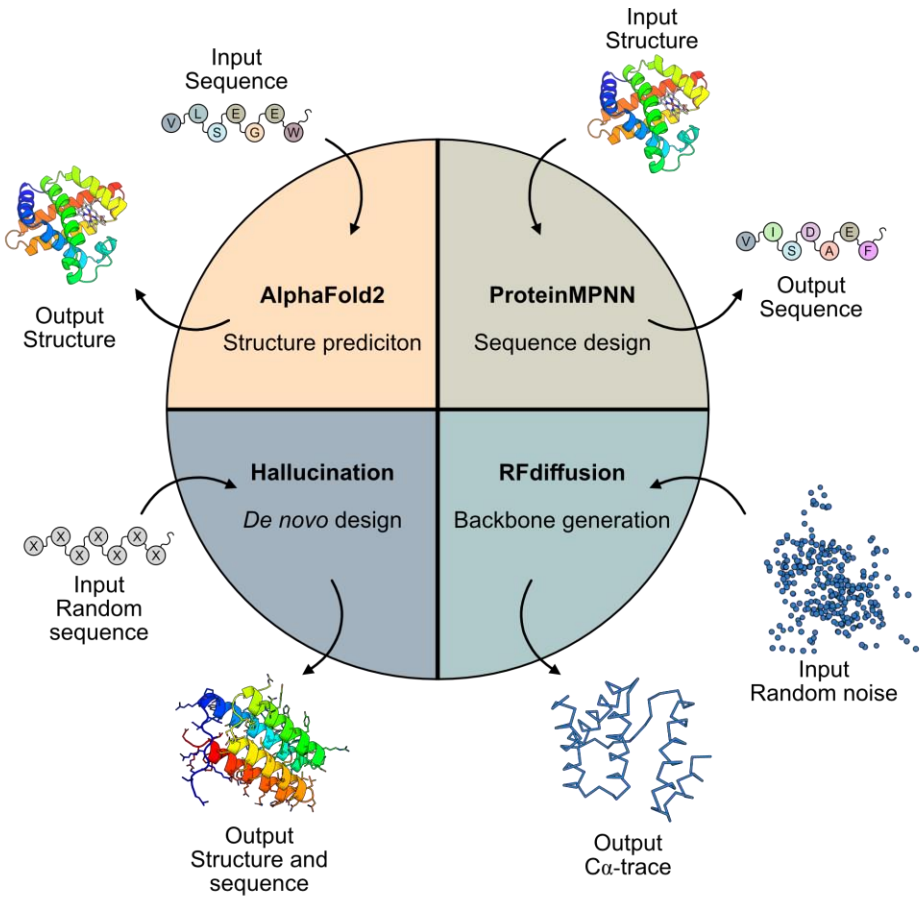


Figure 4: Overview of different machine learning approaches used in protein design. AlphaFold2 is used for structure prediction and takes a protein sequence as input and predicts the corresponding structure (*4*). ProteinMPNN is used for sequence design and takes a protein structure as input and predicts its amino acid sequence (*67*). Hallucination generates *de novo* proteins by optimizing random sequences of certain lengths and produces structures and sequences (*68*). RFdiffusion is used for backbone generation as it generates Cα-traces by denoising random noise (*69*). For AlphaFold2 and ProteinMPNN the displayed structure is myoglobin (PDB: 1MBN, cartoon representation, rainbow coloring N- to C-terminus). For Hallucination and RFdiffusion Hallucination ColabNotebook v.1 or RFdiffusion ColabNotebook v1.1.1. was used to generate the corresponding output (Hallucination: structure in cartoon representation, rainbow coloring N- to C-terminus; RFdiffusion: Noise and Cα-trace colored in blue).

3.1 AlphaFold2 – Solving the protein folding problem?

Determining the structure adopted by a specific sequence without time-consuming and labor-intensive experiments is a long-lasting dream in protein biochemistry as the sequence-structure-relationship holds valuable insights into protein function, evolutionary relationship and biological mechanisms (*5*). The need for reliable structure prediction tools is highlighted by the growing discrepancy between the number of known protein sequences and experimentally validated structures. This discrepancy continues to grow, impacted by high throughput metagenomic experiments identifying billions of sequences per experiment, outpacing the growth of deposited structures in the PDB (*70*). To bypass the slow process of experimental structure validation and help close the gap, robust and accurate structure prediction tools are a necessity. In this context, the scientific community has been organizing the Critical Assessment of protein Structure Prediction (CASP) competition since 1994 in which participants can validate their developed tools via blind prediction (*71*).

In 2018, DeepMind, an AI-focused company owned by Google, participated in the CASP13 competition for the first time with its structure prediction tool AlphaFold (*72*). AlphaFold outperformed competitors in the prediction of novel folds with a considerable enhancement in the overall performance compared to previous assessments. Its success was mainly based on two key concepts: (1) utilizing co-evolutionary analysis to map residue co-variation in protein sequence to physical contacts in protein structures and (2) employing deep neural networks to find patterns in sequences and co-evolutionary couplings to build up two-dimensional contact maps (*73*). However, these approaches are not unique to AlphaFold, as neural networks are commonly used in structure predictions tools, driven by the continual growth of the PDB. The real leap forward happened with the development of AlphaFold2 for CASP14 in 2020, where DeepMind entirely reworked its existing model, introducing a novel neural network architecture (*74*). This new architecture consists of two major blocks: (1) the Evoformer module, which generates a first array, that represents a processed multiple sequence alignment (MSA), and a second array, that represents residue pair interactions, utilizing attention-based components, (2) and the structure module, which utilizes these representations to construct a structure (*4*). This rework dramatically enhanced the performance of AlphaFold2, enabling accurate and reliable structure prediction at atomic resolution. Additionally, AlphaFold2 evaluates the reliability of its own prediction using an internal confidence metric known as the predicted local distance difference test (pLDDT), which correlates with the actual local distance difference test when compared to an experimentally solved structure.

The success of AlphaFold2 had a huge impact on biological research (*75*). DeepMind made the code public and established the AlphaFold protein structure database in collaboration with the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) (*76*). By now, it contains over 200 million predicted structures, covering entire proteomes and much of the UniProt database. The AlphaFold protein structure database starts to close the gap between known sequences and structures and provides researchers with fast and easy access to structural data. Despite AlphaFold2's achievements – John Jumper and Demis Hassabis were even recognized with the Nobel Prize in Chemistry for it (*31*) – it has certain limitations, such as the absence of ligands, DNA and ribonucleic acid (RNA), the training on only single chain proteins or no consideration of dynamics. However, updated or newer versions like AlphaFoldMultimer or AlphaFold3 overcame these limitations to some extent (*77–79*).

One ongoing debate is whether AlphaFold2 has truly solved the protein folding problem or not, as it predicts static structures and does not provide insights into protein folding or the underlying energy landscape and dynamic (*75*). Nevertheless, there is no doubt that AlphaFold2 has deepened our understanding of the sequence-structure relationship, accelerated research across numerous fields and started the AI-driven era in protein design.

3.2 Inversion of protein structure prediction

Various deep learning approaches, such as conformation sampling or sequence design, were already in use in protein design before the release of AlphaFold2 (*80–82*), but its release intensified the usage. Inspired by AlphaFold2's success and its basic principles, the structure prediction tool RoseTTAFold was developed (*83*). Whereas AlphaFold2 employs a two-track architecture, in which the sequence and contact map processing are completed before the generation of the atomic coordinates, RoseTTAFold features a three-track architecture. This architecture enables the bidirectional information flow between all levels of data allowing a more integrated data processing. While AlphaFold2 outperforms RoseTTAFold in structure prediction tasks, RoseTTAFold had a great impact on protein design as its flexibility and adaptability provides an excellent framework for various protein design tasks.

The potential of structure prediction networks for *de novo* protein design was first demonstrated with trRosetta (*84*), a predecessor of RoseTTAFold, by the inversion of the prediction to generate new protein sequences and structures (*68*). In this approach, called hallucination, the process begins by passing a random sequence into the neural network that predicts an initial contact map (Figure 5). As a random sequence is unlikely to fold into a defined structure, the initial contact map shows no pattern. To improve the sequence, the network applies an iterative Monte Carlo sampling strategy: one residue is mutated at a time, and the corresponding contact map is re-evaluated (*85*). Through repeated cycles, the algorithm aims to optimize the contact map via its loss-function, gradually sharpening the map's features to resemble those of structured proteins. After sufficient optimization steps, the resulting contact maps showcases characteristics similar to those predicted for naturally occurring sequences. This method generated novel proteins with diverse sequences and characteristics of idealized proteins. To advance this approach toward functional protein design, Wang *et al.* developed constrained hallucination (*86*). Unlike hallucination, which generates proteins without constraints, this method is trained to incorporate a predefined functional site with minimal distortion into the hallucinated protein. The final methodology replaced trRosetta with RoseTTAFold and added multiple loss functions, regarding the preservation of the functional site, distinguishing it from the original approach. Additional problem-specific loss functions are possible, enabling a broader spectrum of possible design tasks. This approach enables the design of functional proteins by preserving key functional sites from known proteins and hallucinating stabilizing scaffolds around them, as demonstrated in the successful design of both metal- and protein-binding proteins. A particularly notable feature of constrained hallucination is its ability to simultaneously design both sequence and structure, in contrast to most design approaches that focus either on sequence or backbone level.

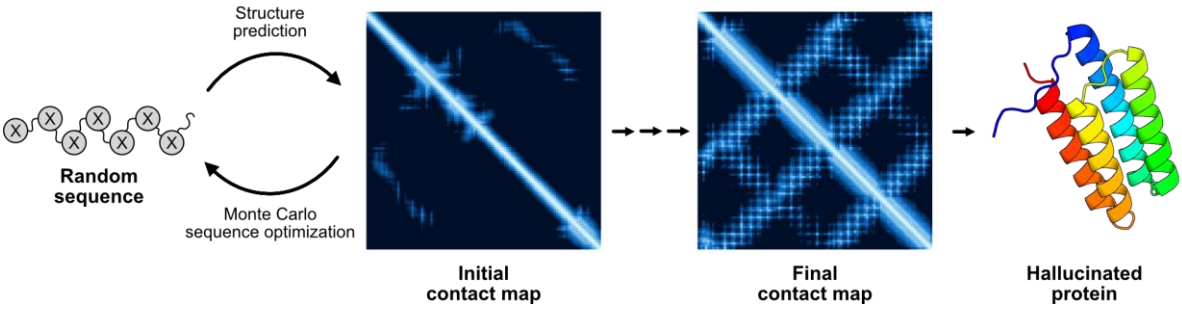


Figure 5: Principle of hallucination. The structure of a random sequence is initially predicted, resulting in a contact map with weak features. The network iteratively optimizes the sequence using Monte Carlo sampling to enhance contact map definition. After sufficient iterations the final contact map exhibits stronger features, corresponding to well-defined protein structures. Hallucination ColabNotebook v.1 was used to generate the hallucinated protein (structure in cartoon representation, rainbow coloring N- to C-terminus). Contact maps were calculated using the ProteinTools server (*87*).

3.3 Neural networks specialized in backbone or sequence design

For the specified task of either backbone or sequence design, multiple deep learning methods have been developed (*88–92*). Two of the most used and well-established tools are RFdiffusion for backbone design and ProteinMPNN for sequence design. These are commonly combined with AlphaFold2, which completes a possible computational design pipeline (Figure 6) and outperforms the joint design by constrained hallucination (*67*, *69*). RFdiffusion is based on a generative diffusion model with RoseTTAFold fine-tuned for denoising tasks to generate highly diverse protein backbones (*69*). The underlying denoising diffusion probabilistic model, commonly used in image generation, offers some advantages for protein design tasks, as it generates diverse structural output while allowing guidance at each denoising step through conditioning toward specific design objectives. The generation of a protein structure happens iteratively: starting from a random residue frame, RFdiffusion predicts a partially denoised version. This prediction is then used to update the previous frame, which serves as the input for the next iteration. Through multiple cycles of denoising, the model progressively refines the structure, producing a plausible protein structure from the initial random noise. For the design of functional proteins, the same strategy used for constrained hallucination can be applied: functional sites are specified in the beginning, preserved throughout the entire diffusion process and supported by the generated scaffold. In addition to this strategy, RFdiffusion can take advantage of its underlying network architecture to generate proteins conditioned on various other design objectives, such as specific symmetries, topologies, or binder design. Across these tasks, RFdiffusion has demonstrated high experimental success rates and structural accuracy. It is important to note, however, that RFdiffusion generates only the Cα-trace of the designed proteins, with sequence design subsequently performed by ProteinMPNN (*67*).

ProteinMPNN, based on a message-passing neural network (MPNN) trained on the PDB, addresses the inverse protein folding problem by taking an input backbone and assigning an optimized sequence to it or to be more precise predicting its sequence. Therefore, a backbone encoder module computes graph nodes and edge features based on the distances between N, Cα, C, O, and virtual Cβ; relative Cα–Cα–Cα frame orientations and rotations; and backbone dihedral angles. The subsequent decoder module assigns an amino acid identity at one position and updates the data representation in an autoregressive manner using the context of previously generated amino acids and calculates a probability distribution from which a new amino acid type is predicted (*44*). *In silico* analysis demonstrated that ProteinMPNN achieves a higher native sequence recovery rate than sequence design with Rosetta. Additionally, structure predictions indicated that the assigned sequences more strongly encode the intended structure than the native sequence. These findings are supported by experimental studies validating ProteinMPNN's high success rate across various design targets including monomers, cyclic oligomers and protein-protein interfaces (*67*, *93–95*). The design workflow combining RFdiffusion and ProteinMPNN is completed by structure prediction using AlphaFold2. This step is necessary for validating whether the designed sequence is likely to fold into the target structure or not. Additionally, AlphaFold2 enables effective filtering of candidate sequences for experimental characterization based on structural similarity to the design model and its confidence score.

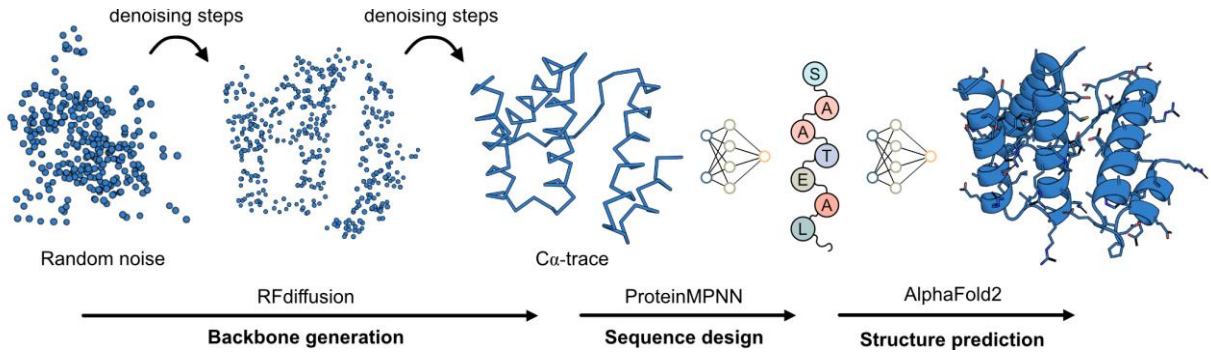


Figure 6: Design workflow with RFdiffusion, ProteinMPNN and AlphaFold2. The backbone generation of RFdiffusion starts with random noise, which is denoised in multiple steps to form a Cα-trace of a protein. The design workflow continues with sequence design with ProteinMPNN, which assigns a possible sequence to the Cα-trace. The designed sequence is used for structure prediction via AlphaFold2 to verify if the designed sequence folds back to the intended structure. Example structure generated via RFdiffusion ColabNotebook v1.1.1.

Despite RFdiffusion's and ProteinMPNN's remarkable success rate and outperformance of traditional physics-based methods, the original versions were limited to protein identities. However, updated versions – RFdiffusionAA and LigandMPNN – were modified to incorporate small molecules, DNA and RNA into the design process (*96*, *97*). These new versions achieve remarkable results in the design of DNA or small molecule binders, but the success rates are lower than for the original versions and their design tasks. Notably, the advance is still ongoing, and the mentioned methods may be already surpassed, highlighting the extraordinary pace of innovation in the field of protein design.

3.4 Possibilities and future challenges for AI

The pace of new milestones in protein design increased significantly in recent years due to the integration of AI. However, several challenges remain. One key concern, if neural networks continue to dominate the field, is the black box problem – our inability to fully comprehend how these models make decisions. While this lack of transparency did not hinder the success of neural networks in protein design, as demonstrated by the methods discussed and their achievements, it contradicts one fundamental objective of protein design: to deepen our basic understanding of proteins. One possible solution is the use of explainable AI (XAI), which focuses on two aspects: interpretability, understanding the model's decision-making process, and explainability, understanding how the model learns during training (*98*). As this improved transparency is crucial not only in protein design but also in safety critical AI applications, such as autonomous driving or medical diagnoses, the field is likely to expand rapidly (*99*). As we have observed during other eras of protein design, advances in adjacent fields often influence the progress of protein design; it might be that XAI will play a role in advancing our fundamental understanding of proteins.

Beyond the black box problem of neural networks, there are still challenges inherent to protein design that must be tackled. One major challenge is the integration of flexibility and conformational dynamics into the design process. Successfully incorporating these features would be highly advantageous, particularly for the design of functional proteins such as antibodies or enzymes (*28*). In parallel, new methods for evaluating these aspects during the design process are needed, as most current approaches rely on static structural models for scoring. Advancements in this area could pave the way for designing responsive protein switches, enabling precise intracellular assembly or even the regulation of gene expression to influence the phenotype (*27*, *28*). Despite remaining challenges, protein design continues to advance rapidly, with increasingly complex proteins (*29*, *100–103*) being designed for specific applications using refined methodologies (*104–106*). As a result, protein design is already proving to be a powerful strategy – and holds even greater promise – for addressing today's scientific and biomedical challenges.

4.1 The TIM-barrel fold – Nature's most versatile fold

Although *de novo* protein design is often associated with the generation of novel folds, such as Top7, the reconstruction of existing protein folds from first principles provides a valuable test of our fundamental understanding of proteins. One outstanding challenge is nature's most versatile fold, the TIM-barrel fold (*107*, *108*). It was identified in 1976, as studies on triosephosphate isomerase (TIM) solved a crystal structure, giving the newly observed topology its name (*109*). More descriptively, the fold is referred to as $(\beta/\alpha)_8$-barrel, as its structure consists of eight parallel β-strands forming a central β-barrel, surrounded by eight α-helices, building a closed, barrel-like architecture (Figure 7) (*110*). The ubiquitous TIM-barrel fold has become a model system in protein science including protein evolution and folding (*16*) and by now many key features of this fold have been unraveled (*108*). Detailed analysis of the central β-barrel revealed a shear number of eight for all TIM barrels, describing the residue shift which is required to return to the same starting point when following a hydrogen-bonded path perpendicular to the strands around the barrel (*111*). Another defining aspect of TIM barrels is the spatial separation of stability and function. Stability is provided by the lower region of the barrel – referred to as stability face – which consists of the hydrophobic barrel core, N-terminal ends of the β-strands and the connecting αβ-loops (*112*, *113*). In contrast, the upper region is referred to as catalytic face, as the C-terminal ends of the central β-strands as well as elongated βα-loops form cavities and anchor catalytic residues making the TIM-barrel fold one of the most versatile enzyme scaffolds in nature.

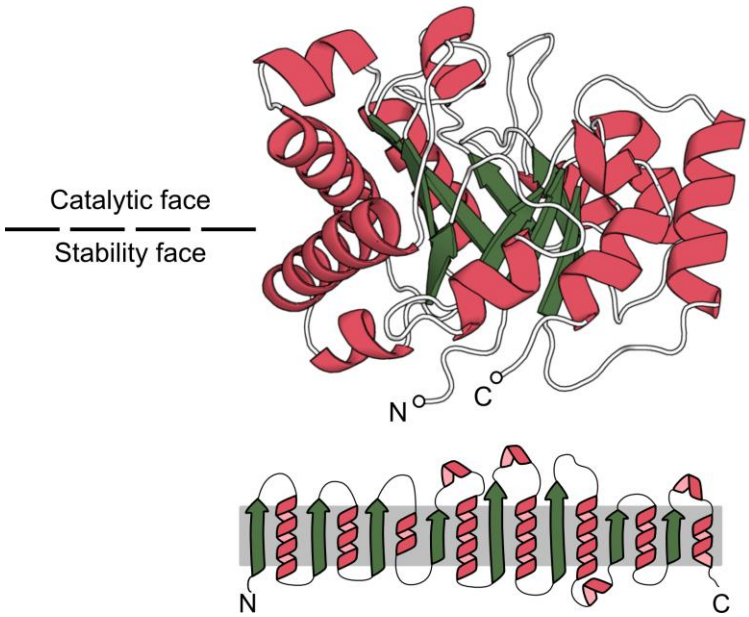


Figure 7: Overview of the TIM-barrel fold. The fold is shown as a three-dimensional structure (PDB: 3UWU) and as a two-dimensional topology scheme. β-sheets are shown in green, α-helices in red and loops in white in the structure and as a line in the scheme. Eight repeating βα-units (highlighted with a grey box in the topology) form a barrel-like structure with termini at the stability face, and additional secondary structure elements within the catalytic face (*112*, *113*).

TIM barrels are present in six out of seven Enzyme Commission (EC) classes, with translocases being the only exception (*110*). Given their enzymatic versatility, it is not surprising that almost 10% of all known enzymes adopt this fold. Despite this enormous number, common patterns can be observed as 85% of these enzymes participate in metabolic reactions (*108*). Additionally, a preference for negatively charged moieties in their substrates, such as a phosphate group, is common, eventually mediated by a positive potential induced by smaller α-helices within elongated βα-loops at the catalytic face (*112*). An outstanding example of the TIM barrel's enzymatic potential is triosephosphate isomerase, an essential key enzyme in glycolysis (*114*). It catalyzes the isomerization of dihydroxyacetone phosphate (DHAP) to glyceraldehyde 3-phosphate (GAP) ensuring the catabolism of DHAP and a net yield of adenosine triphosphate (ATP) from anaerobic glucose metabolism (*115*). This isomerization is catalyzed diffusion limited, reaching catalytic efficiency close to its maximum possible value of 10^9 M^{-1} s^{-1} making TIM a model for enzyme studies since its discovery (*116*). Now 199 crystal structures from 42 different species are available reflecting the long-standing interest in this enzyme and fold (*117*).

4.2 The race for the first *de novo* TIM barrel

Given its biological abundance and outstanding enzymatic potential, the TIM-barrel fold is an attractive target for protein design, particularly for enzyme design. Already in the early 1990s, initial attempts started to design artificial TIM barrels relying on purely statistical analysis of a small set of natural ones. These attempts resulted in proteins with a high content of secondary structure, but which most likely adopt a molten-globule-like state (*118–121*). With increasing availability of structural data on natural TIM barrels and advancements in computational methods, later design attempts explored backbone generation based on parametric features derived from the natural data set (*122*, *123*). These designed proteins again exhibited strong secondary structure content and even showed indications of the intended tertiary structure but suffered from low solubility and conformational stability. A follow up study optimized one design by directed evolution, but structural characterization revealed, rather than the intended TIM-barrel topology, a resemblance to a Rossmann-like fold (*124*). Another attempt by Nagarajan and colleagues in the year 2015 introduced the concept of a symmetrical TIM barrel to simplify the design process (*125*). They utilized Rosetta for both design and *ab initio* structure prediction. Despite providing more insights into essential design principles for the desired fold, the resulting proteins displayed again molten-globule like behavior. Although none of these studies succeeded in designing a TIM barrel, the findings provided valuable insights for future design strategies and the challenges encountered emphasized the complexity of the fold.

With further advancements in protein design and a deeper understanding of the principles of idealized proteins (*126*), Huang *et al*. revisited the challenge of a *de novo* TIM barrel in 2016 (*111*). To make the design of a *de novo* TIM barrel achievable, they simplified their approach by targeting an idealized fold with the highest possible symmetry. Based on the alternating pleat of paired β-strands at the center, the

highest feasible symmetry was four-fold. This symmetry in combination with the α/β rule for idealized proteins – which indicates the orientation of the first residue within a β-strand following an α-helix – and the shear number of eight observed in natural TIM barrels significantly influences the design of the inner β-strands (Figure 8) (*126*). By considering all these principles, they determined that only four identical βαβα units, with no strand register-shift within a unit and a shift of two residues between units, can satisfy all prerequisites. As a direct consequence, these properties influence the necessary characteristics of the α-helices of each βαβα-unit. Specifically, the α-helix between two β-strands with a register shift of zero must be longer and more tilted than the α-helix between two β-strands with a register shift of two. Taking all these principles and requirements into account, the design process started with the generation of backbones with a fixed length of the β-strands and varying lengths for each α-helix and loop in the repeat unit using RosettaRemodel. In the next step these subunits were propagated into four successive tandem repeats to form a full TIM barrel. Subsequently, the structure with the most extensively hydrogen bonded cylindrical sheet was chosen for iterative sequence design, ongoing until the sequences converged to a final amino acid composition. As an additional design step, a circular permutation was performed for some designs to introduce an extra loop at the stability face, and two cysteines were introduced to form a stabilizing disulfide bond. Such a design, namely sTIM11, was crystallized and the solved structure revealed the successful design of the first *de novo* TIM barrel. By comparison with the computational model a high level of accuracy, even on the side chain level, was observed, showcasing the successful design strategy and deep understanding of the TIM-barrel fold. An important outcome of the design strategy next to the high accuracy was an entirely novel sequence for sTIM11 as homology searches indicated only a distinct relationship to natural TIM barrels. With these results the first *de novo* TIM barrel without any evolutionary bias was achieved and given the high potential of its fold, it raised the expectations that tailor-made enzymes were within the reach of *de novo* protein design.

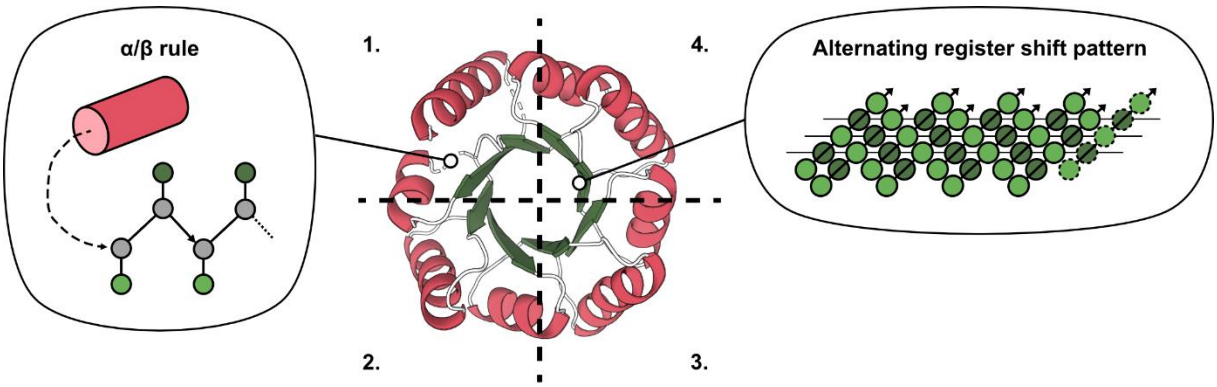


Figure 8: Design principles of the first *de novo* TIM barrel. sTIM11 (PDB: 5BVL) is displayed as cartoon with β-sheets in green, α-helices in red and loops in white. According to the α/β rule, applied to the stability face of sTIM11, the first residue of the β-strand (light green) points away from the previous α-helix (red). sTIM11 is four-fold symmetric, indicated by the dashed line. The inner β-sheet is designed to have no strand register shift within a unit and a shift of two residues between units (dashed β-strand indicates the position of the first β-strand). Not all design principles are shown within the figure. Schemes adapted from Huang *et al.* and Koga *et al*. (*111*, *126*).

4.3 The expanding family of *de novo* TIM barrels

After the successful design of sTIM11 from scratch, further optimization and expansion of the *de novo* TIM-barrel family progressed rapidly. One early optimization addressed two cysteines introduced during the design process to form a stabilizing disulfide bond, however, the bond failed to form as intended. To avoid the potential presence of reactive thiols, both cysteines were reverted to their symmetry-related counterparts, resulting in the variant sTIM11noCys (Figure 9) (*16*). Starting from this variant, the TIM-barrel family was further expanded by an optimization approach focused on hydrophobic repacking (*16*). Stabilizing mutations were introduced in the internal, bottom and top core of the barrel and combinatorially tested. The resulting set of *de novo* TIM barrels, referred to as DeNovoTIMs (Figure 9), showed an increased thermostability, with certain designs remaining folded at 100 °C. In a separate approach inspired by natural TIM barrels, a salt bridge cluster was introduced at the bottom of several previous designed *de novo* TIM barrels (*127*). These variants were denoted with the suffix 'SB' to indicate the presence of the introduced salt bridge cluster (Figure 9). While the insertion of these salt bridge clusters did not affect thermostability, it impacted conformational stability and improved the crystallization properties, thereby contributing further to the optimization and expansion of the *de novo* TIM-barrel family.

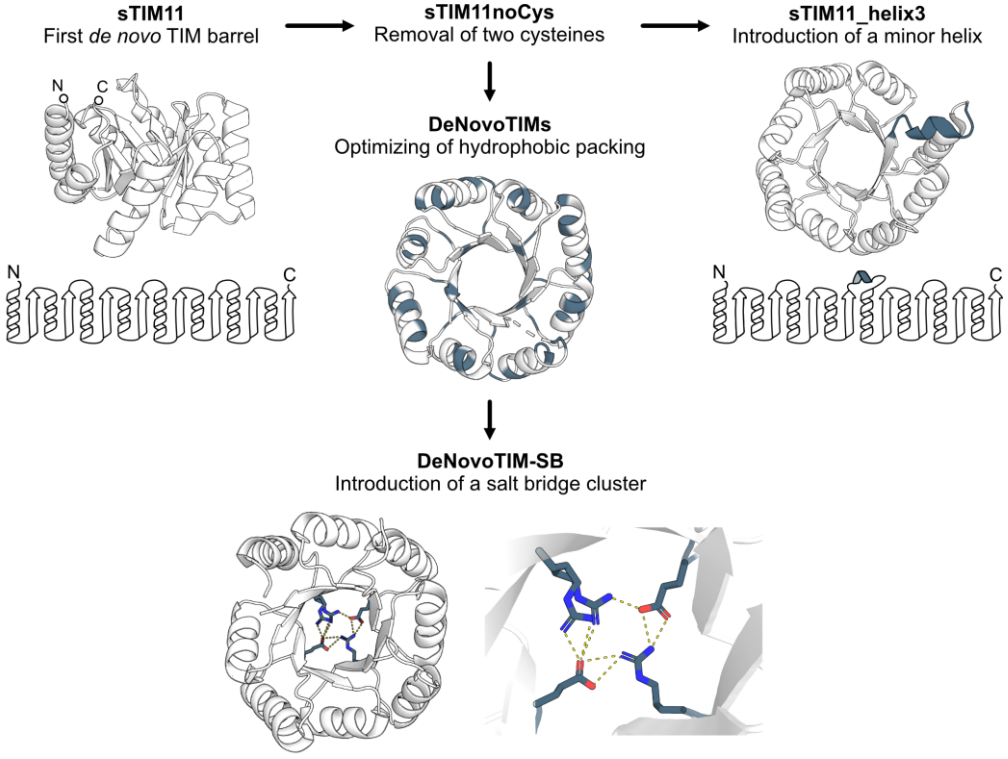


Figure 9: Overview of selected optimized and modified *de novo* TIM barrels. All structures are displayed as cartoon representations in white and each optimization or modification in comparison to sTIM11 is colored in blue. The same coloring scheme is applied to the topology schemes. The first modification of sTIM11 (PDB: 5BVL) was the removal of two cysteines resulting in sTIM11noCys. sTIM11noCys was used to introduce a minor helix at the catalytic face, resulting in sTIM11_helix3 (PDB: 7A8S). In addition, sTIM11noCys was used as a starting point to optimize the hydrophobic repacking of the barrel resulting in multiple designs referred to as DeNovoTIMs (Exemplary structure: DeNovoTIM13; PDB:6YQX). Multiple DeNovoTIMs and sTIM11 were used to introduce a salt bridge cluster at the bottom of the barrel denoted with the suffix SB in their names (Exemplary structure: sTIM11-SB; PDB:7OSU).

As a model system in protein design, the TIM-barrel fold served often as a challenging design target to validate newly developed design pipelines, resulting in additional family members with diversified sequences (*92*, *128*, *129*). The diversity on the structural level was enhanced by the generation of a *de novo* TIM barrel with an ovoid shape deviating from the established circular architecture of previous ones (*130*). Despite the growing number of *de novo* TIM barrels, most retained the highly idealized fold of the original sTIM11 with minimal loops. One notable exception is the extended *de novo* TIM barrel sTIM11_helix3 (Figure 9). This *de novo* TIM barrel is based on sTIM11noCys but has an additional minor helix at the catalytic face. Similar small helices are frequently observed in natural TIM barrels, where they are often involved in phosphate binding. The rationale behind this design was that introducing a secondary structure element at the catalytic face could increase the available surface area and potentially form a pocket, thereby facilitating downstream functionalization with ligand binding or enzymatic activity (*131*). Another outstanding example of diversification and even downstream functionalization was achieved by Caldwell *et al*. (*132*). Utilizing the established four-fold symmetry of *de novo* TIM barrels, they divided the barrel into two parts and fused it with a *de novo* designed ferredoxin. Through homodimerization a TIM barrel with a big cavity on the catalytic face was generated. By introducing additional mutations within this pocket, they enabled highly specific lanthanide binding. A follow-up study demonstrated the potential of this lanthanide-binding protein as a photoenzyme (*133*). The bound lanthanide enables radical C-C bond cleavage of 1,2-diols upon visible-light irradiation, making this the first – and to date, only – enzymatic active *de novo* TIM barrel. Despite this remarkable example, progress toward functional *de novo* TIM barrels with tailor-made reactions is slower and more challenging than expected.

5 Aim of this thesis – Tackling the limitations of *de novo* TIM barrels

After decades of unsuccessful attempts, the design of sTIM11 – the first *de novo* TIM barrel – was a milestone for *de novo* protein design achieving the generation of nature's most prevalent fold without any evolutionary background (*111*). Soon after, variants of sTIM11 with enhanced thermodynamic stability, improved crystallization properties, diversified sequence or oval shape were designed, further expanding the *de novo* TIM-barrel family (*16*, *69*, *127*, *130*). However, while the design of sTIM11 initially raised hopes to achieve tailor-made enzymes and despite one outstanding example of introduced lanthanide binding and photoactivity (*132*), the functionalization of *de novo* TIM barrels have yet to meet expectations. The challenges and reasons for the slow progress become apparent by a comparison between *de novo* and natural TIM barrels (Figure 10). While natural TIM barrels employ extended loops, secondary structure elements, or even additional domains at the catalytic face to form pockets and anchor catalytic residues, *de novo* TIM barrels lack these features due to their highly idealized structure with minimal loops. Moreover, the possible structural diversity of the catalytic face is further decreased by a

circular permutation common to many *de novo* TIM barrels reducing the number of designable βα-loops on the catalytic face in contrast to their natural counterparts. Therefore, structural modifications to deviate from the idealized architecture of *de novo* TIM barrels are crucial to support functionalities such as small molecule binding or enzymatic activity.

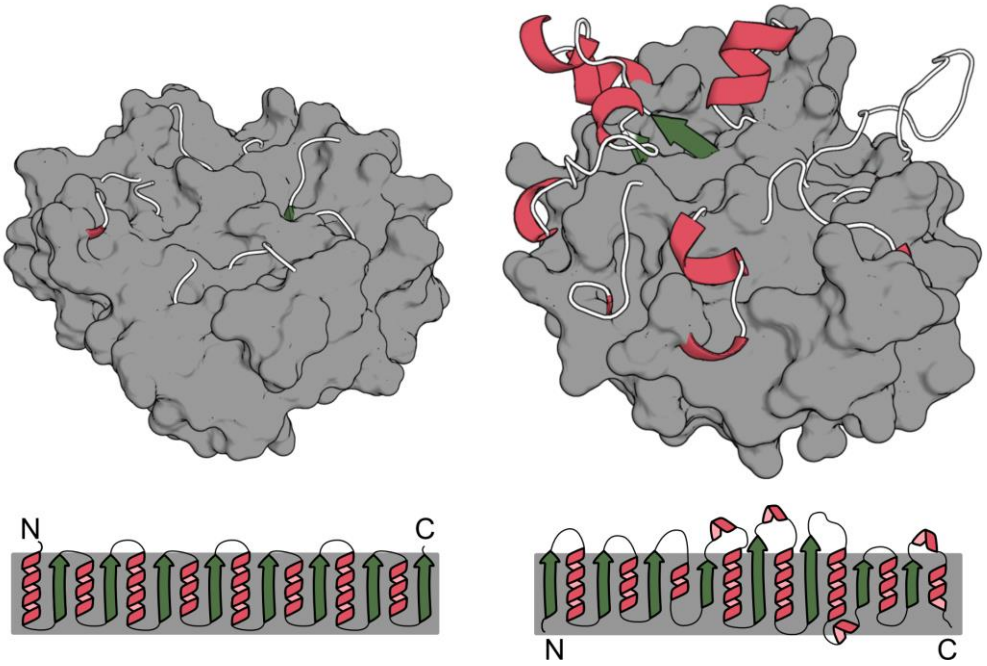


Figure 10: Comparison of *de novo* TIM barrels with natural ones. sTIM11 (left structure, PDB: 5BVL) and triosephosphate isomerase (right structure, PDB: 3UWU) are shown as cartoon representation and topology. β-sheets are shown in green, α-helices in red and loops in white in the structure and as a line in the topology scheme. The surface of the core barrel (highlighted with a grey box in the topology) is displayed in grey for both structures. The idealized structure of sTIM11 results in limited surface area and no pockets as it features only short loops in addition to the barrel architecture. In natural TIM barrels, additional structural elements and elongated loops at the catalytic face enable the formation of functional pockets. sTIM11 is circular permutated in comparison to the natural TIM barrel.

This thesis aims to overcome the limitations of *de novo* TIM barrels and advance their functionalization. Hereby, the focus is on the structural diversification of existing *de novo* TIM barrels by introducing secondary structure elements to generate pockets for possible binding or enzymatic activity. Various *de novo* TIM barrels were generated which anchor structural extensions and provide pockets for the incorporation of functional residues in a downstream step. Obtained insights from these *de novo* TIM barrels were applied to generate a new set of *de novo* TIM barrels with tailor-made extensions for a specific enzymatic activity directly, achieving multiple active *de novo* TIM barrels and providing the framework for future functionalized *de novo* TIM barrels.

Synopsis

Extending a *de novo* TIM barrel with rational designed motifs – αTIMs

A potential strategy for functionalizing *de novo* TIM barrels involves a two-step approach: first, introducing structural extensions in the scaffold protein to form a suitable pocket, and second, incorporating functional residues into this pocket to enable a specific activity. Previous work by Wiese *et al.* attempted to generate such a pocket by introducing a small helix at the catalytic face (*131*). However, the crystal structure of the modified barrel revealed that the introduced helix deviated from the intended motif, and its limited size was insufficient to generate a pocket, hindering progress toward functional *de novo* TIM barrels. Nonetheless, these initial attempts provided valuable insights. We hypothesized that achieving pocket formation requires larger structural extensions that are as precisely controlled as possible to avoid unintended deviations.

In our first publication, we applied these principles by introducing larger coiled coils, which offer rational designability and tunable size based on their heptad repeat sequence (*134*). To generate suitable coiled coils for insertion, we selected coiled-coil sequences from literature and used Rosetta *ab initio* structure prediction for modeling (Figure 11). Since the project started during the third era of *de novo* protein design – before AlphaFold2 and comparable AI-based tools – all computational methods relied on physics-based approaches. Since the obtained full-length coiled-coil motif was disproportionately large compared to the TIM barrel, we truncated the sequence slightly. The predicted structure of the shortened motif did not resemble a coiled coil but adopted a helix-loop-helix motif, however based on the high-quality, as indicated by a funnel shaped landscape, we proceeded with the fragment incorporation. Using RosettaRemodel, we tested different βα-loops of sTIM11 for the insertion and identified the $\beta_4\alpha_5$-loop as the most suitable one. As the inserted motif revealed high conformational flexibility with respect to the TIM barrel, we constructed a continuous motif with the TIM barrel's outer α-helix rigidifying the extension. To restore the originally intended coiled-coil topology we performed sequence optimization of the motif based on the characteristic typical heptad sequence pattern. However, this optimization did not restore the intended topology, though it further stabilized the motif. Based on this workflow we selected two designs – αTIM1 and αTIM2 – which differ in the amino acid sequence in their extension. In addition, we rationally designed a third variant, αTIM3, by introducing a helix-capping motif into the extension of αTIM1 to further stabilize it. Our computational workflow utilized sTIM11 as the starting scaffold, however further optimized variants were developed in the meanwhile. For experimental characterization, we introduced our designed extensions into sTIM11noCys. Initially, we tested αTIM1, αTIM2 and αTIM3, but only αTIM2 could be produced in sufficient yields and showed indication of a successful formation of the intended motif. Based on these early experimental insights, we identified αTIM2 as the most promising design and subsequently introduced its extension a second time. Utilizing the four-fold symmetry of the TIM barrel and given the positions of the termini,

we incorporated the motif in the $\beta_2\alpha_3$- and $\beta_6\alpha_7$-loop without further optimization generating αTIM2-2. We then proceeded with the experimental characterization of αTIM2, featuring one extension, and αTIM2-2, featuring two extensions.

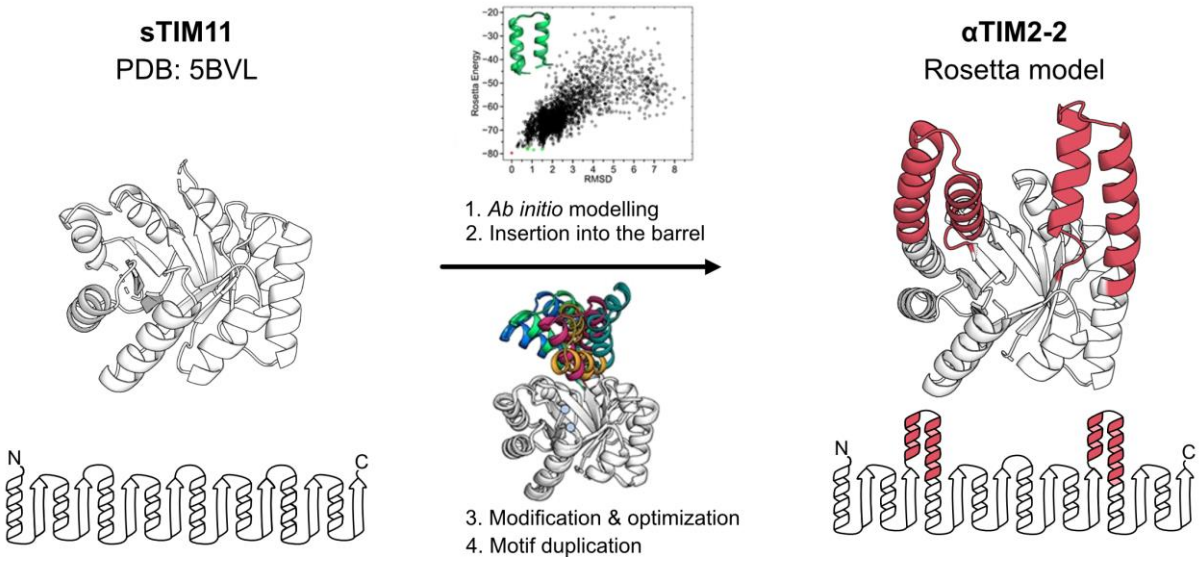


Figure 11: Design overview of the αTIMs. The starting scaffold for the design workflow was sTIM11 (PDB: 5BVL), displayed as cartoon representation and topology in white. The design pipeline started with the *ab initio* modelling of a small helix-loop-helix motif resulting in a good prediction based on the funnel shaped landscape (figure above the arrow). This motif was inserted in sTIM11 resulting in models with a high conformational flexibility of the inserted fragment (figure below the arrow). To reduce this flexibility a continuous helix with the outer helix of the barrel was constructed and the motif was further optimized. After initial experimental characterization the designed motif was duplicated in other regions of the barrel without further modifications resulting in αTIM2-2, displayed as cartoon representation and topology with all extensions in comparison to sTIM11 highlighted in red. Figure adapted from Kordes *et al.* (*134*).

For those two proteins, size exclusion chromatography with multi-angle light scattering (SEC-MALS) measurements showed a stepwise increase in the hydrodynamic radius, while CD spectroscopy revealed an increase in α-helical content with each extension compared to sTIM11noCys, supporting a correct incorporation of the helix-loop-helix motifs. Although the melting temperature remained comparable to the original scaffold, each additional introduced motif reduced the $\Delta G_{25\,°C}$ indicating a modest destabilization associated with the extensions (Figure 12A). Apart from CD measurements, we did not obtain structural data, such as X-ray crystallography, NMR or SAXS, to validate the formation of the helix-loop-helix motif. However, with the release of AlphaFold2 in the meanwhile, our Rosetta designs were supported by the AlphaFold2 predictions, showing high confidence and only minor deviations within the angles of the extensions (Figure 12B). Additionally, the neural network PUResNET, designed for ligand-binding site prediction, identified potential binding sites formed by the inserted motif above the barrel, underscoring the potential for downstream functionalization with ligand binding or substrate recognition sites (Figure 12C). However, several factors may complicate downstream functionalization. The pockets are relatively solvent exposed and not tailor-made for a defined function and no exact atomic coordinates are available due to the absence of an experimentally solved structure. Nevertheless, even if the generated proteins are not directly suitable for functionalization, they provided valuable insights for future functionalization attempts. We learned that the size and identity of the extensions

were sufficient to contribute to pocket formation, the selected insertion points and loops resulted in well-behaving proteins and *de novo* TIM barrels can successfully accommodate multiple structural extensions.

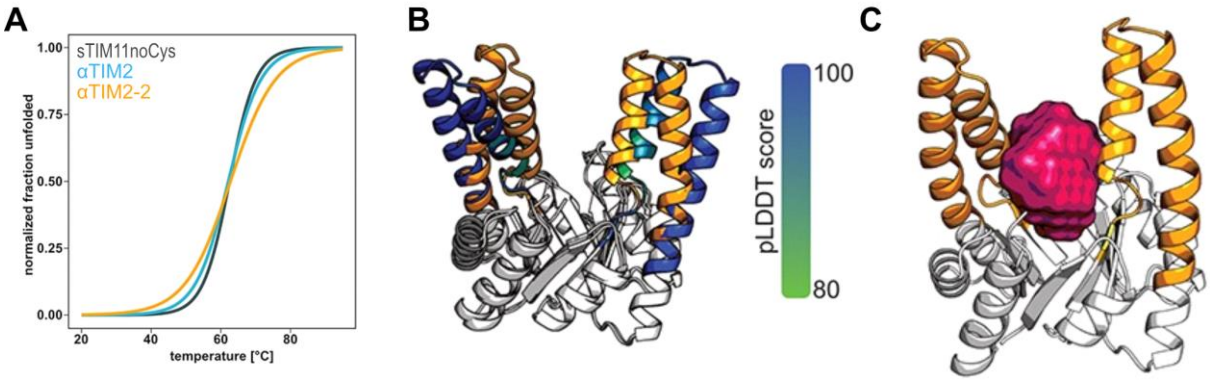


Figure 12: Overview of the characterization of the αTIMs. A. The designed proteins αTIM2 (blue) and αTIM2-2 (yellow) showed a progressive destabilization with each extension, compared to the base scaffold sTIM11noCys (grey), as observed in melting curves monitored by CD. **B.** An AlphaFold2 prediction (barrel in white and extensions colored according to the pLDDT score) indicated the correct formation of the inserted fragment but showed minor deviations in the angles of the extensions in comparison to the Rosetta model (yellow). **C.** A pocket (displayed as surface in red) was predicted by PUResNET within the Rosetta model enabling downstream functionalization. Figures taken from Kordes *et al.* (*134*).

Structural extensions in the era of deep learning – HalluTIMs

The design of the αTIMs was conducted during the third era of *de novo* protein design, but by its ending, the field entered the fourth era with newly available and powerful tools like AlphaFold2, constrained hallucination and ProteinMPNN. Inspired by the structural diversity observed in hallucinated proteins, we applied constrained hallucination to generate a new set of *de novo* TIM barrels referred to as HalluTIMs and described in my second publication (*135*). This set was diversified with two or three extensions at the catalytic face to further enhance the surface area and likelihood of pocket formation (Figure 13). Based on our knowledge from the αTIMs regarding well-behaving insertion points and required extension size, we selected suitable fragment lengths and insertion points within the chosen base scaffold sTIM11-SB. We initially chose constrained hallucination due to its ability to generate diverse secondary structure elements. However, the method predominantly sampled α-helical extensions similar to those in the αTIMs, likely a result of our insertion points within helices. Nevertheless, given the successful pocket formation observed in the αTIMs, we proceeded with the sampled motifs and refined them in a second round of constrained hallucination. As this refinement did not improve the pLDDT score of the best designs, we proceeded with sequence optimization of the extensions via ProteinMPNN, followed by structural validation with AlphaFold2. This approach resulted in designs with high pLDDT scores. From these, six candidates were selected for experimental characterization. Three of the selected designs featured extensions in the $\beta_2\alpha_3$- and $\beta_6\alpha_7$-loops, while the remaining three anchored an additional extension in the $\beta_4\alpha_5$-loop. These designs were named HalluTIMX-X, where the first 'X' indicates the number of extensions and the second distinguishes individual proteins within each category.

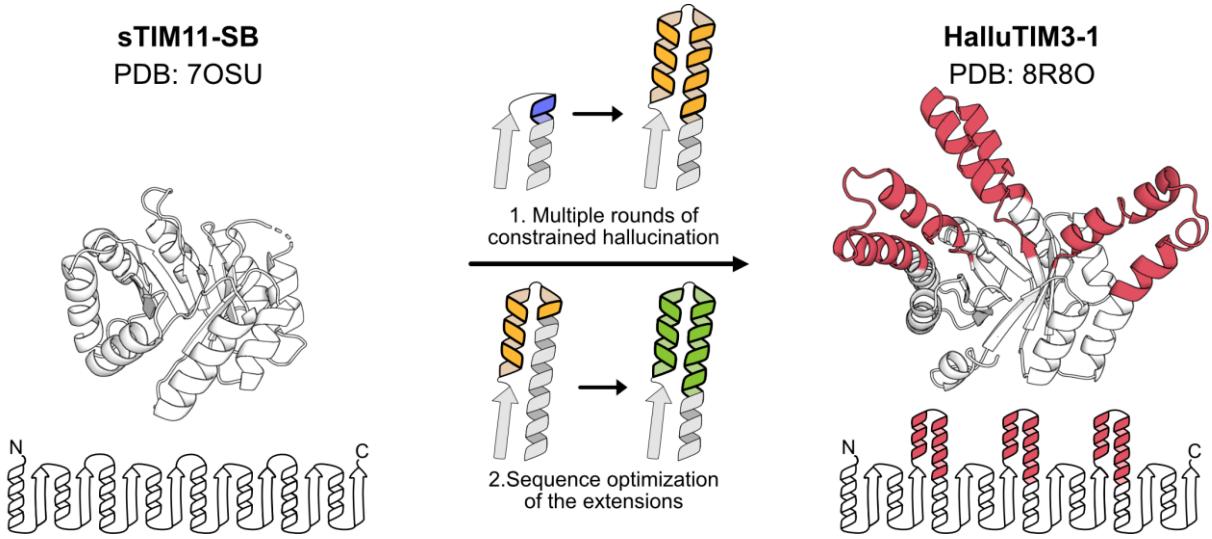


Figure 13: Design overview of the HalluTIMs. The starting scaffold of the design workflow was sTIM11-SB (PDB: 7OSU), displayed as cartoon representation and topology in white. To extend the barrel, two or three insertion regions in the outer helices were chosen and extensions were hallucinated using constrained hallucination (scheme above the arrow, insertion regions in blue, resulting extension in yellow). Multiple rounds of constrained hallucination were performed to refine certain areas of the extensions and subsequently, the entire extensions were used for sequence optimization (scheme below the arrow, refined extension regions in yellow, sequence optimized regions in green). This workflow resulted in multiple HalluTIMs, such as HalluTIM3-1 shown as carton representation (PDB: 8R8O) and topology with all extensions in comparison to sTIM11-SB highlighted in red. Figure adapted from Beck *et al*. (*135*).

The experimental characterization of the HalluTIMs revealed high levels of soluble expression and purification yields across all designs representing a significant improvement over the αTIMs, of which only one initial design resulted in sufficient yields for its characterization. SEC-MALS measurements showed for all designs, except one, a monomeric elution behavior and increase in hydrodynamic radius compared to sTIM11-SB. While CD spectroscopy revealed an increase in α-helicity in comparison to the base scaffold, no stepwise increase with each additional extension was observed. In contrast, to our experience from the αTIMs, the incorporation of the structural extensions did not lead to a destabilization but to an increase in the melting point and $\Delta G_{25\,°C}$ for some designs (Figure 14A). With the crystallization and structure determination of HalluTIM2-2 and HalluTIM3-1, we confirmed the successful formation of the structural extensions, whereby one extension in HalluTIM2-2 was not entirely resolved. Our design HalluTIM2-2 was in close agreement with the crystal structure, whereas HalluTIM3-1 exhibited deviations in the angles of the extensions (Figure 14B). As these discrepancies occurred in regions with a high number of crystal contacts, we conducted SEC-SAXS measurements to investigate their structure in solution. This analysis revealed that the crystal structure, despite its crystal contacts, more closely reflects the solution state than the AlphaFold2 prediction (Figure 14C). However, no structure or prediction of HalluTIM2-2 and HalluTIM3-1 fully agrees with the experimental data indicating considerable flexibility in the introduced extensions. Despite the indicated flexibility, we sought to assess the potential for downstream functionalization using PUResNET to predict potential binding sites within the crystalized HalluTIMs. Hereby, we identified in both HalluTIMs a potential

pocket with variable size. HalluTIM2-2 exhibited a smaller pocket, while HalluTIM3-1 formed a larger one due to the additional extension. These findings highlight the capacity of the HalluTIMs to serve as a scaffold for accommodating ligands of varying sizes.

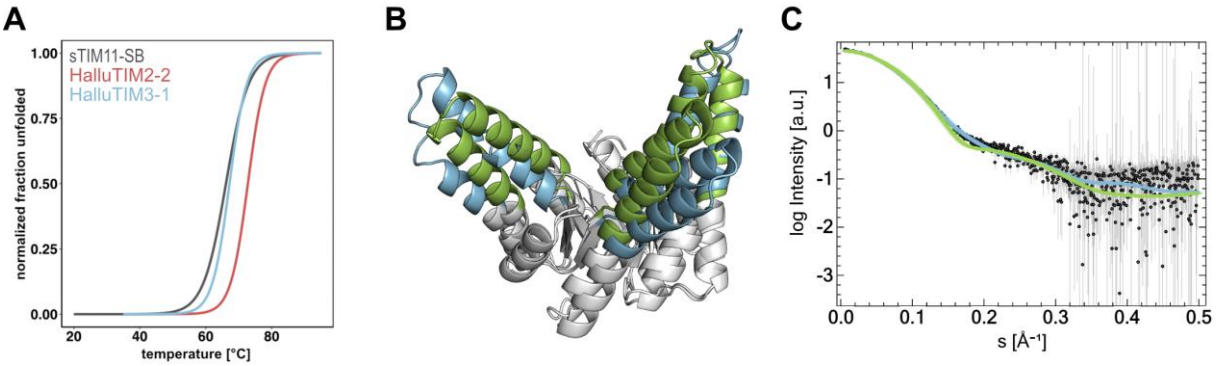


Figure 14: Overview of the characterization of the HalluTIMs. A. The designed proteins HalluTIM2-2 (red) and HalluTIM3-1 (blue) showcased an increased thermostability in comparison to the original scaffold sTIM11-SB (black) within melting curves followed by CD. **B.** A comparison between the crystal structure (barrel in white and extensions in blue) and AlphaFold2 (barrel in white and extensions in green) of HalluTIM3-1 revealed deviations in the angles of the extensions. **C.** SEC-SAXS measurements of HalluTIM3-1 indicated that the crystal structure (blue line) is in closer agreement with the experimental data (black dots) than the AlphaFold2 prediction (green line), whereby both fits are not in perfect agreement with the data suggesting a higher flexibility of the extensions. Figures taken from Beck *et al.* (*135*).

The successful design of the HalluTIMs represents an improvement over the αTIMs in terms of experimental success rate and potential for downstream functionalization. An even higher surface increase is achieved due to the insertion of more extensions and atomic coordinates are available due to multiple solved crystal structures. Additionally, the observed stabilization instead of destabilization indicates a higher robustness of the HalluTIMs eventually necessary for downstream functionalization. Despite representing a valuable steppingstone toward the functionalization of *de novo* TIM barrels, the downstream functionalization of HalluTIMs is likely highly challenging based on the considerable flexibility of the extensions.

Introduction of an enzymatic activity in a minimal TIM barrel – KempTIMs

The introduction of a binding function and especially enzymatic activity requires high accuracy. Realizing diverse functions within a preformed pocket of a scaffold protein is highly unlikely (*136*). Consequently, unlocking the full potential of *de novo* TIM barrels requires moving beyond the conventional two-step strategy of pocket formation followed by downstream functionalization. To address this, we developed a strategy that integrates the design of both a tailor-made pocket and its corresponding function in a single step. This led to the design of a new set of *de novo* TIM barrels, described in detail in our third manuscript (*137*). As proof of principle, we selected the Kemp elimination, a benchmark reaction in computational enzyme design involving carbon-based proton transfer. Accordingly, we named this set KempTIMs. To generate it, we developed a new workflow called CANVAS (Customizing Amino-acid Networks for Virtual Active-site Scaffolding) which

combines neural networks for backbone and sequence design with the protein design software Triad for enzyme design (Figure 15) (*47*). As an additional step to facilitate functionalization, we applied a circular permutation to a *de novo* TIM barrel to increase the number of designable βα-loops on the catalytic face. For this purpose, we used DeNovoTIM6-SB and selected two different permutation points, generating two new variants – NT6-CP1 and NT6-CP2 – that closely resemble the termini positions of natural TIM barrels. In addition to these circularly permuted designs, we included a *de novo* TIM barrel with low sequence identity to sTIM11 and its derivatives, aiming to further diversify our starting points.

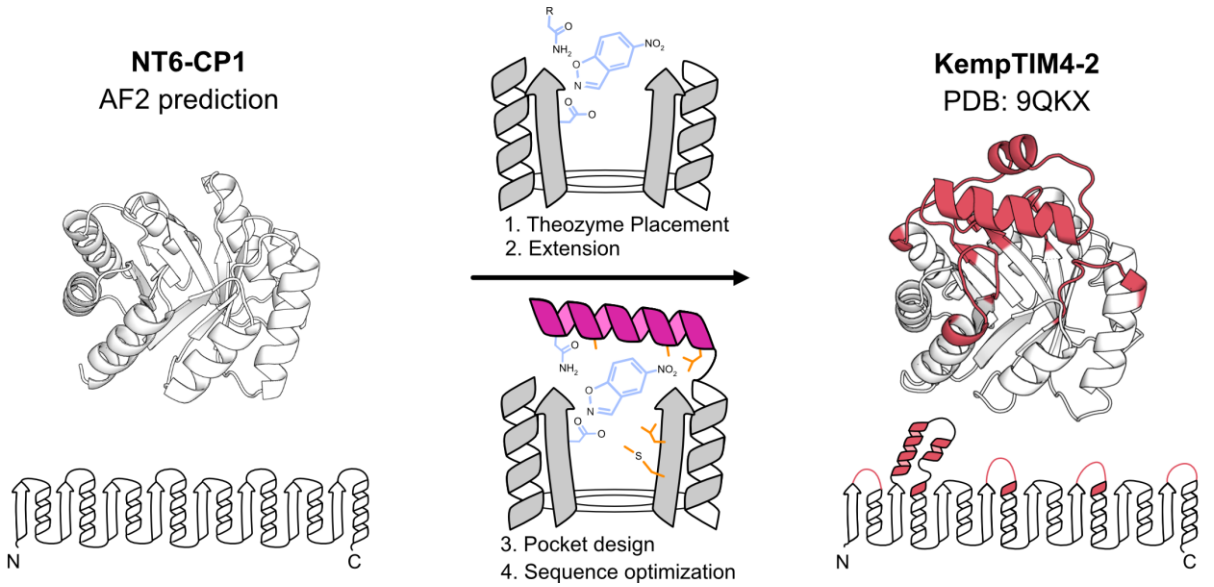


Figure 15: Design overview of the KempTIMs. Starting from a minimal *de novo* TIM barrel, such as the AlphaFold2 (AF2) prediction of NT6-CP1 (cartoon and topology representation in white), we placed one catalytic residue of the theozyme inside the barrel and the second one above the barrel without any connection to the TIM barrel (TIM-barrel scheme above the arrow, theozyme in blue). Subsequently, we extended the minimal TIM barrel to connect the second residue and build up an active site, which was optimized to support the transition state (TIM-barrel scheme below the arrow, theozyme in blue, extension in magenta, residues optimized during pocket design in orange). After initial experimental characterization a sequence optimization of the whole barrel was performed resulting in KempTIM4-2, shown as cartoon representation (PDB: 9QKX) and topology with all extensions in comparison to the minimal TIM barrel highlighted in red. Figure adapted from Beck *et al*. (*137*).

Based on the finding of rather flexible helical extensions in the HalluTIMs, we realized that the backbone design needs to be guided by the desired reaction to generate a suitable pocket. This guidance was achieved by the placement of one catalytic residue of the theozyme, a computational model of an idealized enzyme active site with catalytic groups arranged to stabilize a transition state, above the barrel not connected to the structure. By connecting the residue to the structure using RFdiffusion and ProteinMPNN, we generated extensions anchoring the catalytic residue and forming a pocket for the enzymatic reaction simultaneously. Several of these lids were composed of multiple extensions, consisting of a major helical extension and elongated loops, displaying greater structural variance than the relatively idealized extensions of the αTIMs and HalluTIMs and more closely resembling the structural extensions of natural TIM barrels. After designing the pocket with Triad to achieve the optimal environment for the reaction and performing extensive computational filtering, we identified one active

design during experimental characterization (Figure 16A), showcasing comparable activity to previously designed Kemp eliminases (*138*). This protein, named KempTIM4, is the first *de novo* TIM barrel with an enzymatic function based on a tailor-made extension. Despite being active, KempTIM4 displayed several unfavorable properties, including low expression levels and purification yields, a higher random coil content in CD measurements than expected, major destabilization compared to an idealized TIM barrel and a tendency to aggregate at higher concentrations. To address these issues, a sequence optimization of the whole protein, excluding the active site, was performed using ProteinMPNN, resulting in a stabilized variant named KempTIM4-2. This optimized variant showed improved expression, solubility, structural integrity, and thermal stability. However, it also showed altered enzymatic kinetics, characterized by the absence of saturation (Figure 16A) and reduced dependency on the second catalytic residue, as observed by only a slight decrease in activity in the alanine mutant. Unlike the HalluTIMs, which did not require a sequence optimization of the entire protein, the increased complexity of the extensions and a destabilization of the core due to the pocket design may have led to this necessity. Given the improved properties of the optimized variant, incorporating a sequence optimization of the entire barrel early in the design strategy may be advantageous for the generation of future functional *de novo* TIM barrels.

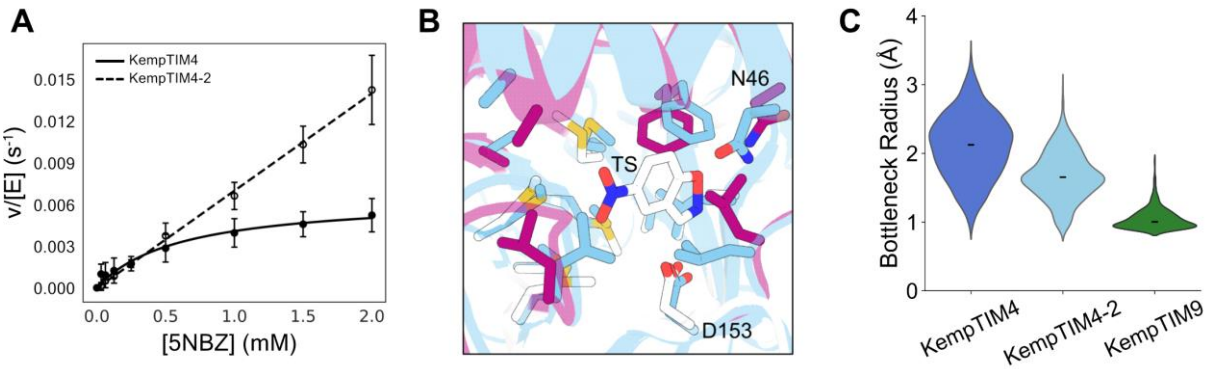


Figure 16: Overview of the characterization of the KempTIMs. A. Enzyme kinetics of KempTIM4 (datapoints as filled symbols and fit as solid line) and KempTIM4-2 (datapoints as open symbols and fit as dashed line); KempTIM4 exhibits substrate saturation within the solubility limit while KempTIM4-2 does not reach saturation. **B.** A comparison between the pocket of the crystal structure (blue) and the designed pocket (minimal TIM barrel and lid colored white and magenta, respectively) revealed correct positioning of the catalytic residue in the barrel (D153), but displacement of the second catalytic residue anchored on the lid (N46) and suboptimal preorganization of the active site. **C.** MD simulations indicated a smaller bottleneck radius as the key difference between active (KempTIM4 and KempTIM4-2) and the non-active design (KempTIM9). Figures taken from Beck *et al*. (*137*).

With the improved properties of KempTIM4-2 we were able to solve its crystal structure. The entire lid was resolved in the structure and in close agreement with the AlphaFold2 prediction underscoring that our design pipeline generated stable extensions. Modest activity can likely be explained by a slight shift of the lid and hence the position of the anchored catalytic residue as well as a poor preorganization within the pocket (Figure 16B). This poor preorganization is likely to contribute to the low affinity toward the substrate and the reason why no density of the transition state analogue was observed in the active site despite the presence in the crystallization condition. But the active site and substrate affinity can be tackled by downstream optimization via directed evolution or computational methods utilizing

the solved crystal structure. Since we observed a high number of crystal contacts within the extended regions, we performed SEC-SAXS to investigate their behavior in solution. This measurement indicated that the lid likely adopts a conformation in solution comparable to its crystal structure, as models lacking the lid or featuring a displaced version showed poorer fits to the experimental data.

Having demonstrated that our design pipeline generated stable extensions, we tried to understand why KempTIM4 and its optimized variant were the only active designs. To investigate this, we conducted MD simulations comparing the active designs to the inactive variant KempTIM9. This particular design was selected because it displayed high expression levels, a well-folded structure in CD spectroscopy and a high melting temperature, which ruled out misfolding as the cause of inactivity. Structurally, KempTIM9 featured a single, larger extension consisting of a helix and an elongated loop, in contrast to KempTIM4, which contained multiple insertions. The MD simulations revealed that only the original design KempTIM4 maintained an active state conformation of both catalytic residues with the transition state analogue. In contrast, the inactive design and KempTIM4-2 showed weaker contacts of the second catalytic residue, anchored on the lid, explaining the reduced dependence on this residue in the case of KempTIM4-2. The key difference between the active and inactive variants was a significantly narrower entrance to the active site in case of the inactive one that likely restricts substrate binding and led to inactivity (Figure 16C). While other potential factors for inactivity cannot be entirely excluded, ensuring greater active-site accessibility, a feature of highly active Kemp eliminases, during lid generation should be considered for future design pipelines to enhance the success rate and catalytic efficiency of the active variants. Although higher catalytic efficiency within the first round of designs is desirable, we successfully designed enzymatically active *de novo* TIM barrels and gained valuable insights for future design strategies.

Overall findings and journey ahead

As is common in protein design, results must be interpreted in the context of existing knowledge and available methodologies. The αTIMs were generated during the third era of protein design, relying primarily on Rosetta and targeting a challenging design objective back then. Consequently, the experimental success rate was low, and no experimentally solved structures were obtained. However, newer AI-based techniques such as AlphaFold2 and PUResNET indicated that the designed extensions and pockets were successfully formed. With the ongoing AI revolution in protein design and the integration of methods such as constrained hallucination into the design pipeline, it is now possible to generate *de novo* TIM barrels with motifs similar to the αTIMs – without rationally guided design – achieving significantly higher experimental success rates and marking a major leap forward in design capabilities. The resulting HalluTIMs showed improved behavior over the αTIMs and demonstrated potential for downstream functionalization, supported by both solved crystal structures and predicted binding pockets. However, the ongoing advances in protein design make it possible to overcome the

process of pocket formation and downstream functionalization and enable one-step functionalization, as demonstrated by the KempTIMs. While the design of the KempTIMs represents a significant step forward, their design would not have been possible without the foundational work and insights gained from the αTIMs and HalluTIMs showcasing a principle of protein design as each design deepened the understanding for the functionalization (Figure 17).

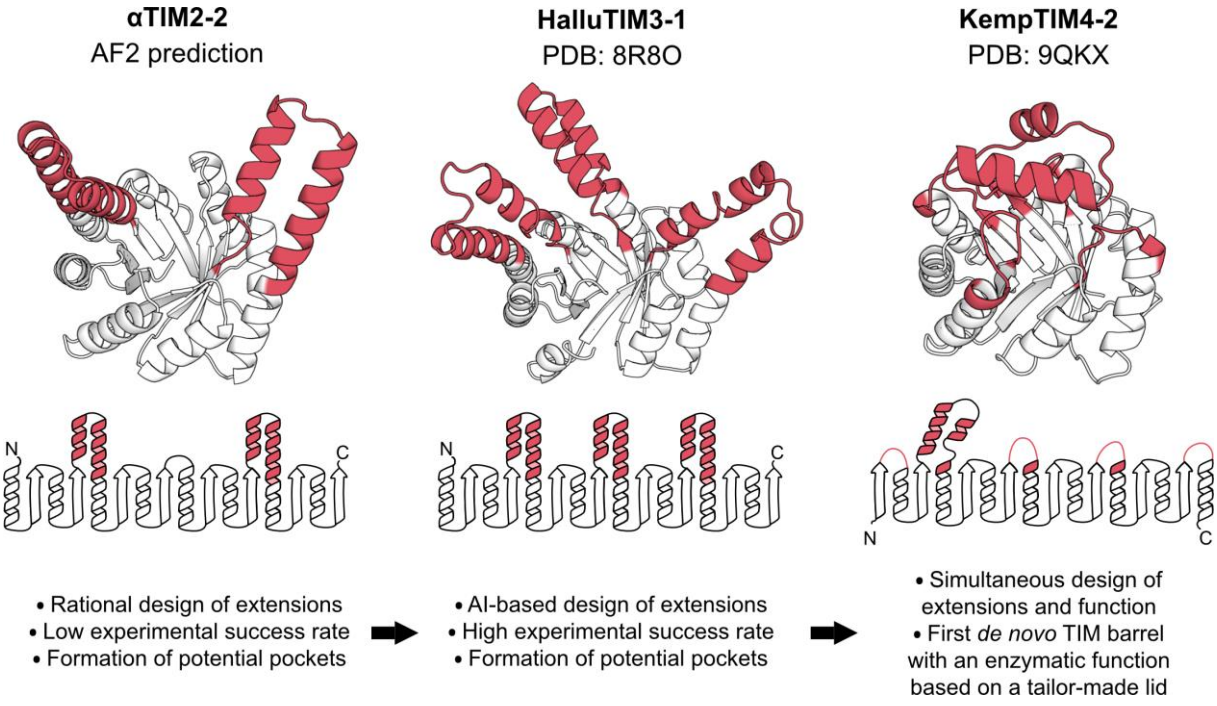


Figure 17: Overview of the designed sets of *de novo* TIM barrels. Each protein is shown as cartoon representation and topology. Structural extensions are colored red and the unmodified parts of the TIM barrel white. Important details of each set are summarized as bullet points. The increased diversity and complexity within the structural extensions are observable with each further design generation.

This thesis aimed to overcome the limitations of *de novo* TIM barrels and advance their functionalization – and successfully met these objectives. The αTIMs and HalluTIMs reduced the limitations of *de novo* TIM barrels as they formed pockets available for downstream functionalization, while also providing crucial insights that enabled the design of enzymatically active *de novo* TIM barrels, the KempTIMs. However, this marks not the end but rather the beginning of the journey for functional *de novo* TIM barrels, given the enormous potential of the TIM-barrel fold. As natural TIM barrels catalyze nearly every known type of reaction, there is immense potential to introduce a variety of functions into *de novo* TIM barrels. This is the start for a future family of enzymatically active *de novo* TIM barrels and the exploration of the full potential of nature's most versatile fold, free from evolutionary bias.

Author Contributions

Publication 1: Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs

Sina Kordes*, Julian Beck*, Sooruban Shanmugaratnam, Merle Flecks, Birte Höcker
Protein Engineering, Design and Selection doi: https://doi.org/10.1093/protein/gzad012
* equal contribution

Sina Kordes and Birte Höcker worked on the conceptualization. Sina Kordes performed the Rosetta design calculations and cloning of the designs. Initial expression, purification and biochemical characterization were performed by Sina Kordes and Merle Flecks. Sooruban Shanmugaratnam and I refined expression, purification, and characterization protocols. I conducted the computational analysis via AlphaFold2 and PUResNET. Sina Kordes and Birte Höcker wrote the first manuscript. Birte Höcker, Sooruban Shanmugaratnam and I led the manuscript's editing and revision. Birte Höcker acquired funding for this study.

Publication 2: Diversifying *de novo* TIM barrels by hallucination

Julian Beck, Sooruban Shanmugaratnam, Birte Höcker
Protein Science, doi: 10.1002/pro.5001

Birte Höcker and I worked on the conceptualization. I performed all *in silico* work. I also carried out expression, purification and biochemical characterization. I set up and screened crystallization conditions. Crystal preparation, data processing, structure building and deposition were done together with Sooruban Shanmugaratnam. SAXS data analysis was also performed together with Sooruban Shanmugaratnam. We curated the data and visualized it together. All authors contributed to writing the first manuscript and to its editing and revision. Birte Höcker acquired funding for this study.

Manuscript 3: Customizing the Structure of a Minimal TIM Barrel to Craft a *De Novo* Enzyme

Julian Beck*, Benjamin J. Smith*, Niayesh Zarifi, Emily Freund, Roberto A. Chica, Birte Höcker
Manuscript under review at *Nature Chemical Biology* (Manuscript-ID: NCHEMB-A250400931-T)
Preprint available on *bioRxiv*, 2025.01.28.635154
* equal contribution

Birte Höcker, Roberto A. Chica, Niayesh Zarifi, and I worked on the conceptualization. I designed the additional circular permutation of starting proteins. The *in silico* design pipeline CANVAS was carried out by Niayesh Zarifi, Benjamin J. Smith (supervised by Niayesh Zarifi), Emily Freund (supervised by me), and myself to generate the tested proteins. We performed the cloning, expression, purification and biochemical characterization except the inhibition assay. The inhibition assay was conducted by Benjamin J. Smith. I carried out the sequence optimization of the best protein variant. Cloning, expression, purification and biochemical characterization were performed by Benjamin J. Smith and

me. I set up and screened crystallization conditions, prepared crystals and processed data. Finalizing the structure was done with support from Roberto A. Chica and Birte Höcker. Deposition of the structure was done by me. I prepared SAXS samples and performed the data analysis. MD simulations and analysis were conducted by Benjamin J. Smith. Data visualization was done by Benjamin J. Smith, Roberto A. Chica and me. Manuscript preparation was led by Benjamin J. Smith, Roberto A. Chica, Birte Höcker and me. Finalization and submission of the manuscript was completed by Birte Höcker. Funding was acquired by Birte Höcker and Roberto A. Chica.

Publication 1

Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs

Sina Kordes*, Julian Beck*, Sooruban Shanmugaratnam, Merle Flecks, Birte Höcker
Protein Engineering, Design and Selection doi: https://doi.org/10.1093/protein/gzad012
* equal contribution

Protein Engineering, Design and Selection, 2023, **36**, 1–8
https://doi.org/10.1093/protein/gzad012
Advance access publication date 14 September 2023
Original Article

peds
protein engineering design & selection
OXFORD

Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs

Sina Kordes[1,2,†], Julian Beck[1,2,†], Sooruban Shanmugaratnam[1], Merle Flecks[1] and Birte Höcker[1,*]

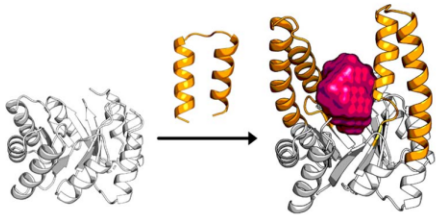
[1]Department of Biochemistry, University of Bayreuth, Bayreuth 95447, Germany
*To whom correspondence should be addressed. E-mail: birte.hoecker@uni-bayreuth.de

[2]Present address: Proteros Biostructures GmbH, 82515 Martinsried, Germany
[†]Sina Kordes and Julian Beck contributed equally to this work.
Edited by: Dr. Roberto Chica

Abstract

Computational protein design promises the ability to build tailor-made proteins *de novo*. While a range of *de novo* proteins have been constructed so far, the majority of these designs have idealized topologies that lack larger cavities which are necessary for the incorporation of small molecule binding sites or enzymatic functions. One attractive target for enzyme design is the TIM-barrel fold, due to its ubiquity in nature and capability to host versatile functions. With the successful *de novo* design of a 4-fold symmetric TIM barrel, sTIM11, an idealized, minimalistic scaffold was created. In this work, we attempted to extend this *de novo* TIM barrel by incorporating a helix-loop-helix motif into its $\beta\alpha$-loops by applying a physics-based modular design approach using Rosetta. Further diversification was performed by exploiting the symmetry of the scaffold to integrate two helix-loop-helix motifs into the scaffold. Analysis with AlphaFold2 and biochemical characterization demonstrate the formation of additional α-helical secondary structure elements supporting the successful extension as intended.

Graphical Abstract



Keywords: ($\beta\alpha$)8-barrel, computational protein design, helix-loop-helix, TIM-barrel

Introduction

De novo protein design aims to expand the protein universe by creating new sequences with predefined properties. Much progress has been made over the years in the design of specific topologies from scratch, including all-α (Regan and Degrado, 1988; Doyle *et al.*, 2015), all-β (Dou *et al.*, 2018; Marcos *et al.*, 2018) and $\alpha\beta$ proteins (Huang *et al.*, 2016) by recapitulating natural folds or even creating completely new topologies (Kuhlman *et al.*, 2003; Pan *et al.*, 2020; Yang *et al.*, 2021; Minami *et al.*, 2023). One particular protein fold has challenged the field for many years: the $(\beta\alpha)_8$- or TIM-barrel fold. Ubiquitous among enzymes it is one of the most common protein topologies in nature and catalyzes a multitude of reactions (Sterner and Höcker, 2005). The TIM-barrel fold is adopted by about 10% of all known enzymes and is present in six out of seven Enzyme Commission (EC) classes. It is organized in eight alternating $\beta\alpha$-subunits that form a central eight stranded, parallel β-barrel surrounded by eight α-helices (Banner *et al.*, 1975; Maes *et al.*, 1999; Wierenga, 2001). A key feature of this fold is the spatial separation of stability and catalytic function. Protein stability is built up by the hydrophobic core of the barrel and the $\alpha\beta$-loops at the N-terminal end of the β-strands ('stability face') (Urfer and Kirschner, 1992; Vijayabaskar and Vishveshwara, 2012), while the catalytically active residues are located at the C-terminal end of the β-strands ('catalytic face') (Nagano *et al.*, 2002). Substrate binding usually occurs *via* a cavity formed at the central surface of the β-sheet, supported by the $\beta\alpha$-loops on the top of the barrel (Thoma *et al.*, 2000).

Due to its recurring use and diversification in nature, the TIM-barrel fold is a highly interesting design scaffold. For more than two decades researchers endeavored to identify the main structural determinants of this fold to create a TIM barrel from scratch (reviewed in (Romero-Romero *et al.*, 2021b)). Finally, Huang *et al.* (Huang *et al.*, 2016) succeeded in constructing an idealized TIM barrel, called sTIM11, using the physics-based Rosetta software suite. Since then, optimized versions of sTIM11 have been designed resulting in an expanded *de novo* TIM-barrel family (Romero-Romero *et al.*, 2021a; Kordes *et al.*, 2022). Recently

Received: July 31, 2023. **Revised:** September 4, 2023. **Accepted:** September 5, 2023.

the repertoire has been further extended to include an ovoid shaped TIM barrel using Rosetta design (Chu *et al.*, 2022) and the application of a learned potential neural network to furnish the sTIM11 structure with new sequences (Anand *et al.*, 2022). Similarly state of the art AI-based methods like AlphaFold2 (Jumper *et al.*, 2021) and proteinMPNN (Dauparas *et al.*, 2022) have been applied by Goverde *et al.* (Goverde *et al.*, 2023) using sTIM11 as a starting point to create *de novo* TIM barrels with a low sequence similarity to the existing variants. Despite the growing number of *de novo* TIM barrels, all still miss the crucial feature of cavities, pockets or extended loops on the 'catalytic face' suitable for ligand binding or catalytic activity compared to natural TIM barrels.

For the construction of a functionalized *de novo* TIM barrel extended surfaces or pockets are an essential prerequisite. Different approaches could be followed to achieve this, ranging from elongation of unstructured loops, *via* the introduction of small structured elements, up to the insertion of complete domains. Wiese *et al.* (Wiese *et al.*, 2021) diversified sTIM11 by the incorporation of a small helix into one of the $\beta\alpha$-loops. In another work, a designed ferredoxin fold was fused to a split TIM barrel variant to create a homodimer with a cavity that could be functionalized by introducing a metal-binding site (Caldwell *et al.*, 2020). These two approaches showcase the ability of *de novo* TIM barrels to serve as a starting point for the introduction of further extensions and functions.

In this work we aimed to extend the family of *de novo* TIM barrels by adding larger secondary structure elements to the 'catalytic face'. We decided to follow up on the small helix insertion of Wiese *et al.* and introduce an elongated helical motif. To obtain a suitable secondary structure fragment, we started with the *ab initio* modeling of *de novo* antiparallel coiled coils, that resulted in a stable helix-loop-helix motif. This fragment was incorporated and optimized to construct a continuous helix with the barrel using RosettaRemodel (Huang *et al.*, 2011).

Taking advantage of the symmetry of sTIM11, two copies of the motif were inserted analogously, creating a TIM barrel with even larger surface area and pocket like feature (Fig. 1). Experimental characterization of these αTIMs showed an increase in α-helical content as well as hydrodynamic radius. The data emphasizes the formation of extended helical structures and AlphaFold2 analysis furthermore supports that this purely physics-based approach led to successful designs of extended *de novo* TIM barrels.

Materials and Methods

Design protocol

All modelling and design steps were performed with the Rosetta molecular modelling suite using the scoring function ref2015 (Leaver-Fay *et al.*, 2011; Alford *et al.*, 2017). The command line options can be found in the supplementary information. An overview of the design protocol is given in Figure 2.

Antiparallel coiled coil – Rosetta ab initio

A fragment for insertion was created by structure prediction using Rosetta *Ab Initio* (Simons *et al.*, 1997; Raman *et al.*, 2009). An antiparallel *de novo* coiled-coil from Myszka and Chaiken (Myszka and Chaiken, 1994) was used as a starting point. The sequence was adapted to fit our requirements: N- and C-termini were changed from Cys to Val, the sequence was tailored to two heptad repeats for each helix and the loop length was reduced to three residues (Gly-Gly-Pro). Fragment files for *ab initio* prediction were created using the Robetta server (http://old.robetta.org/fragmentsubmit.jsp) (Raman *et al.*, 2009; Song *et al.*, 2013). The predicted models were tested for convergence by calculating the RMSD (root mean square deviation) for all decoys against the lowest scoring model. The best scored model was used for subsequent design steps.

Insertion and optimizations – RosettaRemodel

Insertion of the generated fragment into sTIM11 (PDB ID: 5BVL (Huang *et al.*, 2016)) was performed with RosettaRemodel (Huang *et al.*, 2011). Flags for RosettaRemodel can be found in the supplementary information. After each remodel calculation, the models were relaxed and scored (Nivón *et al.*, 2013). For each subsequent optimization of the inserted fragment or the transition region the secondary structure and amino acid identity in the blueprints were adjusted and restricted. Models for experimental characterization were chosen based on Rosetta score and geometrical features of the inserted fragment, in particular proper connection and a continuous α-helix extending from the motif all the way through the barrel.

Biochemical materials

All reagents were purchased from Sigma-Aldrich or Carl Roth in analytical grade unless otherwise stated. All solutions were prepared with deionized water. Oligonucleotides were ordered from Eurofins Genomics. Enzymes for cloning purposes were obtained from New England Biolabs GmbH. Cloned constructs were verified by sequencing at myGATC or Eurofins Genomics. All other constructs were ordered readily cloned in a pET21b(+) vector from BioCat.

Cloning methods

αTIM1 to 3

The initial constructs with a single inserted secondary structure element were cloned using oligonucleotides with overhangs annealing to the flanking regions at the insertion site of the TIM barrel. We used the construct sTIM11noCys as starting point for the extensions as before (Wiese *et al.*, 2021) since the two cysteines in sTIM11 did not form the originally intended disulfide bridge (Huang *et al.*, 2016). In initial polymerase chain reactions (PCR) fragments of the 5′ half of the αTIM genes were generated by using a T7 promotor primer and the reverse primer of the secondary structure element, while the 3′ half was constructed by using the forward primer for the secondary structure element with the T7 terminator primer. The PCR reactions were loaded on a 1% agarose gel and fragments with correct sizes were cut and purified using the NucleoSpin Gel and PCR Clean-Up Kit from Macherey-Nagel. In a subsequent PCR reaction the fragments were joined at their matching overhangs and further amplified using the T7 primers. PCR fragments were again purified as described. Fragments were then digested for 1 h at 37°C using NdeI and XhoI restriction enzymes. Accordingly, pET21b(+) was linearized using the same enzymes. After another purification step, insert and vector were ligated using T4 DNA ligase and an incubation at 4°C overnight. Chemically competent *E. coli* Top10 (Novagen) cells were transformed with the ligation reaction and incubated at 37°C overnight on agar plates. Colonies were picked and cultivated at 37°C, shaking at 180 rpm in LB media with ampicillin (100 μg ml^{-1}) for

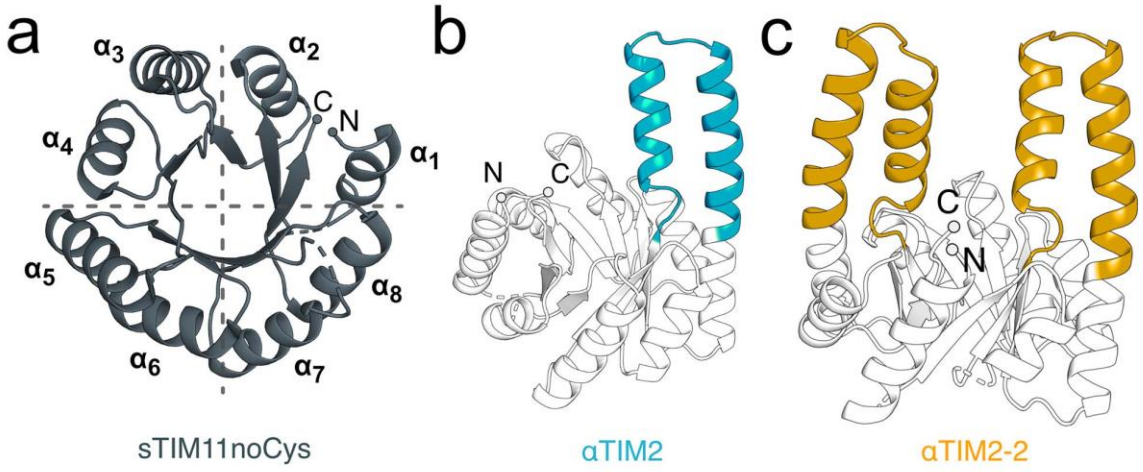


Fig. 1. Base scaffold and models of the extended TIM barrels. Structures are shown as cartoon representation with their N- and C-termini highlighted as dots. (**a**) Crystal structure of sTIM11noCys (PDB ID: 6YQY (Romero-Romero *et al.*, 2021a)). The fourfold symmetry is indicated by dashed lines and α-helices are denoted according to their order. (**b**) The Rosetta model of αTIM2 with the base scaffold displayed as outline and the introduced helix-loop-helix motif highlighted in solid. (**c**) The Rosetta model of αTIM2–2 with the base scaffold displayed as outline and the two introduced fragments as solid cartoon.

plasmid preparation using the NucleoSpin Plasmid EasyPure Kit from Macherey-Nagel. The cloned genes were verified by sequencing.

Overexpression and protein purification

E. coli BL21(DE3) cells (Novagen) were transformed with the plasmids, plated on agar plates containing 100 μg ml^{-1} ampicillin and incubated over night at 37°C. Single colonies were picked to inoculate LB precultures supplemented with ampicillin (100 μg ml^{-1}) and incubated at 37°C at 180 rpm overnight. 1 l of Terrific Broth (TB) were inoculated with 10 ml of the preculture and grown at 37°C until an OD_{600} of 0.8–1.0. Then overexpression was induced by addition of isopropyl-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. The expression cultures for sTIM11noCys and αTIM1, 3 and 2–2 were incubated for 4 h at 30°C. αTIM2 was incubated at 20°C overnight. Cells were harvested by centrifugation (Beckman Coulter Avanti J-26 XPI, JLA-8.1000, 15 min, 4000 g, 15°C) and pellets resuspended in 20 ml buffer A (50 mM NaP pH 8.0, 150 mM NaCl, 20 mM imidazole). Cells were lysed by sonication (Branson Ultrasonic Sonifier 250, output 4, duty cycle 40%, 3 × 3 min with 5 min pauses) and then centrifuged (Beckman Coulter Avanti J-26 XPI, JA-25.50, 1 h, 40 000 g, 15°C).

All purification steps were performed at room temperature. The supernatant was loaded onto a preequilibrated HisTrapHP column (5 ml, Cytiva Life Sciences) coupled to an ÄKTApure system (Cytiva Life Sciences). Target protein was eluted with a 30% step wash against buffer B (50 mM NaP pH 8.0, 150 mM NaCl, 500 mM imidazole). The peak fractions were pooled and further purified on a HiLoad 26/600 Superdex 75 preparative grade column (Cytiva Life Sciences) preequilibrated with buffer C (50 mM NaP pH 8.0, 150 mM NaCl).

Size exclusion chromatography-multi angle light scattering (SEC-MALS)

SEC-MALS measurements were performed using a Superdex 75 Increase 10/300 GL column (Cytiva Life Sciences) connected to an Agilent 1260 Infinity II HPLC system, coupled to a miniDAWN multi-angle light scattering (MALS) detector and an Optilab differential refractive index detector (dRI) (Wyatt Technology). All experiments were performed in buffer C with 0.02% NaN_3 at room temperature and a flow rate of 0.8 ml min^{-1} using a protein concentration of 1 mg ml^{-1} and an injection volume of 100 μl. Data collection and analysis were performed with the ASTRA 8.0.2.5 software (Wyatt Technology). A BSA standard sample at 2 mg ml^{-1} (Pierce) was used for MALS detector normalization, correction of peak alignment and band broadening. The signal from the dRI detector was used as concentration source for molar mass determination.

Far-UV circular dichroism

Circular Dichroism (CD) spectra were collected with a Jasco J-710 after dialyzing samples into 10 mM NaP pH 8.0 and adjusting sample concentration to 0.2 mg ml^{-1}. Far-UV spectra were recorded in the range of 190–260 nm at room temperature in a 1 mm cuvette, with a 1 nm bandwidth, 1 s response time and scanning speed of 100 nm min^{-1}. 10 accumulations were measured per sample to obtain the final spectrum. Normalization was done by subtraction of a buffer spectrum and conversion to mean residue molar ellipticity (Θ_{MRE}) using:

$$\Theta_{MRE} = \frac{\Theta \times MW}{10 \times d \times c \times n}$$

where Θ is the measured ellipticity in mdeg, MW the molecular weight in Da, d the path length in cm, c the protein concentration in mg ml^{-1} and n the number of residues in the protein.

Thermal unfolding followed by circular dichroism

Thermal unfolding curves of the samples were measured in the temperature range of 20–95°C at 222 nm with 1 nm bandwidth, 1 s response time and a heating rate of 1.0°C min^{-1}. Measured unfolding curves were analyzed with the Denatured

Protein function of SpectraAnalysis 1.53.07 (Jasco), where dependencies in the initial and final baselines were fitted and subtracted before unfolding parameters were determined from three different measurements and averaged. $\Delta G_{25°C}$ values were calculated from the obtained values for ΔH and ΔS by using the Gibbs-Helmholtz-Equation with T = 298 K.

Results

Originally, we aimed to introduce coiled coils into the TIM barrel, based on their well understood rational design principles and the high versatility of applications (Oakley and Hollenbeck, 2001). Myszka and Chaiken (Myszka and Chaiken, 1994) proposed a sequence of a two-stranded, intramolecular, antiparallel coiled coil and showed experimental evidence for this arrangement. Using this sequence, we performed Rosetta *ab initio* predictions and obtained models of the intended coiled coil topology confirmed by Socket2 (Kumar and Woolfson, 2021). As a coiled coil of 56 residues in length is too large in relation to the TIM barrel, we shortened the sequence to two heptad repeats and abridged the connecting loop. The modified sequence was again applied to Rosetta *ab initio* predictions converging to low scoring designs (Fig. 2 - *Ab initio*). Analysis of the best scoring models with Socket2, showed that the modified sequence does not correspond to a coiled-coil topology anymore, but that they form a well-defined helix–loop–helix motif, which we decided to use for the incorporation into sTIM11.

As an insertion site for the helix–loop–helix motif, we assumed that the $\beta\alpha$-loops of the second and third quarters are most suitable due to their spatial distance to the termini. We introduced the modelled fragment individually in all four $\beta\alpha$-loops in both quarters of the TIM barrel and identified the $\beta_4\alpha_5$-loop as the best insertion site based on its low Rosetta score. Throughout the loop selection procedure, we observed high conformational flexibility of the inserted motif with respect to the TIM barrel (Fig. 2 - Insertion). To avoid this plasticity and further stabilize the inserted element, we aspired to form a continuous helix with the outer α-helix of the TIM barrel. Therefore, the sequence of this transition region was shortened and restricted to a helical secondary structure. With this strategy, we were able to rigidify the inserted secondary structure element on top of the barrel as all calculated models indicated a similar structural orientation of the insert (Fig. 2 - Optimization).

Since the calculated results pointed to a proper insertion of the extension into sTIM11, we turned to revisiting the helix–loop–helix motif itself for further optimization. As the motif still had the sequence features of heptad repeats typical for coiled coils, we redesigned the inserted fragment using design principles of coiled coils to either further stabilize the extension or even establish this topology (Oakley and Hollenbeck, 2001; Hadley *et al.*, 2008). Consequently, we restricted the residues at the interface of the insert to Ile, Leu, Val, Ala or Gly building up a stable hydrophobic core. In addition, flanking residues were restricted to charged residues to shield the hydrophobic interface from solvent. In combination with these restrictions the sequence of the transition region of the continuous α-helix was not fixed to give more freedom for the optimization of the extension. Although the resulting models did not show a coiled coil topology, the optimization led to further stabilized models. At this point we selected two designs, αTIM1 and αTIM2, for experimental characterization based on the overall score and geometric arrangement. In an additional step the loop within the helix–loop–helix motif was addressed and a helix capping motif introduced (Viguera and Serrano, 1995), resulting in the design αTIM3, which we also took forward for experimental characterization.

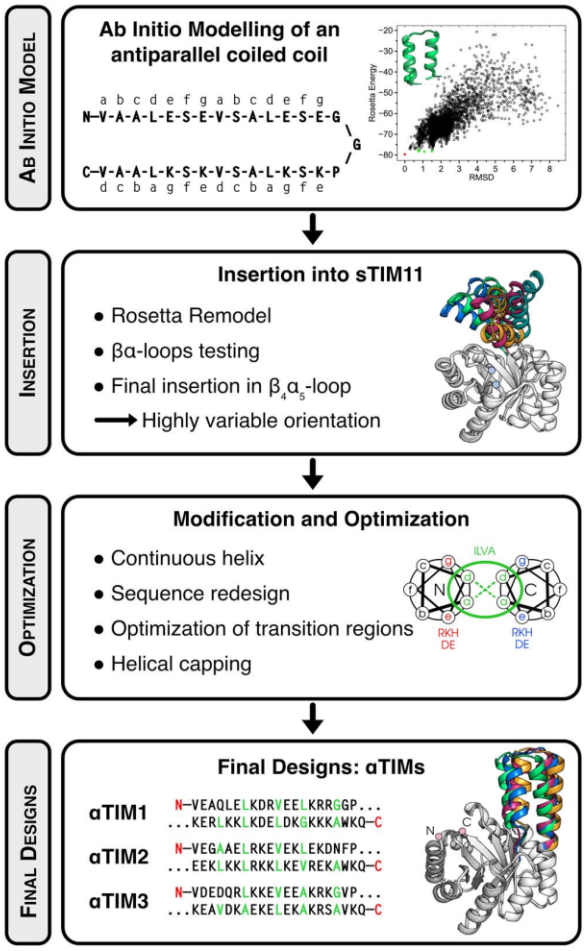


Fig. 2. Design Strategy for the extension of sTIM11. *Ab initio* Model. The structure of a modified sequence of the antiparallel coiled coil from Myszka and Chaiken (Myszka and Chaiken, 1994) was predicted using Rosetta *ab initio* modelling. The sequence is displayed and annotated with the corresponding letters for a helical wheel presentation. Structure prediction resulted in a funnel shaped landscape, with the best scoring design marked in red and the following ones in green. A cartoon representation of the best design is shown. **Insertion.** Design step where the obtained helix-loop-helix was inserted into sTIM11 using RosettaRemodel. Exemplary models with high variable orientation of the inserted fragment are shown as cartoon representation, with extension highlighted in different colors. **Optimization.** Optimization step and rigidification of the inserted fragment. The sequence was optimized by restricting to amino acids as shown in the helical wheel diagram. Residues at the interface are hydrophobic (in green), whereas adjacent residues are polar (shown in red or blue). **Final Designs.** Based on low Rosetta scores in the optimization step, three models (αTIM1, αTIM2 and αTIM3) were chosen for experimental characterization. Exemplary structures are shown on the right. Final sequences of the extensions are displayed, whereby the residues building up the hydrophobic core are highlighted in green and the directionality is indicated by the letters N and C for the termini.

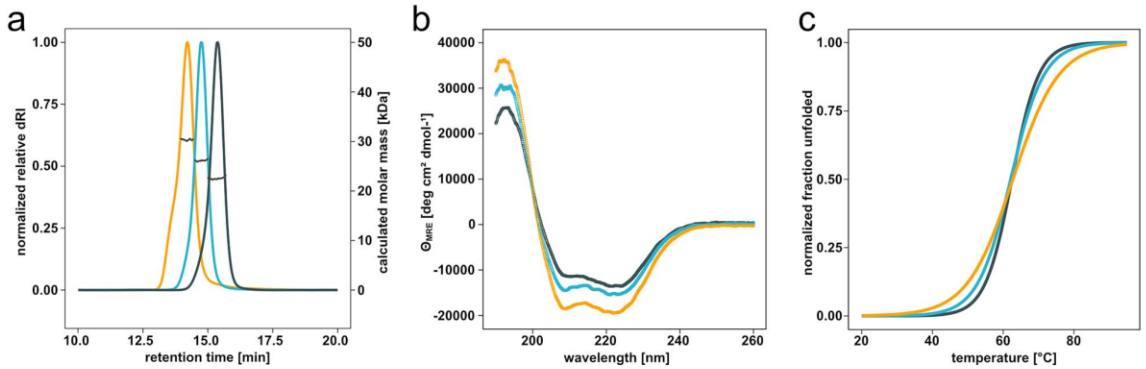


Fig. 3. The designs show the expected experimental features in comparison with the base scaffold. Experimental characterization of αTIM2 (in blue), αTIM2–2 (in yellow) and sTIM11noCys (in black). (**a**) Elution profile of the SEC-MALS measurements showing the normalized relative differential refractive index as solid line and the calculated molar mass as dots within the corresponding peak. The extensions lead to characteristic increase in apparent molecular weight. For calculated molar masses see Table I. (**b**) Far-UV CD spectra of the three constructs show a step-wise increase in α-helicity with each extension. (**c**) Thermal unfolding curves followed by CD at 222 nm highlight the stability of the constructs. For melting points and $\Delta G_{25°C}$ values see Table I.

Table I. Data points for theoretical and experimentally determined molecular weight (MW) using SEC-MALS as well as apparent melting temperature (T_M) and ΔG at 25°C using CD spectroscopy

Construct	Theoretical MW [kDa]	Experimental MW [kDa]	T_M [°C] (n = 3)	$\Delta G_{25°C}$ [kcal mol^{-1}]
sTIM11noCys	22.9	22.6 ± 0.1	62.6 ± 0.2	-6.6
αTIM2	26.4	26.2 ± 0.1	61.8 ± 0.3	-5.4
αTIM2–2	30.1	30.4 ± 0.2	62.5 ± 0.3	-3.9

Of the three constructs αTIM2 could be produced in high yields and initial characterization indicated a successful α-helical extension while αTIM1 and 3 did not yield sufficient soluble protein. Thus, we decided to continue with αTIM2 and to make use of the 4-fold symmetry of our base scaffold to introduce a second helix-loop-helix motif. Since opposite of the fragment insertion site at the $\beta_4\alpha_5$-loop the protein termini are located, we transferred the insertion positions to equivalent sites in the other quarters, namely to the $\beta_2\alpha_3$- and the $\beta_6\alpha_7$-loop, and introduced the extension from αTIM2 twice without any further optimization (Fig. 1). The obtained models showed striking resemblance of the introduced motif to our previous design pipeline and gave a good RosettaScore, which is why we chose to also experimentally characterize this design called αTIM2–2.

For the experimental work we used a cysteine-free variant of sTIM11 as base scaffold as with previous insertions (Wiese *et al.*, 2021) to avoid any undesired effects by the thiol-groups. sTIM11 had originally been designed containing two cysteines at position 9 and 182 to form a disulfide bridge as a structurally stabilizing feature, but the structure (PDB ID: 5BVL) revealed that the intended disulfide bridge is not formed. Based on this scaffold we generated the constructs αTIM1 to 3. Among these single extension designs only αTIM2 was found in the soluble fraction upon heterologous expression in *E. coli*. The αTIM2 protein was purified to homogeneity and characterized. It shows a homogenous peak in SEC-MALS analysis corresponding to monomeric protein and exhibits an increased hydrodynamic radius compared to sTIM11noCys (Fig. 3a, Table I). The experimentally determined molecular weight of αTIM2 (26.2 kDa) corresponds well to the theoretical molecular weight (26.4 kDa). Analysis of the secondary structure content by circular dichroism (CD) spectroscopy of αTIM2 revealed the spectrum of a well folded protein. In comparison to the basic barrel scaffold an increase in α-helical content is observed, which supports the formation of the introduced helix–loop–helix motif (Fig. 3b). The construct αTIM2–2 with the double extension, also expressed soluble in *E. coli*, though with reduced yields in comparison to αTIM2. The protein shows again an increased hydrodynamic radius compared to αTIM2 in SEC-MALS as expected but with a slight tendency to form oligomers (Fig. 3a). We further observed an even higher α-helical signal in CD spectroscopy indicating the presence of two helix–loop–helix motifs (Fig. 3b). To test the stability of the proteins, we followed their thermal unfolding by CD at 222 nm (Fig. 3c). Interestingly, we observed the same melting temperature (~62°C, Table I) for all constructs without a shift compared to the base scaffold. However, the cooperativity changes slightly with each insertion, so that the $\Delta G_{25°C}$ decreases significantly by up to 1.5 kcal mol^{-1} per inserted helix-loop-helix motif.

Experimental attempts to obtain structures were unfortunately not successful. Therefore, we used AlphaFold2 (Colabfold v1.5.0 with default parameters (Jumper *et al.*, 2021; Mirdita *et al.*, 2022)) to obtain structure predictions. The sequences of both αTIM2 and αTIM2–2 were predicted to form TIM barrels with α-helical extensions (Fig. 4). Both predictions show not only a high pLDDT value for the entire protein, but also high confidence within the inserted helix-loop-helix motifs (αTIM2: 89.36 to 98.50, αTIM2–2: 85.75 to 98.47). In addition, the AlphaFold2 predictions match the final Rosetta model quite accurately with RMSD values over all Cα atoms below 1.7 Å and 3.0 Å for αTIM2 and αTIM2–2, respectively. Smaller differences between the Rosetta models and the AlphaFold2 predictions are due to different angles of the extensions. The AlphaFold2 predictions show a straighter helix on the outside of the barrel, whereas

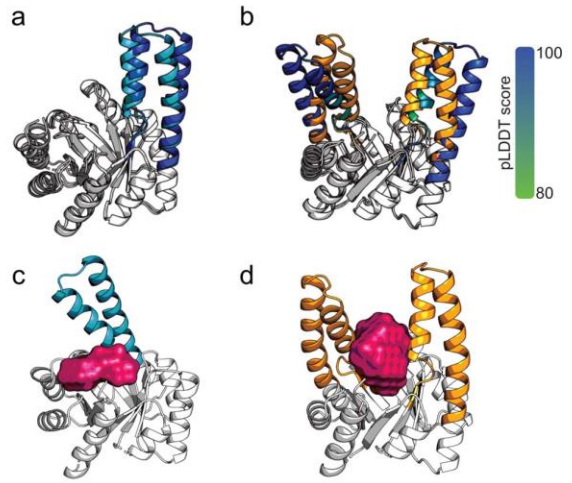


Fig. 4. The Rosetta models compare well to the AlphaFold2 predictions and show predicted pockets. Rosetta models are shown in the same representation and color code as in Figure 1 (extensions of αTIM1 in light blue and of αTIM2 in yellow). AlphaFold2 predictions are shown as cartoon representation and superimposed on the Rosetta models over all Cα atoms. Extensions in the AlphaFold2 predictions are colored corresponding to their pLDDT score according to the color bar ranging from 80 in green to 100 in blue. Predicted pockets by PUResNET are shown as surface representation and colored in red. (**a**) Superposition for αTIM2 (RMSD: 1.67 Å). (**b**) Superposition for αTIM2–2 (RMSD: 2.97 Å). (**c**) Pocket prediction for αTIM2. (**d**) Pocket prediction for αTIM2–2.

the Rosetta models have a higher curvature within this helix. To identify pockets within our designs we applied the AI-based ligand-binding site prediction tool PUResNET (Kandel *et al.*, 2021). Predicted pockets for both constructs are located between the C-terminal end of the inner β-sheet and the helix-loop-helix motifs. In αTIM2–2 the two pockets from both extensions connect and form one pocket with an increased volume (Fig. 4c and d).

Discussion

The TIM-barrel is a ubiquitous protein fold that can harbor diverse functions present in six out of seven EC classes and by this poses an interesting target for protein design as a promising starting point towards tailor-made enzymes. A major progress in the field of *de novo* design was the development of sTIM11 and its descendants providing access to minimalistic TIM barrels without evolutionary bias and with different thermodynamic properties (Huang *et al.*, 2016; Romero-Romero *et al.*, 2021a). The drawback of these designs are their idealized topology and minimalistic loops that lack extended surface area or cavities crucial for the introduction of enzyme function. In this work we diversified the set of *de novo* TIM barrels by inserting a helix–loop–helix motif into sTIM11noCys. To search for structurally similar motifs in nature we performed a DALI database search with αTIM2 (Holm, 2022). Hereby, we found mainly other designed TIM barrels as well as natural TIM barrels of the aldolase family, which have a round shaped barrel core. Of these enzymes the class II fructose-bisphosphate aldolase (*e.g.* PDB ID: 5u7s) shows a somewhat comparable motif. In its eighth $\beta\alpha$-loop there is an $\alpha\alpha$-hairpin, where the second α-helix extends continuously through helix 8 of the barrel and shields the inner β-sheet similar to our motif. However, the aldolase extension differs in length of the hairpin helices, in the length of the connecting loop and in the angle of the motif on top of the barrel. In addition, it is supported by a second motif in the neighboring loop and seems to be involved in dimer formation, which is neither intended nor indicated for αTIM2.

For the introduction of the helix–loop–helix motif in sTIM11noCys we divided our design approach into three individual steps. First, a proposed antiparallel coiled coil (Myszka and Chaiken, 1994) was truncated to match its length to the size of a TIM barrel. Additionally, the connecting loop was shortened to reduce the flexibility between both helices. The resulting sequence was applied to a Rosetta *ab initio* calculation, where a proper folding funnel was obtained and the best scoring model was picked as an extension to be incorporated into sTIM11. Prior to this, we analyzed the model with the Socket2 framework (Kumar and Woolfson, 2021), which revealed that in fact our model does not adopt a proper coiled coil topology anymore, which is likely due to the shortening of the element to two heptad repeats. Nevertheless, we were confident with the modelled secondary structure to move on and in the second design step to test different $\beta\alpha$-loops of sTIM11 to find the most promising insertion site. We identified the $\beta_4\alpha_5$-loop as the lowest scoring insertion point according to the Rosetta calculations. The final design step with RosettaRemodel included a sequence optimization aiming to build a continuous outer helix and stabilization of the fragment. The continuous outer helix is thought to rigidify the insertion, because we observed flexible arrangements of the helix–loop–helix motif on top of the TIM barrel upon modelling. Further, we aimed to improve stability of the helix–loop–helix fragment by applying design strategies found in successful coiled coil designs (Oakley and Hollenbeck, 2001; Hadley *et al.*, 2008). Hereby, we used RosettaRemodel to introduce hydrophobic residues at the interface of the helix–loop–helix fragment and polar residues at the adjacent positions to protect the hydrophobic core from the solvent and by this improving its packing properties. Out of these calculations two designs were chosen based on lowest Rosetta score and geometrical features in the sense of a continuous outer helix. These two designs were called αTIM1 and αTIM2. Additionally, we took a third design (αTIM3) forward with a helix capping motif (Viguera and Serrano, 1995) to check for further stabilizing effects.

An advantage of this approach is the full control of each individual design step. The physics-based strategy allows for rational assessment and intervention during the entire process in contrast to recent AI methods, such as RFdiffusion (Watson *et al.*, 2023). These new tools allow fast and easy generation of *de novo* proteins but with the downside that the design workflow is not as accessible and finely tunable as the presented design pipeline.

sTIM11 was already furnished with a smaller secondary structure element (Wiese *et al.*, 2021), where a PSIPRED predicted helix fragment was introduced into the same $\beta_4\alpha_5$-loop as used in αTIM2. In comparison to the rather small insertion, our αTIMs provide a larger surface area and by this possess the potential to exhibit cavities useful for downstream functionalization. In contrast to the functionalized TFD designs by Caldwell *et al.* (Caldwell *et al.*, 2020) where the TIM barrel was split and the halves fused to a *de novo* designed dimeric ferredoxin fold, our designs retain the

TIM barrel architecture in a monomeric fashion. This should allow to explore a wider range of functions, as we stay closer to the natural topology with its ability to catalyze a wide range of enzymatic reactions (Nagano *et al.*, 2002).

Experimental analysis of the three αTIM variants showed that only αTIM2 could be purified for further measurements. The protein turns out to be a monomer whose α-helical content and hydrodynamic radius is increased compared to sTIM11noCys (Fig. 3), indicating a successful formation of the helix–loop–helix motif. Based on these results, we imagined to increase the surface area by duplication of our αTIM2 extensions and transferring these to different quarters without any modifications thereby taking advantage of the modularity of the scaffold. The resulting construct αTIM2–2 displayed the expected behavior in the experimental analysis (Fig. 3): the hydrodynamic radius as well as the α-helical content is further increased compared to αTIM2. This result highlights the capability of sTIM11noCys to harbor multiple extensions on its 'catalytic face'. The increase in size and α-helical content appear to be additive with each insertion, while thermal denaturation experiments (Fig. 3c) showed that the variants are somewhat destabilized with increasing number of extensions (Table I).

As we were not successful in obtaining an experimental structure, we used AlphaFold2 to predict the structures. These calculations gave predictions with high confidence similar to the models from our computational pipeline. The superimposition shows only differences in the orientation of the extensions (Fig. 4). In fact, there might be some conformational flexibility of the extensions towards the barrel. The combination of our experimental data revealing increased α-helical content and hydrodynamic radius, together with the AlphaFold2 predictions showing TIM barrels with the intended extensions, strongly indicate that the workflow to introduce a larger helix–loop–helix motif into a *de novo* TIM barrel was successful. In addition, the introduction of these structural elements creates pockets on top of the barrel indicated by the PUResNET predictions. Especially the increased pocket volume of αTIM2–2 should be feasible for downstream functionalization like the introduction of a ligand or substrate binding site.

On the path to create tailor-made enzymes the main challenge is the incorporation of a specific active site geometry (Privalov and Makhatadze, 1992; Geitner and Schmid, 2012). The diverse set of extensions created previously and in this work lays a solid foundation for future approaches to tackle this problem.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgments

We thank Sabrina Wischt and Leonie Lutz for technical support, Sergio Romero-Romero and Stefan Klingl for comments on the manuscript, and all members of the Höcker Lab for discussions.

Author contributions

Sina Kordes (Conceptualization-Equal, Data curation-Equal, Investigation-Equal, Methodology-Equal, Visualization-Equal, Writing – original draft-Equal), Julian Beck (Data curation-Equal, Investigation-Equal, Methodology-Equal, Visualization-Equal, Writing – review & editing-Equal), Sooruban Shanmugaratnam (Data curation-Equal, Investigation-Equal, Visualization-Equal, Writing – review & editing-Equal), Merle Flecks (Investigation-Equal) and Birte Höcker (Conceptualization-Equal, Funding acquisition-Equal, Methodology-Equal, Resources-Equal, Writing – original draft-Equal, Writing – review & editing-Equal)

Conflict of interest

None.

Funding

This work was supported through core funding of the University of Bayreuth.

Data availability

All data is made available and the statements have been made according to the PEDS guidelines.

References

Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R. *et al.* (2017) *J. Chem. Theory Comput.*, **13**, 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

Anand, N., Eguchi, R., Mathews, I.I. *et al.* (2022) *Nat. Commun.*, **13**, 746. https://doi.org/10.1038/s41467-022-28313-9.

Banner, D.W., Bloomer, A.C., Petsko, G.A. *et al.* (1975) *Nature*, **255**, 609–614. https://doi.org/10.1038/255609a0.

Caldwell, S.J., Haydon, I.C., Piperidou, N. *et al.* (2020) *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 30362–30369. https://doi.org/10.1073/pnas.2008535117.

Chu, A.E., Fernandez, D., Liu, J. *et al.* (2022) *Biodes. Res.*, **2022**, 1–13. https://doi.org/10.34133/2022/9842315.

Dauparas, J., Anishchenko, I., Bennett, N. *et al.* (2022) *Science*, **378**, 49–56. https://doi.org/10.1126/science.add2187.

Dou, J., Vorobieva, A.A., Sheffler, W. *et al.* (2018) *Nature*, **561**, 485–491. https://doi.org/10.1038/s41586-018-0509-0.

Doyle, L., Hallinan, J., Bolduc, J. *et al.* (2015) *Nature*, **528**, 585–588. https://doi.org/10.1038/nature16191.

Geitner, A.J. and Schmid, F.X. (2012) *J. Mol. Biol.*, **420**, 335–349. https://doi.org/10.1016/j.jmb.2012.04.018.

Goverde, C.A., Pacesa, M., Dornfeld, L.J. *et al.* (2023) *bioRxiv*, 2023.05.09.540044.

Hadley, E.B., Testa, O.D., Woolfson, D.N. *et al.* (2008) *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 530–535. https://doi.org/10.1073/pnas.0709068105.

Holm, L. (2022) *Nucleic Acids Res.*, **50**, W210–W215. https://doi.org/10.1093/nar/gkac387.

Huang, P.S., Ban, Y.E.A., Richter, F. *et al.* (2011) *PloS One*, **6**, 8. https://doi.org/10.1371/journal.pone.0024109.

Huang, P.S., Feldmeier, K., Parmeggiani, F. *et al.* (2016) *Nat. Chem. Biol.*, **12**, 29–34. https://doi.org/10.1038/nchembio.1966.

Jumper, J., Evans, R., Pritzel, A. *et al.* (2021) *Nature*, **596**, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kandel, J., Tayara, H. and Chong, K.T. (2021) *J. Chem.*, **13**, 1–14. https://doi.org/10.1186/s13321-021-00547-7.

Kordes, S., Romero-Romero, S., Lutz, L. *et al.* (2022) *Protein Sci.*, **31**, 513–527. https://doi.org/10.1002/pro.4249.

Kuhlman, B., Dantas, G., Ireton, G.C. *et al.* (2003) *Science*, **302**, 1364–1368. https://doi.org/10.1126/science.1089427.

Kumar, P. and Woolfson, D.N. (2021) *Bioinformatics*, **37**, 4575–4577. https://doi.org/10.1093/bioinformatics/btab631.

Leaver-Fay, A., Tyka, M., Lewis, S.M. *et al.* (2011) *Methods Enzymol.*, **487**, 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

Maes, D., Zeelen, J.P., Thanki, N. *et al.* (1999) *Proteins*, **37**, 441–453. https://doi.org/10.1002/(SICI)1097-0134(19991115)37:3<441::AID-PROT11>3.0.CO;2-7.
Marcos, E., Chidyausiku, T.M., McShan, A.C. *et al.* (2018) *Nat. Struct. Mol. Biol.*, **25**, 1028–1034. https://doi.org/10.1038/s41594-018-0141-6.
Minami, S., Kobayashi, N., Sugiki, T. *et al.* (2023) *Nat. Struct. Mol. Biol.*, **30**, 1–9.
Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* (2022) *Nat. Methods*, **19**, 679–682. https://doi.org/10.1038/s41592-022-01488-1.
Myszka, D.G. and Chaiken, I.M. (1994) *Biochemistry*, **33**, 2363–2372. https://doi.org/10.1021/bi00175a003.
Nagano, N., Orengo, C.A. and Thornton, J.M. (2002) *J. Mol. Biol.*, **321**, 741–765. https://doi.org/10.1016/S0022-2836(02)00649-6.
Nivón, L.G., Moretti, R. and Baker, D. (2013) *PloS One*, **8**, e59004. https://doi.org/10.1371/journal.pone.0059004.
Oakley, M.G. and Hollenbeck, J.J. (2001) *Curr. Opin. Struct. Biol.*, **11**, 450–457. https://doi.org/10.1016/S0959-440X(00)00232-3.
Pan, X., Thompson, M.C., Zhang, Y. *et al.* (2020) *Science*, **369**, 1132–1136. https://doi.org/10.1126/science.abc0881.
Privalov, P.L. and Makhatadze, G.I. (1992) *J. Mol. Biol.*, **224**, 715–723. https://doi.org/10.1016/0022-2836(92)90555-X.
Raman, S., Vernon, R., Thompson, J. *et al.* (2009) *Proteins*, **77**, 89–99. https://doi.org/10.1002/prot.22540.
Regan, L. and Degrado, W.F. (1988) *Science*, **241**, 976–978. https://doi.org/10.1126/science.3043666.
Romero-Romero, S., Costas, M., Silva Manzano, D.A. *et al.* (2021a) *J. Mol. Biol.*, **433**, 167153. https://doi.org/10.1016/j.jmb.2021.167153.
Romero-Romero, S., Kordes, S., Michel, F. *et al.* (2021b) *Curr. Opin. Struct. Biol.*, **68**, 94–104. https://doi.org/10.1016/j.sbi.2020.12.007.
Simons, K.T., Kooperberg, C., Huang, E. *et al.* (1997) *J. Mol. Biol.*, **268**, 209–225. https://doi.org/10.1006/jmbi.1997.0959.
Song, Y., Dimaio, F., Wang, R.Y.R. *et al.* (2013) *Structure*, **21**, 1735–1742. https://doi.org/10.1016/j.str.2013.08.005.
Sterner, R. and Höcker, B. (2005) *Chem. Rev.*, **105**, 4038–4055. https://doi.org/10.1021/cr030191z.
Thoma, R., Hennig, M., Sterner, R. *et al.* (2000) *Structure*, **8**, 265–276. https://doi.org/10.1016/S0969-2126(00)00106-4.
Urfer, R. and Kirschner, K. (1992) *Protein Sci.*, **1**, 31–45. https://doi.org/10.1002/pro.5560010105.
Viguera, A.R. and Serrano, L. (1995) *J. Mol. Biol.*, **251**, 150–160. https://doi.org/10.1006/jmbi.1995.0422.
Vijayabaskar, M.S. and Vishveshwara, S. (2012) *PLoS Comput. Biol.*, **8**, e1002505. https://doi.org/10.1371/journal.pcbi.1002505.
Watson, J.L., Juergens, D., Bennett, N.R. *et al.* (2023) *Nature*, **620**, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.
Wierenga, R.K. (2001) *FEBS Lett.*, **492**, 193–198. https://doi.org/10.1016/S0014-5793(01)02236-0.
Wiese, J.G., Shanmugaratnam, S. and Höcker, B. (2021) *Protein Sci.*, **30**, 982–989. https://doi.org/10.1002/pro.4064.
Yang, C., Sesterhenn, F., Bonet, J. *et al.* (2021) *Nat. Chem. Biol.*, **17**, 492–500. https://doi.org/10.1038/s41589-020-00699-x.

Supplementary Material

Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs

Sina Kordes[1,2], Julian Beck[1], Sooruban Shanmugaratnam[1], Merle Flecks[1] and Birte Höcker[1]

[1]Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany
[2]Current address: Proteros biostructures GmbH, 82515 Martinsried, Germany

Corresponding author: Birte Höcker
Email address: birte.hoecker@uni-bayreuth.de

This file includes:

- Supplementary Methods
- Supplementary Figure S1

Supplementary Methods

Specific options for Rosetta modelling are given in the following. For all calculations Rosetta version 2017.34 was used with the scoring function ref2015.

1. Options used for ab initio modeling

```
-database Rosetta/main/database
-abinitio:relax
-in:file:fasta coiledcoil.fasta
-in:file:frag3 3fragm
-in:file:frag9 9fragm
-nstruct 5000
```

2. Options used for initial fragment insertion

```
-database Rosetta/main/database
-remodel:blueprint blueprint_initial_insertion.txt
-remodel:domainFusion:insert_segment_from_pdb Helix-Loop-Helix.pdb
-remodel:num_trajectory 100
-remodel::quick_and_dirty
-find_neighbors
```

3. Options used for optimization

```
-database Rosetta/main/database
-remodel:blueprint blueprint_ContinousHelix.txt
-remodel:num_trajectory 100
-find_neighbors
-no_optH false
-ex1
-ex2
```

Supplementary Figure

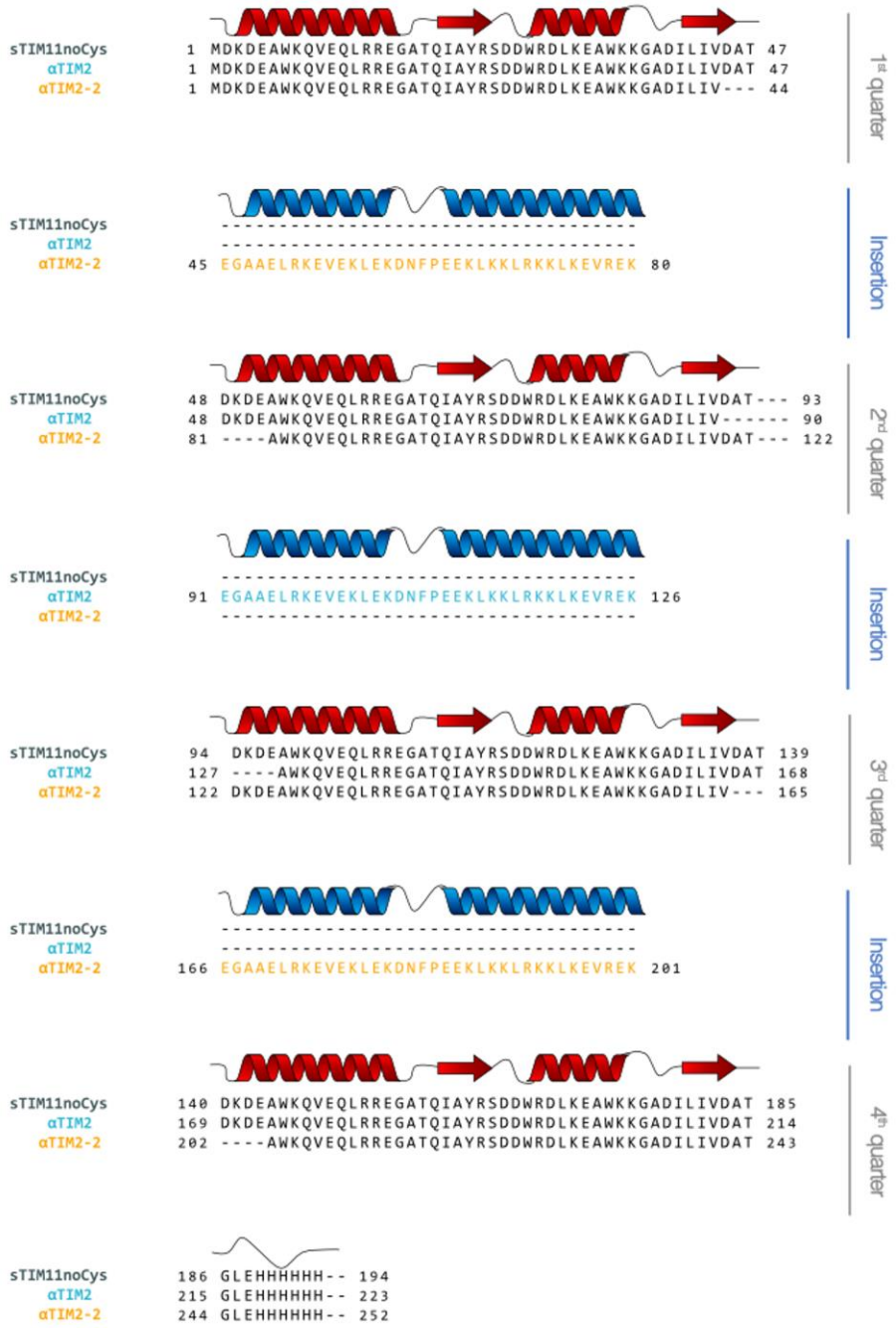


Fig S1. Sequence alignment of base scaffold with αTIMs. Full sequence alignment of the three experimentally characterized constructs. Sequences for the extensions are highlighted according to the color code used in Figure 1. Sequences are divided into the different quarters and the insertion. Secondary structure depictions are red for the base scaffold and blue for the insertions.

Publication 2

Diversifying *de novo* TIM barrels by hallucination

Julian Beck, Sooruban Shanmugaratnam, Birte Höcker
Protein Science, doi: 10.1002/pro.5001

Received: 22 December 2023 | Revised: 26 March 2024 | Accepted: 10 April 2024

DOI: 10.1002/pro.5001

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Diversifying *de novo* TIM barrels by hallucination

Julian Beck | **Sooruban Shanmugaratnam** | **Birte Höcker**

Department of Biochemistry, University of Bayreuth, Bayreuth, Germany

Correspondence
Birte Höcker, Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.
Email: birte.hoecker@uni-bayreuth.de

Funding information
University of Bayreuth

Abstract

De novo protein design expands the protein universe by creating new sequences to accomplish tailor-made enzymes in the future. A promising topology to implement diverse enzyme functions is the ubiquitous TIM-barrel fold. Since the initial *de novo* design of an idealized four-fold symmetric TIM barrel, the family of *de novo* TIM barrels is expanding rapidly. Despite this and in contrast to natural TIM barrels, these novel proteins lack cavities and structural elements essential for the incorporation of binding sites or enzymatic functions. In this work, we diversified a *de novo* TIM barrel by extending multiple βα-loops using constrained hallucination. Experimentally tested designs were found to be soluble upon expression in *Escherichia coli* and well-behaved. Biochemical characterization and crystal structures revealed successful extensions with defined α-helical structures. These diversified *de novo* TIM barrels provide a framework to explore a broad spectrum of functions based on the potential of natural TIM barrels.

KEYWORDS

de novo protein design, hallucination, machine learning, small molecule binding site, $(\beta\alpha)_8$-barrel

1 | INTRODUCTION

Protein space is not limited to the sequences sampled by natural evolution but can be expanded through *de novo* protein design by creating new sequences (Huang, Boyken, & Baker, 2016). Basic principles to design idealized proteins from scratch have been defined, and a wide variety of *de novo* proteins with different topologies have already been generated (Dou et al., 2018; Doyle et al., 2015; Huang, Feldmeier, et al., 2016; Kim et al., 2023; Koga et al., 2012; Marcos et al., 2018; Minami et al., 2023; Pan & Kortemme, 2021; Yang et al., 2021). One important fold is the $(\beta\alpha)_8$- or triose-phosphate isomerase (TIM) barrel, which is ubiquitous in nature and prominent in enzymes (Romero-Romero, Kordes, et al., 2021; Sterner & Höcker, 2005). It is present in all classes of the Enzyme Commission except the translocase class. The structure is composed of eight alternating βα-subunits, forming a central eight-stranded, parallel β-barrel encompassed by eight α-helices (Wierenga, 2001). One of the key characteristics of this fold is the spatial separation of stability and catalytic function. Protein stability is achieved through the hydrophobic core of the barrel and the αβ-loops situated at the N-terminal ends of the β-strands (Vijayabaskar & Vishveshwara, 2012). In contrast, the catalytically active residues are found at the C-terminal ends of the β-strands (Nagano et al., 2002). Typically, substrate binding occurs

Reviewing Editor: Aitziber L. Cortajarena

via a cavity formed at the central surface of the β-sheet, which is supported by elongated βα-loops on the top of the barrel (Thoma et al., 2000).

Since one prominent objective of *de novo* protein design is to create tailor-made enzymes, the TIM-barrel fold is an outstanding target. After decades of attempts to understand the principles of the TIM-barrel fold, Huang, Feldmeier, et al. (2016) succeeded in building the first *de novo* TIM barrel from scratch, named sTIM11, thereby providing a TIM-barrel scaffold that is free from any evolutionary biases paving the way for further investigations into the capabilities of this fold. In a highly rational fashion, the design problem was simplified by the introduction of a four-fold symmetry and a restriction of the design approach based on geometrical constraints derived from the inner β-sheet. Since then, the idealized sTIM11 with its minimal loops was subject to multiple modifications to increase folding, stability, and crystallizability, resulting in a *de novo* TIM-barrel family with over 20 members (Kordes et al., 2022; Romero-Romero, Costas, et al., 2021). Recently, the family of *de novo* TIM barrels was further expanded by a two-fold symmetric design, leading to a distinctive curvature of the central β-barrel and an overall ovoid shape of the barrel (Chu et al., 2022). Amidst the ongoing machine learning revolution and the emergence of AlphaFold2, numerous novel tools have been integrated into the realm of *de novo* protein design, diverging from traditional rational- and physics-based approaches (Jumper et al., 2021). Nevertheless, the TIM-barrel fold remains a promising design target, as new methodologies have already been utilized to expand the *de novo* TIM-barrel family. Notably, Anand et al. (2022) harnessed a potential learned neural network, while Goverde et al. leveraged AlphaFold2 and proteinMPNN to successfully redesign sTIM11 (Dauparas et al., 2022; Goverde et al., 2023; Jumper et al., 2021). These efforts led to a significant expansion of the sequence space of *de novo* TIM barrels and a deviation from the so-far established sequence symmetry.

In addition to redesign approaches, neural networks have shown their ability to generate entirely novel proteins from scratch. An approach called hallucination utilizes the structure prediction software RoseTTAFold for the optimization of random sequences that result in the generation of diverse proteins with a wide range of sequences and predicted structures (Anishchenko et al., 2021; Baek et al., 2021). Expanding on this, two additional approaches called constrained hallucination and inpainting utilize initial information such as functional sites to construct diverse protein frameworks without the need to predefine a fold or secondary structure (Wang et al., 2022). By fine-tuning RoseTTAFold for denoising tasks, a new approach known as RFdiffusion was developed (Watson et al., 2023). This method can tackle multiple protein design tasks, including unconditional and topology-constrained protein monomer design. To showcase the potential of RFdiffusion in generating targeted folds, the authors designed several TIM barrels. However, RFdiffusion only generates backbones, and its sequence design relies on proteinMPNN (Dauparas et al., 2022).

Despite the growing number of *de novo* TIM-barrel structures with these new artificial intelligence (AI) tools, all generated *de novo* TIM barrels still lack the feature of cavities, pockets, or extended loops compared to natural TIM barrels, which exhibit a wide variety of structural elements in their βα-loops. Thus, to create functionalized *de novo* TIM barrels, incorporating structural extensions or hydrophobic pockets becomes essential. Numerous attempts have been made to diversify the idealized structure of sTIM11. The already-mentioned ovoid-shaped barrel was designed with non-structured loops capable of adopting diverse conformations (Chu et al., 2022). In a separate study, Wiese et al. (2021) introduced a small helix into the βα-loops of the barrel. Building on this concept, Kordes et al. (2023) implemented a larger helix–loop–helix motif. In another work, Caldwell et al. (2020) split the TIM barrel and fused a designed ferredoxin fold, creating a homodimer with a cavity which was functionalized downstream with a metal binding site. All these endeavors demonstrate the versatility of *de novo* TIM barrels in accommodating different structural motifs while emphasizing the importance of diversifying their idealized structure.

In this work, we aimed to expand the *de novo* TIM-barrel family by introducing secondary structural elements to enhance its surface area and create a cavity. Taking advantage of state-of-the-art machine learning methods, we hallucinated extensions and optimized the sequences with proteinMPNN, whereby generating *de novo* TIM barrels with two or three helical extensions. These designs were analyzed through biophysical and structural characterization.

2 | RESULTS

2.1 | Constrained hallucination incorporates helical hairpins into sTIM11-SB

For the diversification experiment, we used the *de novo* TIM barrel sTIM11-SB as the base scaffold. This variant contains a stabilizing salt bridge cluster in the lower part of the β-barrel (Kordes et al., 2022). As a method, we applied the constrained hallucination approach from

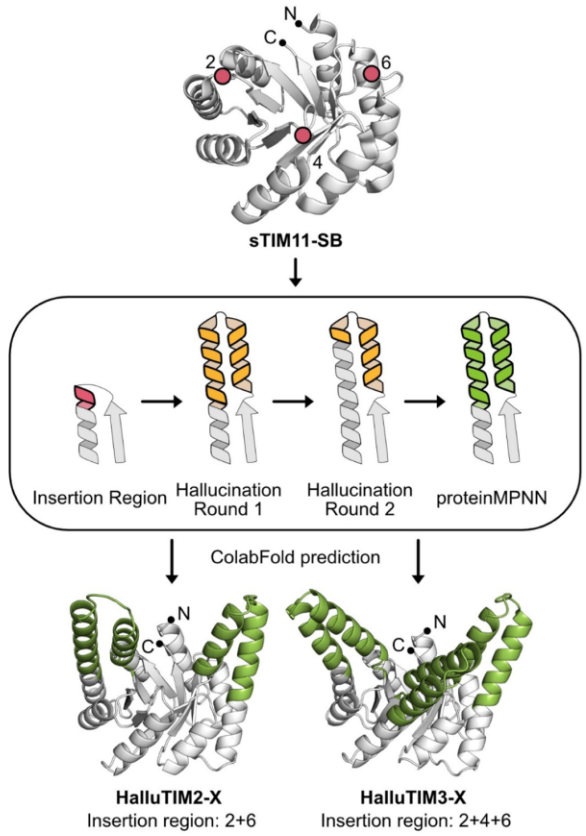


FIGURE 1 Design workflow for the extensions in sTIM11-SB. For the hallucination of extensions sTIM11-SB (PDB-ID: 7OSU), displayed in white and as cartoon representation with black dots highlighting the termini, was used as a base. Three insertion sites were defined within the second, fourth, and sixth loop, marked with a red dot. For the constrained hallucination as shown in the central panel, the first turn of the outer α-helix (in red with thick outline) and the loop to the inner β-strand were used as the insertion region. During round one of constrained hallucination α-helical extensions were obtained on top of the barrel (in yellow with thick outline). Within round two, the newly obtained outer α-helix was kept fixed except for the last turn and only the smaller α-helix was hallucinated again, highlighted with the thicker black outline. After this a sequence optimization of the entire hallucinated fragment was performed (in green with thick outline) and the structure of the designs were predicted with ColabFold. For constrained hallucination, either insertion region one and three or all were used, resulting in HalluTIM2-X with two α-helical extensions or HalluTIM3-X with three extensions (in green).

Wang et al. (2022) and chose as insertion regions the three elongated βα-loops on the C-terminal side of the β-strands (Figure 1). We decided to hallucinate either two extensions in the second and fourth quarter of the barrel or combine these with an additional one in the third quarter opposite to the termini to increase the chances of building up a cavity. The hallucinated fragments within these models turned out to be an elongation of the outer α-helix by multiple turns as well as the generation of a small α-helix above the inner β-strand resulting, in a helix–loop–helix motif (Figure 1). Notably, this topology of the hallucination is present not only in the best but also in most of the designs. To estimate the backbone diversity of the designs, we calculated the TM-score of each design against all others within the initial round of hallucination (Zhang & Skolnick, 2005). The lowest TM-scores within each dataset were found to be approximately 0.81, indicating a low backbone diversity. The highest deviation in the generated structures is found within the region above the inner β-strands. Here, not always a continuous α-helix for the full helix–loop–helix motif is formed but sometimes only a loopy connection to the outer elongated helix. Interestingly, all these designs showed lower pLDDT- scores than the ones with a fully formed helix–loop–helix motif and were thus discarded during the filter process. To further increase the quality of our designs, we performed a second round of constrained hallucination with the top scoring designs. Hereby, the elongated outer α-helix was kept fixed except for the last turn, and the design was focused on the smaller α-helix packed against the elongated one (Figure 1). With this strategy, we were able to improve the average pLDDT of all modeled designs, but the top scoring designs showed only a slight improvement as the original input already had a tight packing of the α-helices against each other. Since the second round of constrained hallucination did not significantly improve the best designs, we did not perform a third round but instead optimized the sequences of the extensions using proteinMPNN (Dauparas et al., 2022) (Figure 1). After the prediction of all generated sequences with ColabFold (Mirdita et al., 2022), six designs were selected for experimental characterization based on the average pLDDT score and the packing of the hallucinated α-helices. We chose three designs for each insertion site combination (Tables S1 and S2). The constructs were named HalluTIMX-X, whereby the first X corresponds to the number of extensions and the second X differentiates the constructs within the same category (Figure 1).

2.2 | Experimentally tested HalluTIM variants show increased helicity and thermostability

After heterologous expression in *Escherichia coli*, all designs were found in the soluble fraction of the cell extract and could be purified to homogeneity. All designs except HalluTIM2-3 showed a homogenous peak

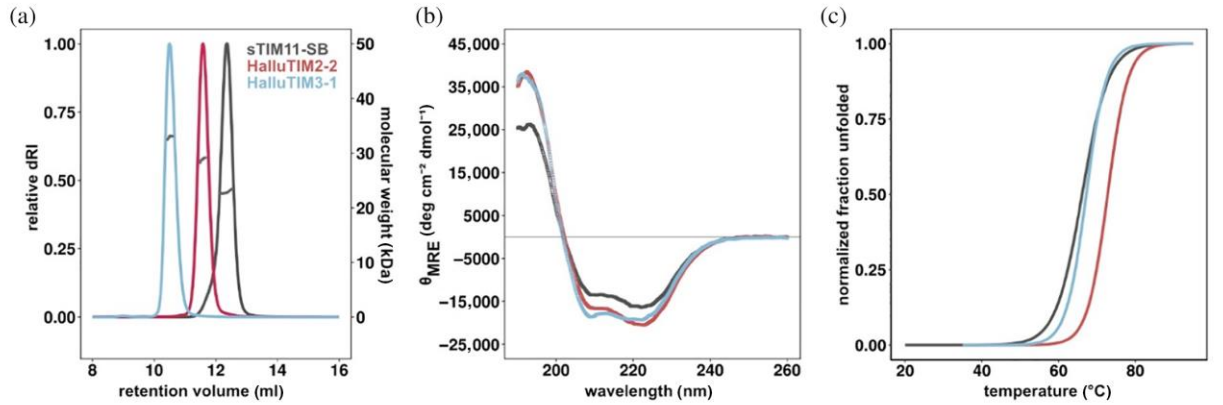


FIGURE 2 Biochemical analysis of HalluTIMs in comparison with the base scaffold. Experimental characterization of HalluTIM2-2 (in red), HalluTIM3-1 (in blue) and sTIM11-SB (in gray). (a) Elution profile of size exclusion chromatography-multi angle light scattering measurements showing the normalized relative differential refractive index as solid line and the calculated molar mass as data points in dark gray within the corresponding peak. With each extension, the hydrodynamic radius and molecular weight increases. For experimentally determined masses see Table S3. (b) Circular dichroism spectra show increases in α-helicity for both HalluTIMs compared to the base scaffold. (c) Thermal unfolding followed by circular dichroism shows an increase in stability of the designs compared to sTIM11-SB. For melting points and $\Delta G_{25^\circ C}$ values, see Table S3. dRI, differential refractive index.

corresponding to monomeric proteins and an increased hydrodynamic radius in comparison to the base construct sTIM11-SB in size exclusion chromatography-multi angle light scattering (SEC-MALS) analysis (Figures 2a and S1). HalluTIM2-3 displayed two species with slightly different hydrodynamic radii. Each experimentally determined molecular weight corresponds well to the theoretical monomeric molecular weight (Table S3). Analysis of the secondary structure content by circular dichroism (CD) spectroscopy revealed the spectra of well folded proteins. In comparison to the basic scaffold sTIM11-SB all HalluTIMs, except HalluTIM2-3 and HalluTIM3-2, showed an increase in α-helicity (Figures 2b and S1), which indicates proper formation of the hallucinated extensions. However, no major differences in the increase in α-helicity between the constructs are observable, despite the introduction of a different number of helical extensions. To investigate if the hallucinated extensions influence protein stability, we followed the thermal unfolding by CD for all proteins (Figures 2c and S1). Interestingly, we observed a similar or even a higher melting temperature for all HalluTIMs, except HalluTIM2-3, in comparison to the base scaffold (Table S3). By calculating unfolding parameters for each protein, we obtained similar or higher $\Delta G_{25^\circ C}$ for all HalluTIMs except HalluTIM2-3 in comparison to sTIM11-SB (Table S3), indicating that the extensions stabilize the entire TIM-barrel protein. Interestingly, these changes in $\Delta G_{25^\circ C}$ are caused mainly by a change in cooperativity. In addition, we checked on the reversibility of unfolding by collecting CD spectra after the melting process (Figure S2) observing that all HalluTIMs maintained the reversibility of the base scaffold.

2.3 | Crystal structures of two HalluTIMs validate the formation of novel extensions

To gain more insights and validate the successful incorporation of the hallucinated extensions, we crystallized HalluTIM2-2 (Protein Data Bank-Identifier (PDB-ID): 8R8N) and HalluTIM3-1 (PDB-ID: 8R8O). The cartoon representations are shown in Figure 3 and the crystallographic details are listed in Table S6. Within the crystal structure of HalluTIM2-2, the α-helical extension at position 1 is resolved entirely; it forms multiple crystal contacts with itself (Figure 3b). The second extension at position 3 is not involved in any crystal contacts, and one helical turn before and after the loop could not be resolved (Figure 3c). In the case of HalluTIM3-1, the crystal structure shows all three intended hallucinated extensions in their entirety, verifying their successful incorporation into sTIM11-SB. One minor deviation between HalluTIM3-1 and the base scaffold is observed within the N-terminal α-helix of the barrel, as these residues do not form a continuous α-helix. Notably, for both crystal structures, a significant number of crystal contacts are formed within the resolved α-helices that had been optimized with proteinMPNN. In the case of HalluTIM3-1, the crystal has an uncommonly high solvent content of 78% (Matthews coefficient: 5.6)

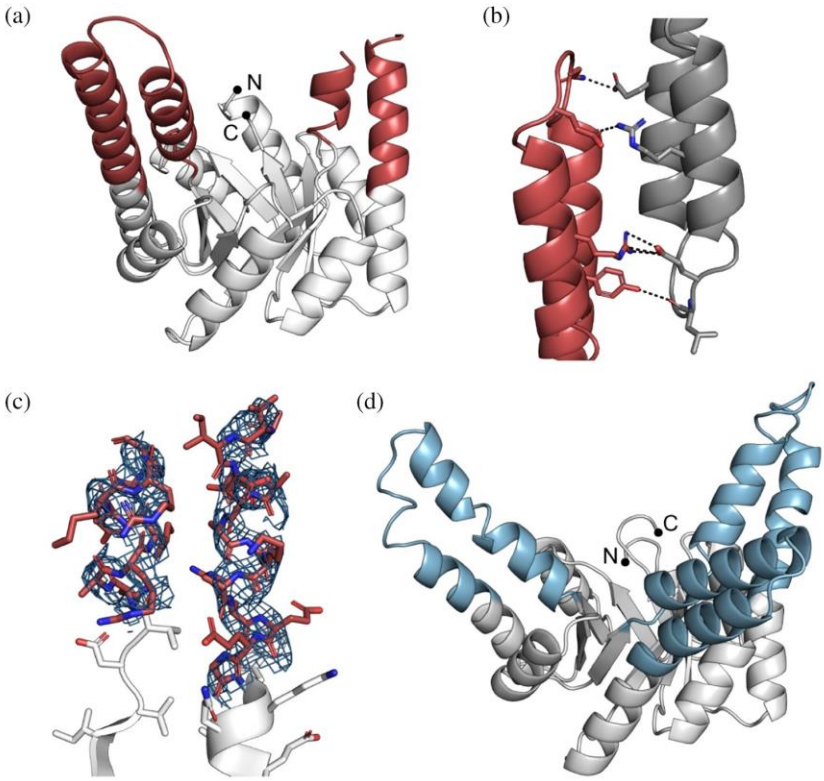


FIGURE 3 Structural details of HalluTIM2-2 and HalluTIM3-1. All structures are displayed in cartoon representation with black dots highlighting the termini. The base scaffold is shown in white. Extensions of HalluTIM2-2 and HalluTIM3-1 are colored in red and blue, respectively. (a) Overall structure of HalluTIM2-2 (chain A, PDB-ID: 8R8N). (b) Resolved helical extension of HalluTIM2-2 forms multiple crystal contacts with its symmetry mate (in gray). Contacts such as polar interactions and hydrogen bonds are shown as black dashed lines. (c) Partially resolved second extension in HalluTIM2-2 shown as stick representation in red with the corresponding electron density in blue (2Fo-Fc map contoured at 1.0 RMSD). (d) Overall structure of HalluTIM3-1 (PDB-ID: 8R8O).

(Figure S3), that may influence the quality of the data and in combination with a certain flexibility of the extensions, lead to the rather noisy diffraction data.

2.4 | Solution states match structures despite crystal contacts

Upon comparison of the obtained crystal structures to corresponding structure predictions using ColabFold, we observed an accurate prediction for HalluTIM2-2 with a root mean square deviation (RMSD) over all Cα atoms of about 1.2 Å but found major differences in the case of HalluTIM3-1 as the RMSD over all Cα atoms is over 4.1 Å (Figure 4a,b). These discrepancies are mainly due to the different angles of the extensions from the barrel core, especially for insertion 1 that does not form a continuous α-helix. The structure prediction shows straighter extensions, whereas two of the extensions in the crystal structure tilt more to the outside. When comparing each individual extension with the corresponding prediction, we observe accurate predictions below 1.0 Å RMSD except for the first extension of HalluTIM3-1, which shows a higher deviation with 2.34 Å (Table S4). To obtain an impression of the protein structure in solution, we measured size exclusion chromatography small angle x-ray scattering (SEC-SAXS) with both constructs and sTIM11-SB. The experimental data indicate globular proteins, whereby both HalluTIMs show a slightly higher flexibility in comparison to the base scaffold (Figure S4). For a comparison with the structures, we calculated a theoretical scatter curve for each crystal structure as well as each prediction and fitted it to the experimental curve (Franke et al., 2017). In the case of HalluTIM2-2, the theoretical scatter curves of both the crystal structure and the predicted structure are in overall agreement with the experimental data, with a χ^2 of 2.6 and 2.4, respectively (Figure 4c). However, around 0.18 $Å^{-1}$ both theoretical scatter curves diverge from the experimental data, suggesting a potential high flexibility in the extensions, which is especially conceivable for the partially resolved one. For HalluTIM3-1, where the crystal structure and prediction differ, we obtained varying qualities of the fits. The theoretical scattering curve of the predicted structure shows a high χ^2 of 8.9, whereas the crystal structure matches the experimental data with a significantly lower χ^2 of 2.0 (Figure 4d), indicating that the crystal structure matches the protein in solution more closely.

Next, we searched for newly introduced pockets within the crystal structures employing the AI-based ligand-binding site prediction tool PUResNET (Kandel

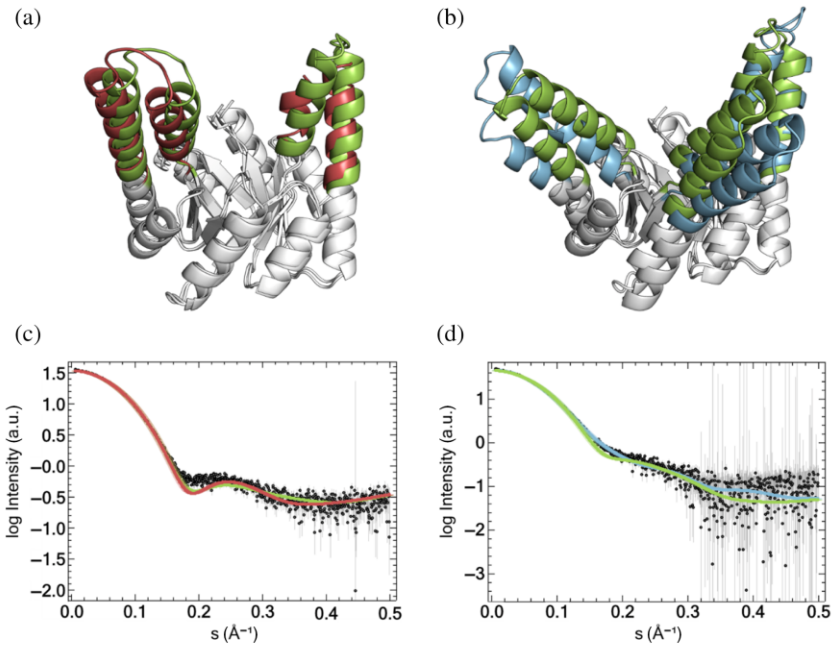


FIGURE 4 Structural comparison and size exclusion chromatography small angle x-ray scattering (SEC-SAXS) analysis of the crystal structures and structure predictions. All structures are displayed as cartoon representation and superimposed over all Cα atoms. The base scaffold is shown in white. Extensions of the structure predictions are colored green. Extensions of HalluTIM2-2 and HalluTIM3-1 are shown in red and blue, respectively. SEC-SAXS experimental scattering data are displayed as black dots. Theoretical scattering curves for the structures are shown in the same color code as in the structural comparison. (a) Superimposition of the experimentally determined and the predicted structures of HalluTIM2-2 (RMSD: 1.2 Å). (b) Superimposition for HalluTIM3-1 (RMSD: 4.1 Å). (c) SEC-SAXS data analysis and comparison to the structures for HalluTIM2-2 (crystal structure χ^2: 2.6, AlphaFold2 prediction χ^2: 2.4). (d) SEC-SAXS data analysis for HalluTIM3-1 (crystal structure χ^2: 2.0, AlphaFold2 prediction χ^2: 8.9).

et al., 2021). When analyzing the starting scaffold, a shallow pocket is predicted near the N- and C-termini above the inner β-sheet (Figure S5A). In contrast, in HalluTIM2-2 and HalluTIM3-1, a major pocket is formed through the introduced fragments above the C-terminal end of the inner β-sheet (Figure S5B,C). These pockets show differences in size with pocket volumes of 1006 Å^3 in HalluTIM3-1 and 2000 Å^3 in HalluTIM2-2 (Table S5).

3 | DISCUSSION

Despite the rapidly growing number of *de novo* TIM barrels, the designs lack the feature of extended surfaces or cavities necessary to introduce catalytic function. We used the recently developed AI-based method of constrained hallucination from Wang et al. to introduce new structural features on top of the TIM barrel topology (Wang et al., 2022). The insertion sites were selected based on the already successful introduction of different secondary structure elements into a descendant of sTIM11 (Kordes et al., 2023; Wiese et al., 2021), whereby not all three insertion sites had been used simultaneously so far. Through the introduction of two or three extensions, we aimed to generate extended surfaces that allow the formation of cavities. The methods of inpainting and constrained hallucination can both be used to generate insertions with comparable quality. We chose constrained hallucination over inpainting as it is stated to lead to increased structural variability (Wang et al., 2022). In our setup, rather than attaining significant structural diversity, we instead observed only the elongation of the outer α-helix plus a second smaller α-helix forming a hairpin located above the barrel. The bias toward helical extensions might be due to the already-existing helix serving as a seed. To generate greater diversity, the newly developed RFdiffusion application might now be utilized to explore more variable insertion sites across all βα-loops, thereby encompassing a broader range of insertion lengths (Watson et al., 2023). As RFdiffusion was only published after we completed our computational workflow, which generated high-quality designs, we did not consider restarting the design process with RFdiffusion. Since helix–loop–helix motifs

can build up a cavity, as demonstrated before (Kordes et al., 2023), we continued with our designs and optimized the sequence of the inserted fragment with proteinMPNN. Sequence optimization was focused on the extensions rather than the entire barrel to preserve the structurally robust scaffold, thereby providing a set of diversified HalluTIM variants.

4 | CONCLUSION

Constrained hallucination in combination with proteinMPNN is a powerful method for the extension of protein loops. Here, we introduced two or three helical insertions into minimal loops in the *de novo* designed TIM barrel sTIM11-SB. Six HalluTIMs were selected for experimental testing. All of them were found in the soluble fraction after expression in *E. coli*, possibly promoted by the preservation of the base scaffold. Moreover, all HalluTIMs showed a monomeric state and an increased hydrodynamic radius compared to sTIM11-SB. Multiple HalluTIMs revealed an increase in α-helicity by CD spectroscopy, indicating the formation of α-helical extensions. Upon analysis of protein stability, we observed that the extensions in some cases even led to stabilization, indicating the robustness of HalluTIMs for further downstream functionalization. As we were able to introduce three extensions, we attempted to introduce an additional fourth extension to build up the cavity further. Following the symmetry of the already successfully introduced extensions, the fourth one would be located at the termini of the TIM barrel. However, any attempt to build a similar extension by elongation of the termini with constrained hallucination was not successful. The introduced extensions did not show any interactions and rather extended separately away from the rest of the protein (Figure S6). This suggests that elongation of the termini is a more challenging design task for constrained hallucination than the other used insertion sites.

Two of the designs could be crystallized and their structures determined, which we consider an incredible success rate. The crystal structures validate the successful incorporation of the hallucinated extensions. A high amount of crystal contacts could be observed within the introduced α-helices. This can be rationalized by the sequence optimization with proteinMPNN, which is suggested to generate protein surfaces more likely to form crystal contacts (Wicky et al., 2022). SAXS measurements support the crystal structure despite variations to the structure predictions. Some variation between crystal and solution structure can, however, be expected due to the inherent flexibility of the elongated helical hairpins.

In another study, we used a highly rational and physics-based approach (Kordes et al., 2023) to incorporate helix–loop–helix motifs into a similar scaffold. Despite entirely different workflows, the resulting designs share similar extensions, pocket formation (Table S5), and the same distinct relationship to natural TIM barrels within a DALI database search, for example, class II fructose-bisphosphate aldolase (Holm, 2022). Differences can be found in the success rate of the two design workflows. The design workflow by Kordes et al. (2023) generated four designs, of which only two showed soluble expression and no structure could be solved. In contrast, our machine learning-based design workflow exclusively produced soluble proteins and two structures could be solved providing structural data necessary for future design of ligand-binding or enzymatic sites.

The TIM barrel sTIM11 was already used for functionalization by fusing one half of the barrel to a *de novo* designed ferredoxin, which dimerizes and binds a lanthanide (Caldwell et al., 2020). In contrast to this functionalized protein, we preserved the TIM-barrel fold in a monomeric fashion, thereby providing a continuous scaffold to explore a broader spectrum of functions based on the potential of natural TIM barrels (Nagano et al., 2002).

5 | MATERIALS AND METHODS

5.1 | Biochemical materials

All reagents were analytical grade from Sigma-Aldrich or Carl Roth, except when indicated. All solutions were prepared with double-distilled water. Constructs were codon optimized by BioCat and ordered already cloned in pET21b(+) vector.

5.2 | Computational extension of a *de novo* TIM barrel

For the modeling and analysis of the extensions into sTIM11-SB (PDB-ID: 7OSU), the constrained hallucination method from Wang et al. (2022) was used. During all design steps, the backbone position and amino acid identity of the residues not involved in the design process were restricted. For an initial round of constrained hallucination different combinations of βα-loops of sTIM11-SB were chosen as insertion sites. For each insertion site, extensions in the range of 25–35 residues were allowed. One-hundred were modeled using 600 steps of gradient descent. The resulting designs were relaxed and scored using Rosetta (Leaver-Fay et al., 2011). Structures of the designs were predicted with AlphaFold2 using the Model

4 weights (Jumper et al., 2021). Designs were filtered based on their average predicted local distance difference test (pLDDT) and Rosetta scores. The best design was passed on for a second round of constrained hallucination. Hereby, the insertion site was chosen between the top of the outer hallucinated α-helix and the end of the β-strand of the TIM barrel. The range of an allowed extension was shortened to 19–26 residues. Modeling and filtering were performed identical to the first round of constrained hallucination. Based on a visual inspection of the top scoring designs, particularly with respect to the transition region from the outer α-helix of the barrel to the extension and the packing of the α-helix extensions against each other, designs were chosen for a sequence optimization with proteinMPNN (Dauparas et al., 2022). For each chosen backbone, 16 sequences with the full protein backbone model and a temperature factor of 0.2 were generated, whereby everything except the extensions were restricted to their original amino acid identities. For all generated sequences, structures were predicted using ColabFold (v1.3.0) with all five model weights (Mirdita et al., 2022). The prediction with the highest average pLDDT score was selected as the final structure prediction for this sequence. Based on these pLDDT scores and visual inspection as described above, designs were chosen for experimental characterization (Tables S1 and S2).

5.3 | Overexpression and protein purification

E. coli BL21(DE3) cells (Novagen) were transformed with plasmid, plated on agar plates containing 100 μg mL^{-1} ampicillin, and incubated over night at 37°C. From these plates, single colonies were picked to inoculate Lysogeny Broth (LB) media supplemented with ampicillin (100 μg mL^{-1}) and incubated at 30°C overnight. For protein expression, 1 L LB was inoculated with 10 mL of the preculture and incubated at 37°C until OD_{600} reached a value of 0.6–0.8. Overexpression was induced by adding isopropyl-β-thiogalactoside to a final concentration of 0.1 mM. Cultures were further incubated at 20°C overnight. On the next day, cells were harvested by centrifugation (Beckman Coulter Avanti J-26 XPI, JLA-8.1000, 15 min, 4000 g, 4°C) and pellets were either frozen at −20°C until usage or directly resuspended in 35 mL of buffer A (35 mM of NaP pH 8.0, 150 mM of NaCl, and 10 mM of imidazole). The resuspended cells were lysed by sonication (Branson Ultrasonic Sonifier 250, output 4, duty cycle 40%, 3 × 3 min) and centrifuged (Beckman Coulter Avanti J-26 XPI, JA-25.50, 1 h, 40,000 g, 4°C). The supernatant was loaded onto a HisTrapHP column (5 mL, Cytiva Life Science) equilibrated with buffer A and coupled to an ÄKTApure system (Cytiva Life Science). After washing with 10 column volumes (CV) of buffer A, the protein was eluted with a linear gradient over 20 CV to 60% buffer B (35 mM of NaP pH 8.0, 150 mM of NaCl, and 500 mM imidazole). Fractions containing the protein were pooled, concentrated with a centrifugal concentrator, and loaded onto a HiLoad 26/600 Superdex 75 preparative grade column (Cytiva Life Sciences) preequilibrated in buffer C (35 mM of NaP pH 8.0, 150 mM of NaCl). Elution was performed with 1 CV buffer C. Fractions with monomeric protein were pooled. For some subsequent experiments, the protein was dialyzed into buffer D (10 mM of NaP, pH 8). Protein concentration was determined photometrically using the absorption at 280 nm. Expression and purification were checked by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE).

5.4 | Size exclusion chromatography-multi angle light scattering

SEC-MALS measurements were performed using a Superdex 75 Increase 10/300 GL column (Cytiva Life Sciences) connected to an Agilent 1260 Infinity II HPLC system, coupled to a miniDAWN MALS detector and an Optilab differential refractive index detector (dRI) (Wyatt Technology). For all experiments, a protein concentration of 2 mg mL^{-1}, a flowrate of 0.8 mL min^{-1}, an injection volume of 100 μL, and buffer C with the addition of 0.02% NaN_3 were used. Data collection and analysis were performed with the ASTRA 8.0.2.5 software (Wyatt Technology). For the analysis of each run, the signal of the dRI detector was used for protein concentration determination. A bovine serum albumin (BSA) standard at 2 mg mL^{-1} was used for MALS detector normalization, correction of peak alignment, peak broadening, and reproducibility.

5.5 | Far-Ultraviolet circular dichroism

CD spectra were collected with a Jasco J-710. Experiments were performed in buffer D using a protein concentration of 0.2 mg mL^{-1}. Far-Ultraviolet-CD spectra were recorded in the range of 190–260 nm at 20°C in a 1 mm cuvette, with a 1 nm bandwidth, 1 s response time, and scanning speed of 100 nm min^{-1}. For each protein, 10 spectra were accumulated. Data were normalized by subtraction of a buffer spectrum and conversion to mean residue molar ellipticity using: $[\theta_{MRE}] = (M \times \theta)/(10 \times d \times c)$ and $M = MW/(n-1)$, where M is the mean

residue weight, MW the molecular weight in Da, n the number of residues in the protein, θ the collected ellipticity in mdeg, d the path length in cm, and c the protein concentration in mg mL^{-1}.

To measure thermostability of the proteins, thermal unfolding was followed by CD at 222 nm. The samples were heated up to 95°C with a rate of 1°C min^{-1}. Measured unfolding curves were analyzed with the Denatured Protein function of SpectraAnalysis 1.53.07 (Jasco). Dependencies in the initial and final baselines were fitted and subtracted before unfolding parameters were determined. Each parameter was determined from measurements of two individually purified samples and averaged. $\Delta G_{25°C}$ values were calculated from the obtained values for ΔH and ΔS by using the Gibbs–Helmholtz equation with $T = 298$ K. In addition, spectra were collected after the heating process at 95°C and after cooling to 20°C with the parameters described above.

5.6 | Crystallization and structure determination

Initial crystallization screens using the sitting drop vapor diffusion method were set up using a Phoenix pipetting robot (Art Robbins Instruments) with commercially available sparse-matrix screens (NeXtal) in 96-well sitting-drop plates (3-drop Intelli-Plates, Art Robbins Instruments). Droplets were pipetted in 1:1, 1:2, and 2:1 ratios of protein: reservoir solution with a protein concentration of 25 mg mL^{-1} for HalluTIM3-1 and 20 mg mL^{-1} for HalluTIM2-2. Plates were incubated at 293 K. Hits for HalluTIM3-1 were obtained in 0.08 M sodium acetate pH 4.6, 1.6 M ammonium sulfate, 20% (v/v) glycerol after 3 days and for HalluTIM2-2 in 0.2 M lithium sulfate, 0.1 M Tris pH 8.6, 25 % polyethylene glycol 8000 after 2 days.

Crystals for HalluTIM2-2 were further optimized using the initial hit and setting up hanging drops in 15-well EasyXtal plates (NeXtal). The best diffracting crystals were obtained in the initial condition composition. Cryoprotection was achieved by the addition of glycerol to a final concentration of 25%.

Crystals for HalluTIM3-1 were further optimized using the initial hit and setting up sitting drops in 48-well MRC Maxi crystallization plates (Swissci). The best diffracting crystals were obtained in 0.08 M sodium acetate pH 4.9, 1.55 M ammonium sulfate, and 20% (v/v) glycerol.

Crystals were manually mounted using cryo-loops on SPINE standard bases and flash-cooled in liquid nitrogen. Diffraction data for HalluTIM3-1 were collected on P13 operated by European Molecular Biology Laboratory (EMBL) Hamburg at the PETRA III storage ring (DESY, Hamburg, Germany) and for HalluTIM2-2 on ID30B at the European Synchrotron Radiation Facility (ESRF) electron-storage ring (Nanao et al., 2022). Measurements were performed at 100 K in single-wavelength mode at 0.9762 Å with a Dectris EIGER X 16 M for HalluTIM3-1 and at 0.8731 Å with a Dectris EIGER2 X 9 M detector for HalluTIM2-2 in fine-slicing mode in 0.1° and 0.05° wedges, respectively, using the *MXCuBE* beamline-control software (Oscarsson et al., 2019). Data were processed with X-ray Detector Software APP3 (*XDSAPP3*) (Sparta et al., 2016) employing *XDS* (Kabsch, 2010). Data quality was assessed by applying *phenix.xtriage* (Liebschner et al., 2019).

Phases were solved by molecular replacement using the respective model as search model with *Phaser* (McCoy et al., 2007). The resulting models were manually rebuilt with *Coot* (Emsley et al., 2010) and refined with *phenix.refine* (Afonine et al., 2012) in an iterative manner. Coordinates and structure factors were validated and deposited in the PDB (Burley et al., 2023) with accession codes 8R8N (HalluTIM2-2) and 8R8O (HalluTIM3-1).

5.7 | Size exclusion chromatography small angle x-ray scattering

SEC-SAXS measurements were performed at the BioSAXS beamline BM29 at the ESRF in Grenoble, France. For all experiments, a protein concentration of 5 mg mL^{-1}, an AdvanceBio Sec 130 Column with a flow-rate of 0.16 ml min^{-1}, an injection volume of 50 µL and buffer C with the addition of 1 mM dithiotreitol (DTT) were used. Data processing of the experimental scattering curves and analysis were performed with the software suite ATSAS 3.2.1 and BioXTAS RAW (Hopkins et al., 2017; Manalastas-Cantos et al., 2021). For each measured protein, a theoretical scattering curve with the crystal structure and the structure prediction was calculated and fitted to the experimental data using CRYSOL with standard parameters (Franke et al., 2017).

AUTHOR CONTRIBUTIONS

Julian Beck: Conceptualization; investigation; methodology; data curation; visualization; writing – original draft; writing – review and editing. **Sooruban Shanmugaratnam:** Investigation; data curation; visualization; writing – original draft; writing – review and editing. **Birte Höcker:** Conceptualization; funding acquisition; writing – original draft; writing – review and editing; resources; methodology.

ACKNOWLEDGMENTS

We acknowledge the beamline staff at DESY for their support during crystal measurements and the beamline staff at ESRF for their support during crystal and SAXS measurements. We further thank Janosch Hennig for providing his expertise in SAXS data analysis and Sabrina Wischt for technical assistance. Support from the Elite Network of Bavaria and its study program "Biological Physics" is gratefully acknowledged. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work was supported through core funding of the University of Bayreuth.

ORCID

Julian Beck https://orcid.org/0009-0007-3555-9890
Sooruban Shanmugaratnam https://orcid.org/0000-0002-2614-6046
Birte Höcker https://orcid.org/0000-0002-8250-9462

REFERENCES

Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, et al. Towards automated crystallographic structure refinement with phenix.Refine. Acta Crystallogr D Biol Crystallogr. 2012;68(4):352–67. https://doi.org/10.1107/S0907444912001308

Anand N, Eguchi R, Mathews II, Perez CP, Derry A, Altman RB, et al. Protein sequence design with a learned potential. Nat Commun. 2022;13(1):1–11. https://doi.org/10.1038/s41467-022-28313-9

Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, et al. De novo protein design by deep network hallucination. Nature. 2021;600(7889):547–52. https://doi.org/10.1038/s41586-021-04184-w

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–6. https://doi.org/10.1126/science.abj8754

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB protein data bank (RCSB.Org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res. 2023;51(D1):D488–508. https://doi.org/10.1093/nar/gkac1077

Caldwell SJ, Haydon IC, Piperidou N, Huang PS, Bick MJ, Sebastian Sjöström H, et al. Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. Proc Natl Acad Sci U S A. 2020;117(48):30362–9. https://doi.org/10.1073/pnas.2008535117

Chu AE, Fernandez D, Liu J, Eguchi RR, Huang P-S. De novo design of a highly stable ovoid TIM barrel: unlocking pocket shape towards functional design. Biodes Res. 2022;2022:9842315. https://doi.org/10.34133/2022/9842315

Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning-based protein sequence design using ProteinMPNN. Science. 2022;378(6615):49–56. https://doi.org/10.1126/science.add2187

Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, et al. De novo design of a fluorescence-activating β-barrel. Nature. 2018;561(7724):485–91. https://doi.org/10.1038/s41586-018-0509-0

Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, et al. Rational design of α-helical tandem repeat proteins with closed architectures. Nature. 2015;528(7583):585–8. https://doi.org/10.1038/nature16191

Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of coot. Acta Crystallogr D Biol Crystallogr. 2010;66(4):486–501. https://doi.org/10.1107/S0907444910007493

Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Cryst. 2017;50(Pt 4):1212–25. https://doi.org/10.1107/S1600576717007786

Goverde CA, Pacesa M, Dornfeld LJ, Georgeon S, Rosset S, Dauparas J, et al. Computational design of soluble analogues of integral membrane protein structures. BioRxiv. 2023. https://doi.org/10.1101/2023.05.09.540044

Holm L. Dali server: structural unification of protein families. Nucleic Acids Res. 2022;50(W1):W210–5. https://doi.org/10.1093/NAR/GKAC387

Hopkins JB, Gillilan RE, Skou S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. J Appl Cryst. 2017;50(5):1545–53. https://doi.org/10.1107/S1600576717011438

Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016;537(7620):320–7. https://doi.org/10.1038/nature19946

Huang P-SS, Feldmeier K, Parmeggiani F, Fernandez Velasco DA, Höcker B, Baker D. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat Chem Biol. 2016;12(1):29–34. https://doi.org/10.1038/nchembio.1966

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2

Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):125–32. https://doi.org/10.1107/S0907444909047337

Kandel, J., Tayara, H., & Chong, K. T. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. Journal of Cheminformatics, 2021;13(1):1–14. doi:10.1186/S13321-021-00547-7

Kim DE, Jensen DR, Feldman D, Tischer D, Saleem A, Chow CM, et al. De novo design of small beta barrel proteins. Proc Natl Acad Sci U S A. 2023;120(11):e2207974120. https://doi.org/10.1073/pnas.2207974120

Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, et al. Principles for designing ideal protein structures. Nature. 2012;491(7423):222–7. https://doi.org/10.1038/nature11600

Kordes S, Beck J, Shanmugaratnam S, Flecks M, Höcker B. Physics-based approach to extend a de novo TIM barrel with rationally designed helix–loop–helix motifs. Protein Eng Des Sel. 2023;36:gzad012. https://doi.org/10.1093/protein/gzad012

Kordes S, Romero-Romero S, Lutz L, Höcker B. A newly introduced salt bridge cluster improves structural and biophysical

properties of de novo TIM barrels. Protein Sci. 2022;31(2):513–27. https://doi.org/10.1002/pro.4249

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74. https://doi.org/10.1016/B978-0-12-381270-4.00019-6

Liebschner D, Afonine PV, Baker ML, Bunkoczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in phenix. Acta Crystallogr D Struct Biol. 2019;75(Pt 10):861–77. https://doi.org/10.1107/S2059798319011471

Manalastas-Cantos K, Konarev PV, Hajizadeh NR, Kikhney AG, Petoukhov MV, Molodenskiy DS, et al. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. J Appl Cryst. 2021;54(Pt 1):343–55. https://doi.org/10.1107/S1600576720013412

Marcos E, Chidyausiku TM, McShan AC, Evangelidis T, Nerli S, Carter L, et al. De novo design of a non-local β-sheet protein with high stability and accuracy. Nat Struct Mol Biol. 2018;25:1028–34. https://doi.org/10.1038/s41594-018-0141-6

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Cryst. 2007;40(4):658–74. https://doi.org/10.1107/S0021889807021206

Minami S, Kobayashi N, Sugiki T, Nagashima T, Fujiwara T, Tatsumi-Koga R, et al. Exploration of novel αβ-protein folds through de novo design. Nat Struct Mol Biol 2023. 2023;118(3):1–9. https://doi.org/10.1038/s41594-023-01029-0

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679–82. https://doi.org/10.1038/s41592-022-01488-1

Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol. 2002;321:741–65. https://doi.org/10.1016/S0022-2836(02)00649-6

Nanao M, Basu S, Zander U, Giraud T, Surr J, Guijarro M, et al. ID23-2: an automated and high-performance microfocus beamline for macromolecular crystallography at the ESRF. J Synchroton Radiat. 2022;29(2):581–90. https://doi.org/10.1107/S1600577522000984

Oscarsson M, Beteva A, Flot D, Gordon E, Guijarro M, Leonard G, et al. MXCuBE2: the dawn of MXCuBE collaboration. J Synchroton Radiat. 2019;26(2):393–405. https://doi.org/10.1107/S1600577519001267

Pan X, Kortemme T. Recent advances in de novo protein design: principles, methods, and applications. J Biol Chem. 2021;296:100558. https://doi.org/10.1016/j.jbc.2021.100558

Romero-Romero S, Costas M, Silva Manzano DA, Kordes S, Rojas-Ortega E, Tapia C, et al. The stability landscape of de novo TIM barrels explored by a modular design approach. J Mol Biol. 2021;433(18):167153. https://doi.org/10.1016/j.jmb.2021.167153

Romero-Romero S, Kordes S, Michel F, Höcker B. Evolution, folding, and design of TIM barrels and related proteins. Curr Opin Struct Biol. 2021;68:94–104. https://doi.org/10.1016/j.sbi.2020.12.007

Sparta KM, Krug M, Heinemann U, Mueller U, Weiss MS. XDSAPP2.0. J Appl Cryst. 2016;49(3):1085–92. https://doi.org/10.1107/S1600576716004416

Sterner R, Höcker B. Catalytic versatility, stability, and evolution of the (βα)8-barrel enzyme fold. Chem Rev. 2005;105(11):4038–55. https://doi.org/10.1021/cr030191z

Thoma R, Hennig M, Sterner R, Kirschner K. Structure and function of mutationally generated monomers of dimeric phosphoribosylanthranilate isomerase from Thermotoga maritima. Structure. 2000;8(3):265–76. https://doi.org/10.1016/S0969-2126(00)00106-4

Vijayabaskar MS, Vishveshwara S. Insights into the fold organization of tim barrel from interaction energy based structure networks. PLoS Comput Biol. 2012;8(5):e1002505. https://doi.org/10.1371/journal.pcbi.1002505

Wang J, Lisanza S, Juergens D, Tischer D, Watson JL, Castro KM, et al. Scaffolding protein functional sites using deep learning. Science. 2022;377(6604):387–94. https://doi.org/10.1126/science.abn2100

Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023;2023:1–3. https://doi.org/10.1038/s41586-023-06415-8

Wicky BIM, Milles LF, Courbet A, Ragotte RJ, Dauparas J, Kinfu E, et al. Hallucinating symmetric protein assemblies. Science. 2022;378(6615):2023–61. https://doi.org/10.1126/science.add1964

Wierenga RK. The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett. 2001;492(3):193–8. https://doi.org/10.1016/S0014-5793(01)02236-0

Wiese JG, Shanmugaratnam S, Höcker B. Extension of a de novo TIM barrel with a rationally designed secondary structure element. Protein Sci. 2021;30(5):982–9. https://doi.org/10.1002/pro.4064

Yang C, Sesterhenn F, Bonet J, van Aalen EA, Scheller L, Abriata LA, et al. Bottom-up de novo design of functional proteins with complex structural features. Nat Chem Biol. 2021;17(4):492–500. https://doi.org/10.1038/s41589-020-00699-x

Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9. https://doi.org/10.1093/NAR/GKI524

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Beck J, Shanmugaratnam S, Höcker B. Diversifying *de novo* TIM barrels by hallucination. Protein Science. 2024;33(6):e5001. https://doi.org/10.1002/pro.5001

Supporting information

Diversifying *de novo* TIM barrels by hallucination

Julian Beck[1], Sooruban Shanmugaratnam[1], Birte Höcker[1]*

[1] Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.

* Corresponding author:

Birte Höcker, e-mail address: birte.hoecker@uni-bayreuth.de

This file includes:

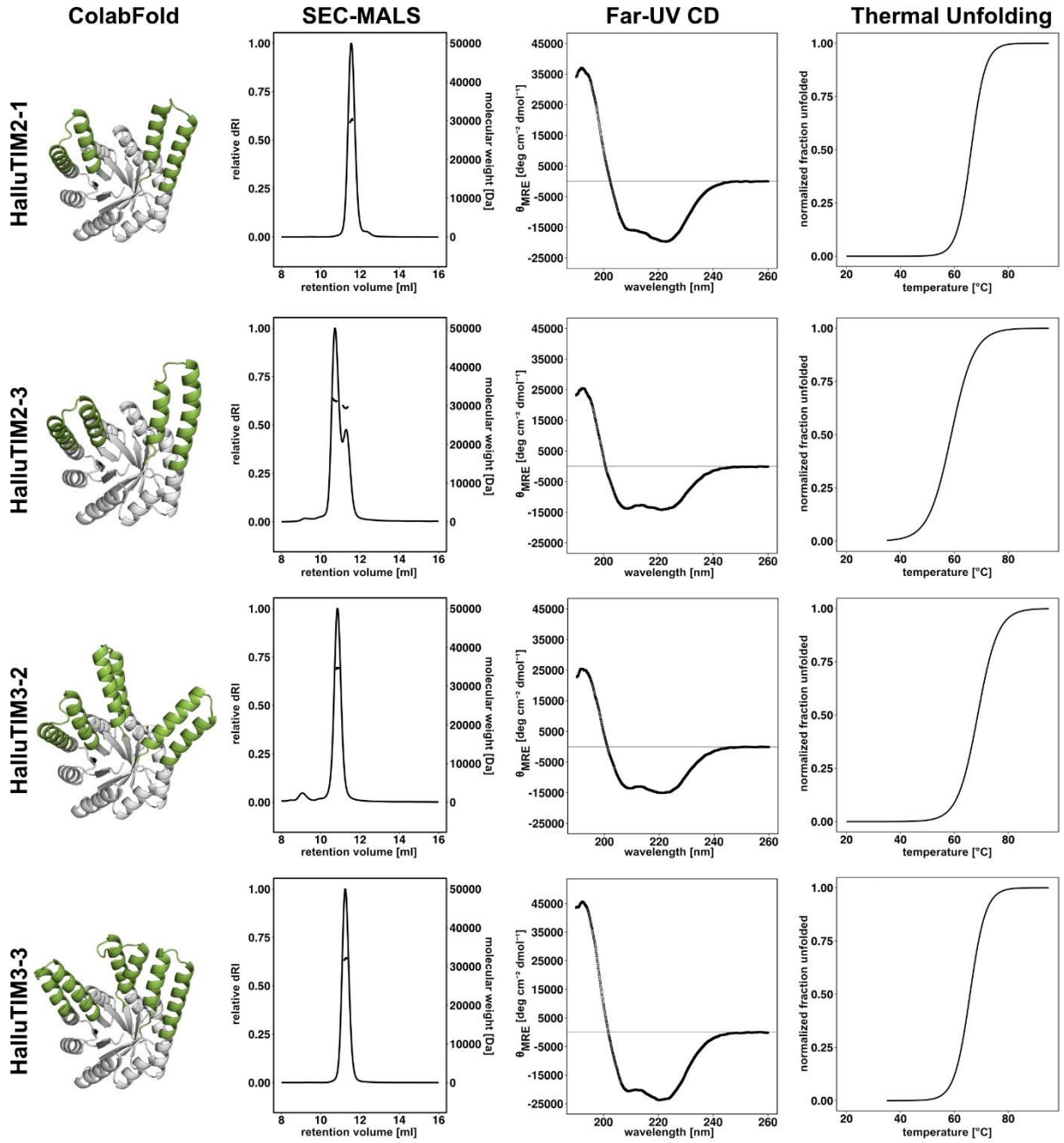


Figure S1. Biochemical characterization of additional HalluTIMs. For each construct the structure prediction with ColabFold and the characterization with SEC-MALS, Far UV-CD and thermal unfolding is shown. Within the structure predictions the base scaffold is shown in white and the extensions in green. Elution profile of the SEC-MALS measurements showing the normalized relative differential refractive index as solid black line and the calculated molar mass as data points in black. Far UV-CD spectra and thermal unfolding are displayed in black. For numerical results of the experimental characterization see Table S3.

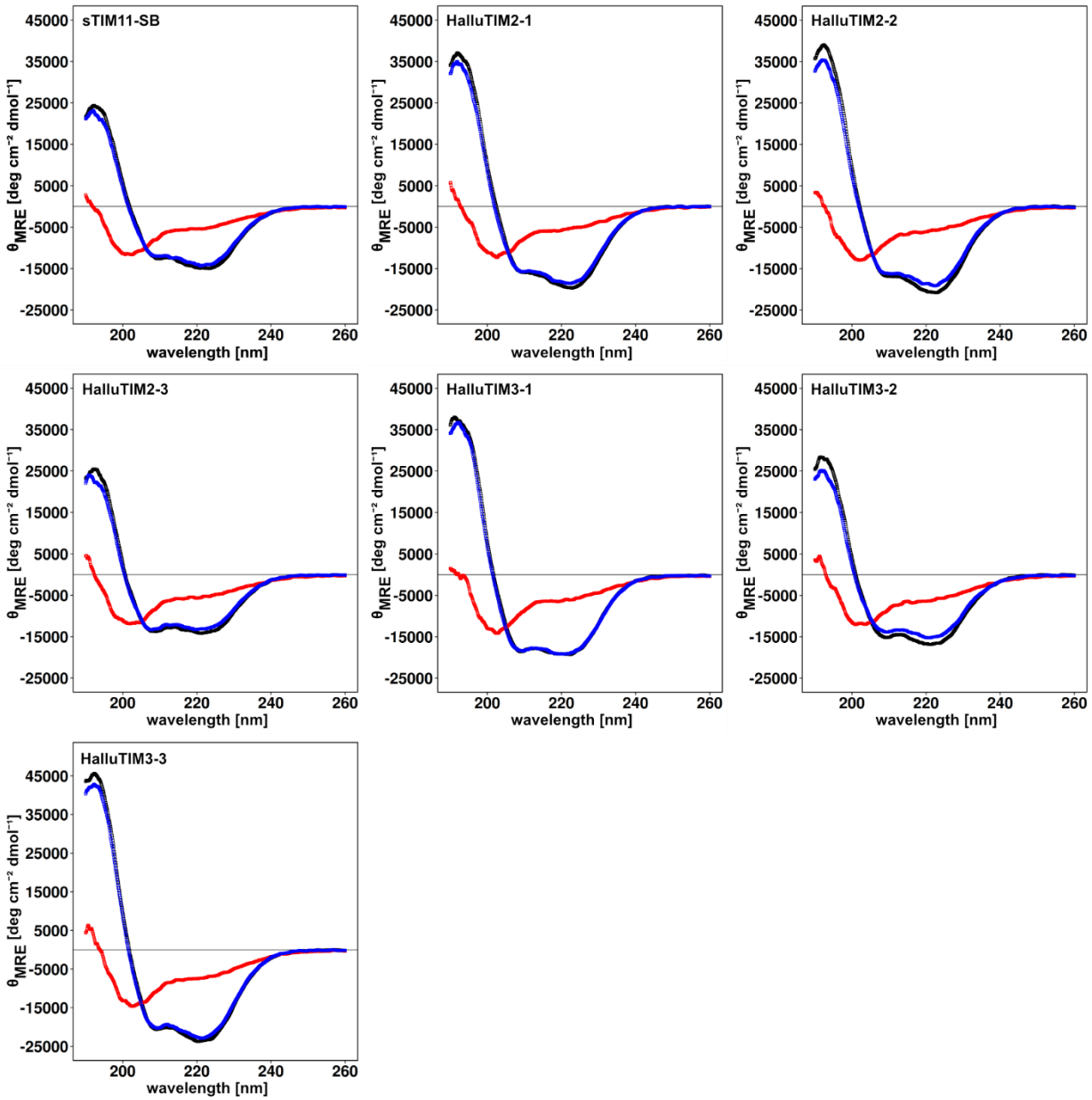


Figure S2. Far UV-CD measurements to determine the reversibility of thermal unfolding. For each construct an initial far UV-CD spectrum is displayed in black, a far UV-CD spectrum at 95 °C in red and a far UV-CD spectrum after cooling down in blue. All HalluTIMs maintain the reversible unfolding behavior of sTIM11-SB.

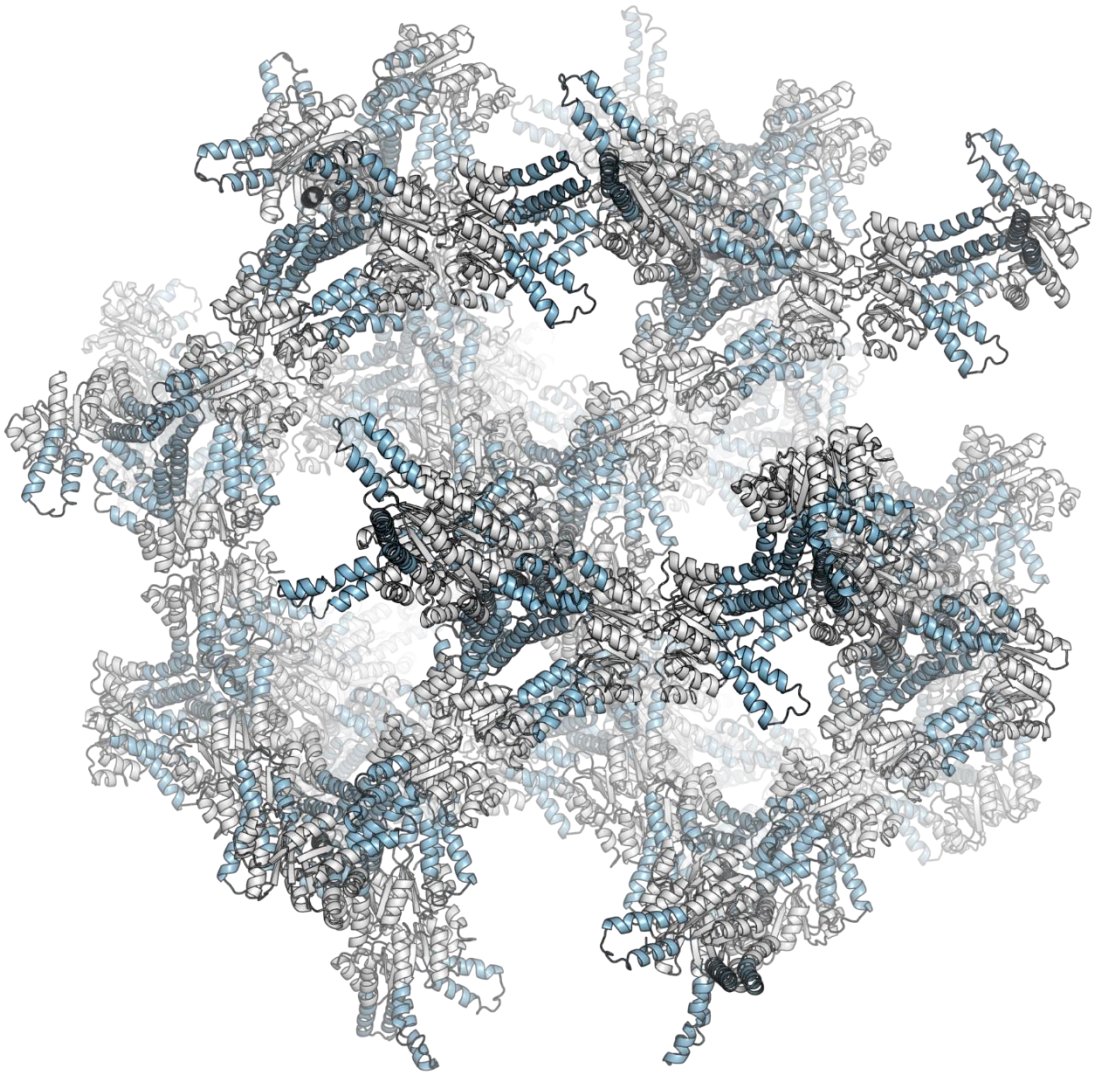


Figure S3. Crystal packing with large void volumes of HalluTIM3-1. Structures are displayed as cartoon representation with the base scaffold colored in white and the extensions in blue. Symmetry mates around 100 Å are shown. Crystal contacts are mainly formed within the extensions resulting in a crystal packing with large void volumes.

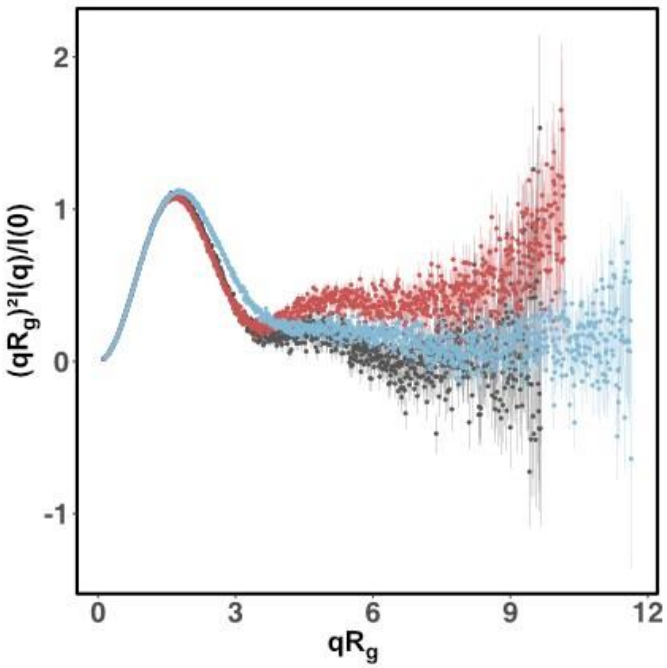


Figure S4. Dimensionless Kratky plot. Datapoints and error bars for sTIM11-SB colored in grey, for HalluTIM2-2 in red and for HalluTIM3-1 in blue. Measurements indicate globular proteins and slightly higher flexibility of HalluTIM2-2 and HalluTIM3-1 in comparison to sTIM11-SB.

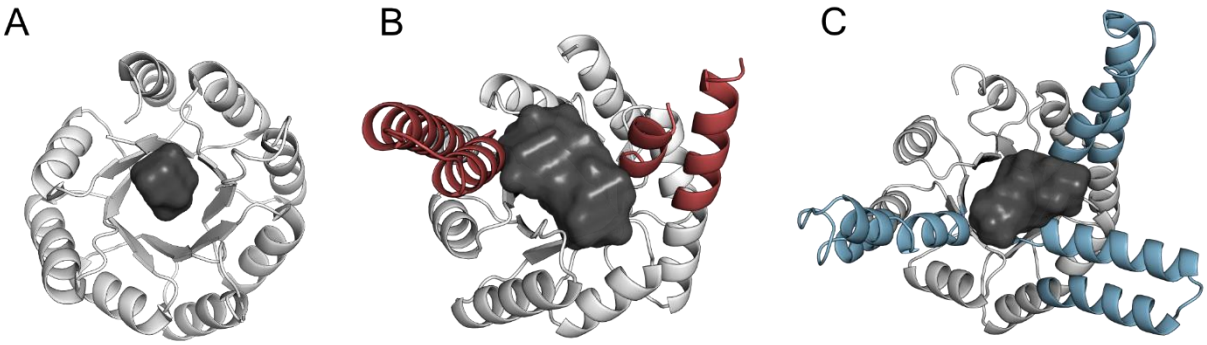


Figure S5. PUResNET pocket prediction. Structures are displayed as cartoon representation. The base scaffold is shown in white. Extensions of HalluTIM2-2 and HalluTIM3-1 are shown in red and blue, respectively. Predicted pockets are displayed as surface representation and colored in black. **A)** Pocket prediction for sTIM11-SB. **B)** Pocket prediction for HalluTIM2-2. **C)** Pocket prediction for HalluTIM3-1.

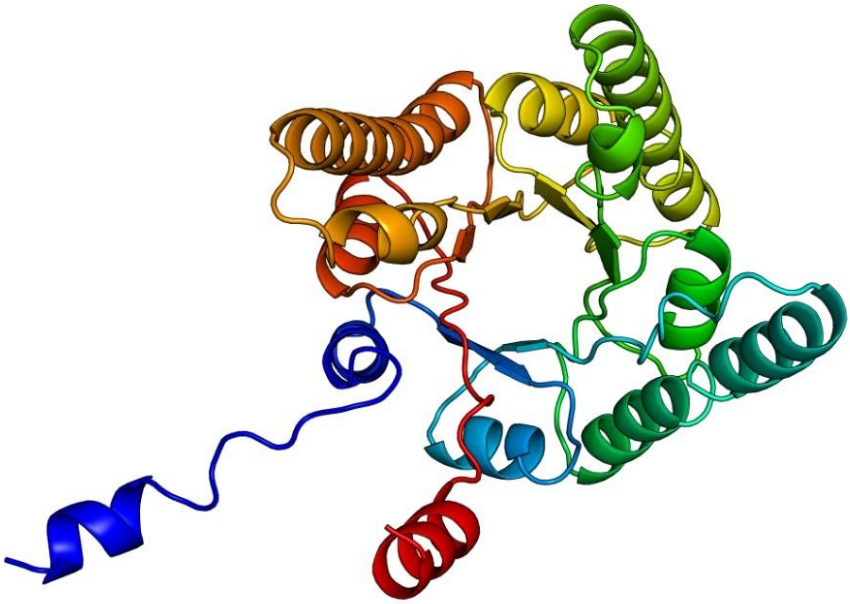


Figure S6. Hallucination of a fourth extension. The predicted structure is displayed as cartoon representation in rainbow coloring. The extension of the termini with constrained hallucination resulted only in non-interacting elongations.

Table S1. Protein sequences (extensions highlighted in bold).

Name	Sequence
HalluTIM2-1	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**VDKADEYRKKAEE VAKKTGNFKPLVDKYLAEAEKARDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWK KGADILIVDATDKNEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DADER VERRKEELKKLGLTDPEVIEKAREEARREA**WKQVEQLRREGATEIAYRSDDWRDLKE AWKKGADILIVDATLEHHHHHH
HalluTIM2-2	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**SSKADDYRARAAA AAKELGNVKPIVDALLAEAKKARDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWK KGADILIVDATDKNEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DVNA RIEKRRKKLAAEGRTDPAVIEAEAAKAREEG**WKQVEQLRREGATEIAYRSDDWRDLK EAWKKGADILIVDATLEHHHHHH
HalluTIM2-3	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DASRASAALQAAK NAKDPKEKEKLLKENQEKAQKIRDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWK KGADILIVDATDKNEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DADDT ADAIRKRAEAEGNKPEYEKKIDEVREKA**WKQVEQLRREGATEIAYRSDDWRDLKEAW KKGADILIVDATLEHHHHHH
HalluTIM3-1	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DASRLREAADAAR AAGEATGDEELIAKAEAYRDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWKKGADI LIV**DGLRRGRIARELERLAKEEGDPALLAAAEAAREAA**WKQVEQLRREGATRIAYRSD DWRDLKEAWKKGADILIV**DNRARLRRAEEEVAETGDPDNEELIRETRERAREEG**WKQ VEQLRREGATEIAYRSDDWRDLKEAWKKGADILIVDATLEHHHHHH
HalluTIM3-2	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**DARRKRRAADAAE ARGKATGDPEAIAVGQAYRDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWKKGADI LIV**DGVTRRGRARRLRRAAEAEGDPELLAEARALREEA**WKQVEQLRREGATRIAYRS DDWRDLKEAWKKGADILIV**DARTLLRRAREEVAAEGRPDDPELIEKTIAEAREEA**WK QVEQLRREGATEIAYRSDDWRDLKEAWKKGADILIVDATLEHHHHHH
HalluTIM3-3	MDKDEAWKQVEQLRREGATRIAYRSDDWRDLKEAWKKGADILIV**VSKAIEWRAEEAK ALAAGDKEAAAKAAAAAKQARDEA**WKQVEQLRREGATEIAYRSDDWRDLKEAWKKG ADILIV**ESGEDRTRRRAIELGLFDPNNPEVQKAREEAKQEA**WKQVEQLRREGATRIAY RSDDWRDLKEAWKKGADILIV**DAKSLEEKAEKLLKEAKKRNDPELEKKAEELKKEA** WKQVEQLRREGATEIAYRSDDWRDLKEAWKKGADILIVDATLEHHHHHH

Table S2. Parameters for the final designs. Length in number of residues of each extension in the design (from N- to C-terminus). RMSD over all Cα-atoms for the proteinMPNN input structure and the AlphaFold prediction. pLDDT for each design.

Design	Extension (# res)	RMSD (Å)	pLDDT
HalluTIM2-1	39/35	1.93	95.2
HalluTIM2-2	39/35	1.76	95.7
HalluTIM2-3	39/33	1.74	94.4
HalluTIM3-1	35/35/36	1.93	94.6
HalluTIM3-2	35/35/36	1.93	94.9
HalluTIM3-3	37/35/35	2.25	95.0

Table S3. Biochemical and thermodynamic properties. Data points for theoretical and experimentally determined molecular weight (MW) using SEC-MALS as well as apparent melting temperature (T_M) and ΔG at 25 °C using CD spectroscopy.

construct	Theoretical MW	Experimental MW [kDa]	T_M [°C] (n=2)	$\Delta G_{25\,°C}$ [kcal mol^{-1}]
sTIM11-SB	22.93	22.90 ± 0.05	65.5 ± 0.7	-7.3 ± 0.8
HalluTIM2-1	29.77	30.10 ± 0.03	65.7 ± 0.5	-10.4 ± 0.2
HalluTIM2-2	29.08	28.90 ± 0.03	71.7 ± 1.1	-12.3 ± 0.1
HalluTIM2-3	29.23	31.30 ± 0.09 / 29.50 ± 0.18	58.3 ± 1.0	-5.1 ± 0.1
HalluTIM3-1	31.96	33.00 ± 0.03	66.5 ± 0.7	-10.5 ± 0.4
HalluTIM3-2	32.00	34.60 ± 0.21	68.4 ± 0.3	-7.2 ± 0.8
HalluTIM3-3	32.16	32.00 ± 0.06	65.1 ± 0.7	-8.9 ± 0.6

Table S4. Structural comparisons. Structural alignment over all Cα-atoms each extension from the crystal structures with the ones in the corresponding ColabFold predictions. Numbering of the helices from N- to C- terminus.

	Structural alignment	**Number of Cα-atoms**
HalluTIM2-2	Extension 1: 0.44 Å	39 of 39 Cα-atoms
	Extension 2: 0.80 Å	22 of 22 Cα-atoms
HalluTIM3-1	Extension 1: 2.34 Å	35 of 35 Cα-atoms
	Extension 2: 0.65 Å	35 of 35 Cα-atoms
	Extension 3: 0.96 Å	36 of 36 Cα-atoms

Table S5. Comparison of pocket volumes in extended *de novo* TIM barrels. Pocket volumes are calculated with ChimeraX (Meng *et al.*, 2023), the ones for αTIM2 and αTIM2-2 derive from Kordes *et al.* (2023).

Construct	**Pocket volume (Å^3)**
sTIM11-SB	316
HalluTIM2-2	2000
HalluTIM3-1	1006
αTIM2	750
αTIM2-2	2127

Reference: Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science*, *32*(11), e4792. doi: 10.1002/PRO.4792

Table S6. Data collection and refinement statistics of HalluTIM2-2 and HalluTIM3-1. Statistics for the highest-resolution shell are shown in parentheses.

	HalluTIM2-2	**HalluTIM3-1**
Wavelength	0.8731	0.9762
Resolution range	47.63 - 2.55 (2.641 - 2.55)	38.23 - 2.15 (2.227 - 2.15)
Space group	P 21 2 21	P 64 2 2
Unit cell	92.71 71.2 88.64 90 90 90	122.36 122.36 165.62 90 90 120
Total reflections	80245 (8237)	1586693 (164305)
Unique reflections	19247 (1921)	40404 (3948)
Multiplicity	4.2 (4.3)	39.3 (41.6)
Completeness (%)	97.11 (96.36)	99.54 (95.95)
Mean I/sigma(I)	8.98 (0.70)	17.65 (0.46)
Wilson B-factor	91.96	67.89
R-merge	0.07381 (1.654)	0.1358 (5.774)
R-meas	0.08514 (1.89)	0.1376 (5.845)
R-pim	0.04118 (0.8896)	0.02208 (0.8997)
CC1/2	0.997 (0.302)	0.999 (0.342)
CC*	0.999 (0.681)	1 (0.714)
Reflections used in refinement	19246 (1879)	40401 (3790)
Reflections used for R-free	958 (92)	2008 (186)
R-work	0.2465 (0.3853)	0.2453 (0.4822)
R-free	0.3060 (0.4170)	0.2598 (0.4945)
CC(work)	0.960 (0.474)	0.939 (0.587)
CC(free)	0.866 (0.050)	0.936 (0.504)
Number of non-hydrogen atoms	3800	2289
macromolecules	3777	2192
ligands	15	47
solvent	8	50
Protein residues	461	262
RMS(bonds)	0.003	0.002
RMS(angles)	0.5	0.44
Ramachandran favored (%)	97.57	97.69
Ramachandran allowed (%)	1.77	1.92
Ramachandran outliers (%)	0.66	0.38
Rotamer outliers (%)	5.45	1.9
Clashscore	7.27	6.76
Average B-factor	110.94	91.96
macromolecules	110.96	91.72
ligands	112.21	106.54
solvent	102.33	89.13
Number of TLS groups	2	3

Manuscript 3

Customizing the Structure of a Minimal TIM Barrel to Craft a *De Novo* Enzyme

Julian Beck*, Benjamin J. Smith*, Niayesh Zarifi, Emily Freund, Roberto A. Chica, Birte Höcker
Manuscript under review at *Nature Chemical Biology* (Manuscript-ID: NCHEMB-A250400931-T)
Preprint available on *bioRxiv*, 2025.01.28.635154
* equal contribution

Customizing the Structure of a Minimal TIM Barrel to Craft a *De Novo* Enzyme

Julian Beck,[1,†] Benjamin J. Smith,[2,3,†] Niayesh Zarifi,[2,3] Emily Freund,[1] Roberto A. Chica,[2,3,*] Birte Höcker[1,*]

[1] Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.

[2] Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[3] Center for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[†] These authors contributed equally.

*** Corresponding authors**

Birte Höcker, E-mail: birte.hoecker@uni-bayreuth.de

Roberto A. Chica, E-mail: rchica@uottawa.ca

Keywords

De novo enzyme design, computational protein design, enzymes, TIM barrel, Kemp elimination

Abstract

The TIM barrel is the most prevalent fold in natural enzymes, supporting efficient catalysis of diverse chemical reactions. While *de novo* TIM barrels have been successfully designed, their minimalistic architectures lack structural elements essential for substrate binding and catalysis. Here, we present CANVAS, a computational workflow that introduces a structural lid into a minimal *de novo* TIM barrel to anchor catalytic residues and form an active-site pocket for enzymatic function. Starting from two *de novo* TIM barrels, we designed nine variants with distinct lids to form active sites for the Kemp elimination. Experimental testing identified one active variant with catalytic efficiency comparable to previously reported Kemp eliminases, and mutational analyses validated the essential role of the designed catalytic residues. Sequence optimization of this variant improved solubility and stability, enabling X-ray structure determination, which confirmed the designed lid structure. This study reports the first enzymatically active *de novo* TIM barrel and establishes a platform for designing enzymes from minimal protein scaffolds.

Introduction

The TIM barrel is one of the most versatile enzyme folds. It is present in six of the seven enzyme classes and supports catalysis at diffusion-limited rates[1-3]. While idealized *de novo* TIM barrels have been successfully designed[4-7], none have been converted into functional enzymes due to their minimalistic architecture, which lacks key structural elements required for catalysis, such as cavities, pockets, and extended loops. Efforts to address these deficiencies have led to the introduction of secondary structural elements to create rudimentary pockets[8-10]. However, these pockets are too solvent exposed and undefined to support the microenvironments required for catalysis. Therefore, novel strategies are needed to craft custom active sites and endow *de novo* TIM barrels with enzymatic function.

A computational approach for active-site scaffolding

Here, we introduce CANVAS (Customizing Amino-acid Networks for Virtual Active-site Scaffolding), a computational workflow for transforming minimal proteins such as *de novo* TIM barrels into functional enzymes (Figure 1a, Supplementary Figure 1). The process begins with a minimal TIM barrel template and a theozyme, which is a computational model of an idealized enzyme active site with catalytic groups arranged to stabilize a reaction's transition state[11]. Using the protein design software Triad[12], a catalytic amino acid from the theozyme is placed onto the TIM barrel template, and the transition state is built from its side chain to maintain the catalytic geometry[13]. A second catalytic residue is then built from the transition state, ensuring proper geometry for catalytic interaction. The alpha carbon of this second residue is thus positioned in the empty space above the TIM barrel catalytic face. This allows it to serve as an anchor for constructing a custom lid to form the active-site pocket, which is created using the generative AI tool RFdiffusion[14] followed by sequence design with ProteinMPNN[15] and evaluation with AlphaFold2[16]. The active site is further optimized with Triad to build a pocket complementary to the transition state[17]. Designs are then filtered using key enzyme design criteria[18], including solvent-accessible surface area of the transition state, active-site preorganization, energy, and catalytic contact geometry.

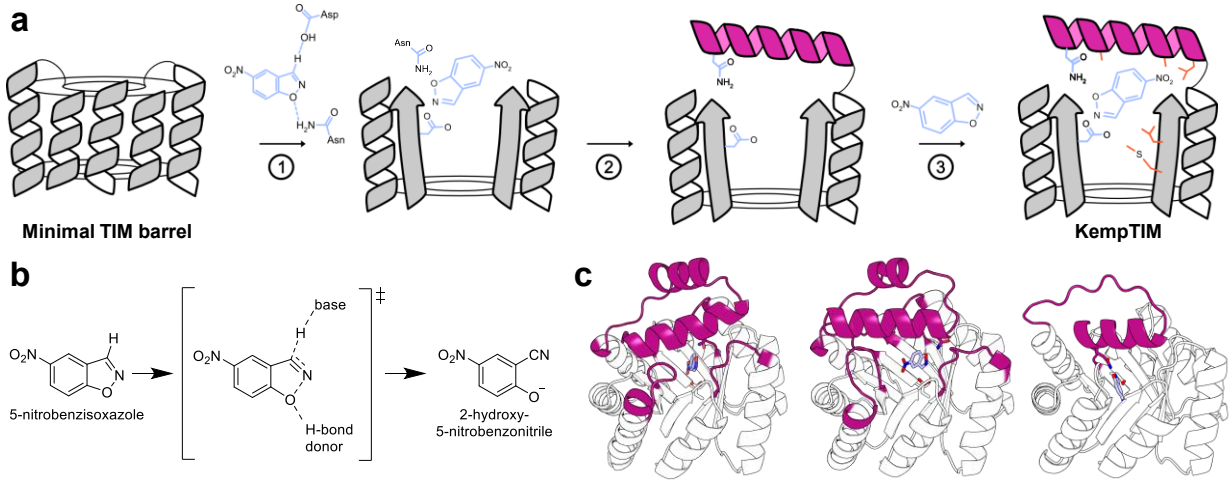


Figure 1. *De novo* enzyme design using CANVAS. (a) CANVAS follows three general steps: (1) placement of the theozyme (blue) onto the minimal TIM-barrel scaffold using a single catalytic residue as an anchor; (2) construction of a custom lid (magenta) to form the active-site pocket and anchor the second catalytic residue; and (3) active-site repacking to further stabilize the transition state. (b) Kemp elimination reaction. (c) Structure of three distinct lids (magenta) tailored to the Kemp elimination, designed using CANVAS. Transition state and designed catalytic residues are shown as sticks.

Applying CANVAS to minimal TIM-barrel proteins

We applied CANVAS to the Kemp elimination (Figure 1b), a model organic transformation commonly used as a benchmark for *de novo* enzyme design[17-19]. The reaction's theozyme features an Asp as catalytic base and an Asn or Gln as H-bond donor to stabilize negative charge buildup on the oxygen at the transition state. We used the crystal structures of two *de novo* TIM barrels (PDB IDs: 7MCD[20] and 7OSV[5]) as templates. Since 7OSV has its N- and C-termini on the catalytic face of the TIM barrel, we applied an *in silico* circular permutation based on the inpainting method[21] to generate two modified structures termed

NT6-CP1 and NT6-CP2 that contain termini on the stability face (Supplementary Figure 2). Starting from these structures, we used CANVAS to generate a distinct lid for each template tailored to the Kemp elimination (Figure 1c). The lid diffused onto 7MCD consists of a single 26-residue fragment introduced in βα loop 7 (Supplementary Figure 3). By contrast, the designed lids in NT6-CP1 and NT6-CP2 comprise a major fragment of 33 or 39 residues introduced into βα loops 2 or 6, respectively, to scaffold the catalytic H-bond donor, as well as three or four elongated loops to enhance interactions with the primary fragment. The resulting designs, termed KempTIMs, include variants 1–3 obtained from NT6-CP2, variants 4–5 from NT6-CP1 and variants 6–9 from 7MCD, each with a unique active-site configuration (Supplementary Figure 4) and sequence (Supplementary Table 1), which yielded high AlphaFold2 pLDDT scores (Supplementary Figure 5).

Kinetic assays reveal an active Kemp eliminase

Of the nine designs, seven expressed in soluble form in *E. coli* but only five could be purified to homogeneity (Supplementary Table 2). Notably, the NT6-CP1 template could not be expressed, in contrast to NT6-CP2 and 7MCD. Enzymatic assays revealed a single active design, KempTIM4, which followed Michaelis-Menten kinetics with substrate saturation (Figure 2a), yielding k_{cat}, K_M, and k_{cat}/K_M values of 0.0066 ± 0.0006 s^{-1}, 0.6 ± 0.1 mM, and 11 ± 3 M^{-1} s^{-1}, respectively. This catalytic efficiency is comparable to those of other *de novo* Kemp eliminases[17,19,22], such as KE07 (k_{cat} = 0.018 s^{-1}, K_M = 1.4 mM, and k_{cat}/K_M = 12.2 M^{-1} s^{-1})[19]. KempTIM4 expressed predominantly as a monomeric protein (Figure 3a,b, Supplementary Table 3) and displayed a mixed αβ secondary structure characteristic of TIM barrels (Figure 3c). Compared to the minimal TIM barrel NT6-CP2, KempTIM4 exhibited increased random coil content, indicated by a stronger circular dichroism (CD) signal at 208 nm. This structural feature aligns with the design of its lid, which includes long loops (Figure 3a, Supplementary Figure 3). KempTIM4 showed a strongly reduced melting temperature relative to NT6-CP2 (Figure 3d), reflecting the destabilizing effects of incorporating a lid and active site onto the minimal TIM barrel scaffold. Mutation of the designed catalytic base and H-bond donor to Ala abolished or substantially reduced activity (Figure 2b) despite size-exclusion chromatograms and CD spectra comparable to KempTIM4 (Supplementary Figure 6, Supplementary Table 3). These results confirm that the loss of activity in these knockout variants is due to mutation, not structural disruption. Moreover, the absence of enzymatic activity in NT6-CP2 demonstrates that catalysis in KempTIM4 is driven by the designed lid and active site.

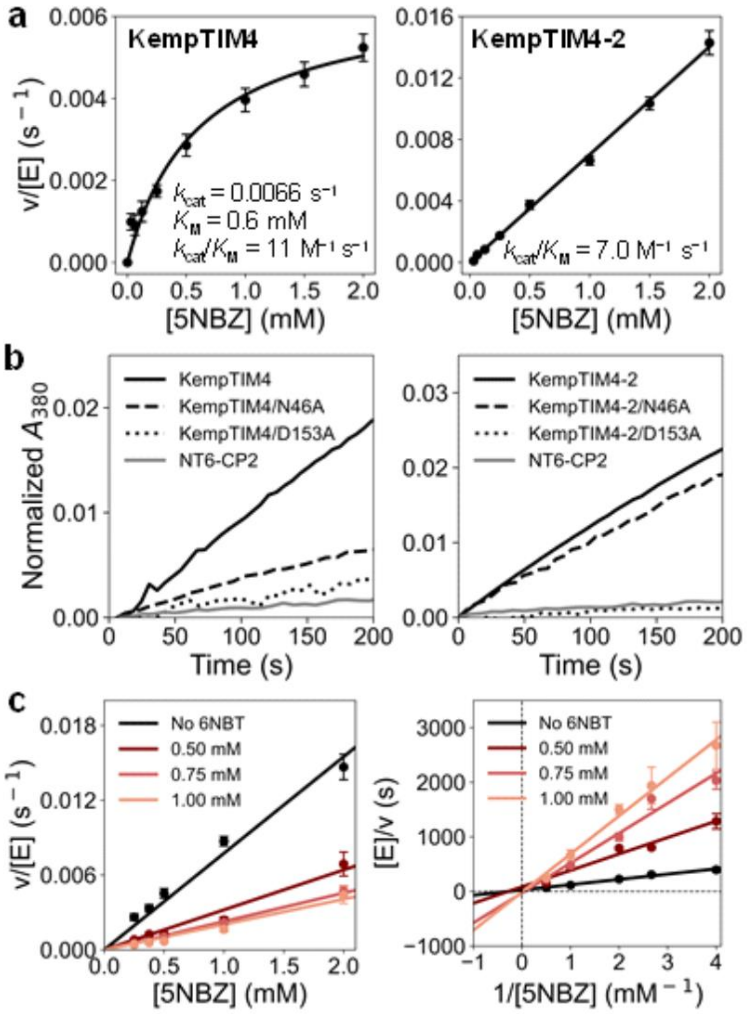


Figure 2. Kinetic characterization of KempTIMs. (a) Michaelis–Menten plots of normalized initial rates as a function of 5-nitrobenzisoxazole (5-NBZ) concentrations. Data represent the average of 18 or nine replicate measurements from five or four independent protein batches for KempTIM4 and KempTIM4-2, respectively (mean ± SEM). (b) Reaction progress curves ([5NBZ] = 2 mM) demonstrate the contributions of designed catalytic residues D153 and N46 to catalysis. (c) Michaelis-Menten (left) and Lineweaver-Burk (right) plots reveal that the transition-state analogue 6NBT acts as a competitive inhibitor of KempTIM4-2. Data represent the average of four replicate measurements from a single protein batch (mean ± SEM).

Redesign yields a more stable enzyme

Although active, KempTIM4 was prone to aggregation at high concentrations. This led us to redesign its sequence for enhanced expression and solubility using ProteinMPNN, which has been shown to produce stable and soluble proteins[23]. In this procedure, identities of active-site residues designed with Triad were preserved to maintain catalytic function, while the rest of the protein was redesigned. The redesigned protein, KempTIM4-2, demonstrated approximately 10-fold higher expression yields compared to KempTIM4 (Supplementary Table 2) and existed exclusively as a monomer with no tendency to form dimers (Figure 3b, Supplementary Table 3). CD spectroscopy showed increased α-helicity compared to KempTIM4 and higher stability, comparable to NT6-CP2 (Figure 3c,d). Despite these improvements, KempTIM4-2 exhibited lower activity than KempTIM4 (k_{cat}/K_M = 7.0 ± 0.1 M^{-1} s^{-1}) and could not be

saturated within the substrate's solubility limit (Figure 2a), even though active-site residues remained unchanged. As with KempTIM4, mutation of the catalytic base in KempTIM4-2 abolished activity (Figure 2b) without disrupting structure (Supplementary Figure 6, Supplementary Table 3). However, mutation of the H-bond donor caused only a small reduction in catalytic activity compared to KempTIM4, suggesting that this designed Asn residue adopts a suboptimal orientation hindering its catalytic function. Nevertheless, inhibition assays using the 6-nitrobenzotriazole (6NBT) transition-state analogue demonstrated competitive binding in the active site (Figure 2c), confirming the integrity of the designed active site pocket.

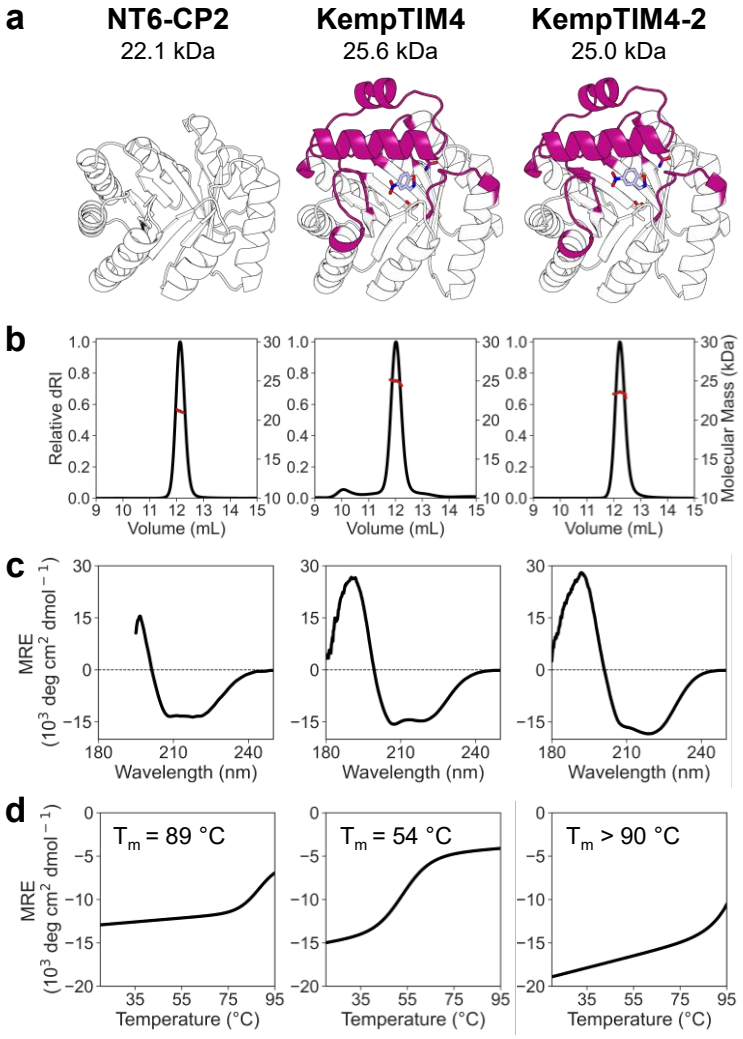


Figure 3. Structural characterization. (a) Computational models of minimal TIM-barrel NT6-CP2 and active KempTIMs with the designed lid shown in magenta and the theozyme as sticks. (b) SEC-MALS indicates that the proteins are predominantly monomeric, with molecular weights close to their expected values. (c) CD spectra reveal a mixed αβ signal characteristic of TIM barrels. MRE: mean residue ellipticity. (d) Melting curves demonstrate that KempTIM4 is substantially destabilized compared to NT6-CP2. Redesign to yield KempTIM4-2 restores stability.

The crystal structure confirms the designed lid and provides active site details

Crystals of KempTIM4-2 were obtained in the presence of 6NBT, and the structure was solved at 2.3 Å resolution (Supplementary Table 4). The structure confirmed the correct folding of the designed lid, with the primary helix-turn-helix fragment and four elongated loops closely matching the design model (Figure 4a). However, the lid was slightly displaced relative to the TIM barrel core, resulting in a 1.8-Å shift of the catalytic H-bond donor N46 Cα atom (Figure 4b), which likely reduces its contribution to catalysis (Figure 2b). While crystal contacts may contribute to this displacement, size-exclusion chromatography coupled with small-angle X-ray scattering (SEC-SAXS) confirmed that the designed lid adopts a similar conformation in solution (Supplementary Figure 7).

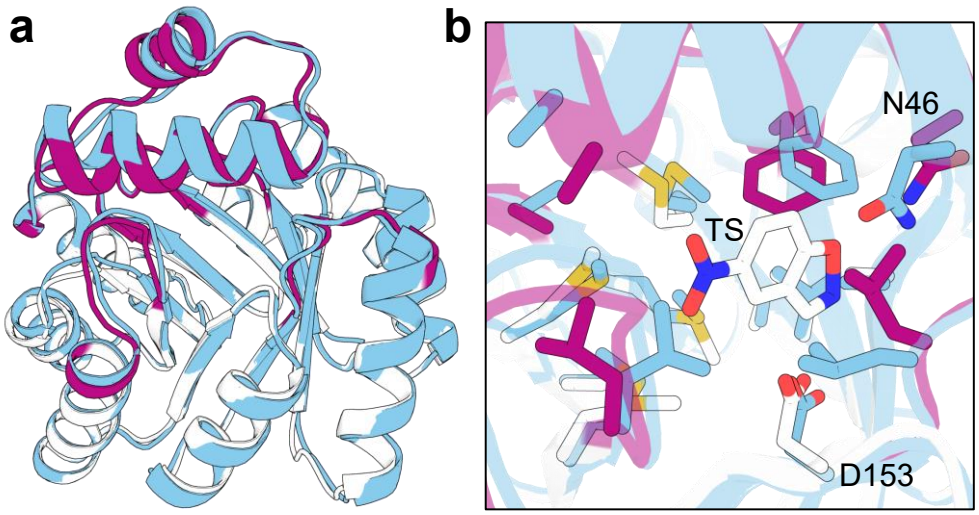


Figure 4. Crystal structure of KempTIM4-2. (a) The crystal structure (blue) aligns well with the design model (minimal TIM barrel and lid colored white and magenta, respectively). The backbone RMSD is 0.86 Å for the full structure and 1.34 Å for the lid. (b) Active-site residues in the crystal structure (blue) and design model (white and magenta) show greater divergence, possibly due to the absence of a bound transition-state analogue in the crystal structure. While the catalytic base D153 adopts a similar rotamer in both structures (side chain RMSD = 0.76 Å), the catalytic H-bond donor N46 adopts a rotameric configuration incompatible with transition-state (TS) stabilization (side chain RMSD = 3.09 Å).

Despite crystallization with 6NBT, no electron density corresponding to the transition-state analogue was observed in the active site; instead, the density was best interpreted as a bound glycerol molecule from the cryoprotectant. This suggests weak binding of the transition-state analogue under crystallization conditions (Methods). Analysis of the active site revealed that the catalytic base D153 adopts a similar rotameric configuration in both the crystal structure and design model, whereas N46 adopts a rotameric state incompatible with transition-state stabilization (Figure 4b). This aligns with the observation that mutation of N46 to Ala in KempTIM4-2 causes only a small reduction in activity, while mutation of D153 abolishes activity (Figure 2b). Additionally, the active site is poorly preorganized, with several amino acid side chains deviating from the conformations designed to stabilize the transition state. These structural discrepancies likely contribute to the modest catalytic efficiency of KempTIM4-2 and provide starting points for future optimization.

Structural Dynamics and Active-Site Accessibility Underpin Catalytic Activity

To explore why only KempTIM4 and KempTIM4-2 were catalytically active, microsecond-timescale molecular dynamics simulations were performed on these designs and the inactive KempTIM9, both in the presence and absence of the transition-state analogue. KempTIM9 was selected as a comparison due to its good expression (Supplementary Table 2), well-folded structure and high melting temperature (67°C, Supplementary Figure 8), which rule out misfolding as the cause of inactivity. The simulations revealed that the helical component of the lids, which anchor the catalytic Asn, were rigid (Figure 5a). This rigidity stabilizes the transition-state analogue within the binding pocket, maintaining catalytic interactions with the catalytic base for much of the simulation, with active variants outperforming the inactive design KempTIM9 (Figure 5b). However, only KempTIM4 consistently maintained catalytic interactions with the H-bond donor. Both KempTIM4-2 and KempTIM9 showed weaker contacts with this Asn residue, with KempTIM4-2 failing to form this interaction throughout the simulation. This result aligns with the lid displacement and unproductive rotameric state of N46 observed in the crystal structure, as well as the small impact on activity when N46 was mutated to Ala in KempTIM4-2 (Figure 2d).

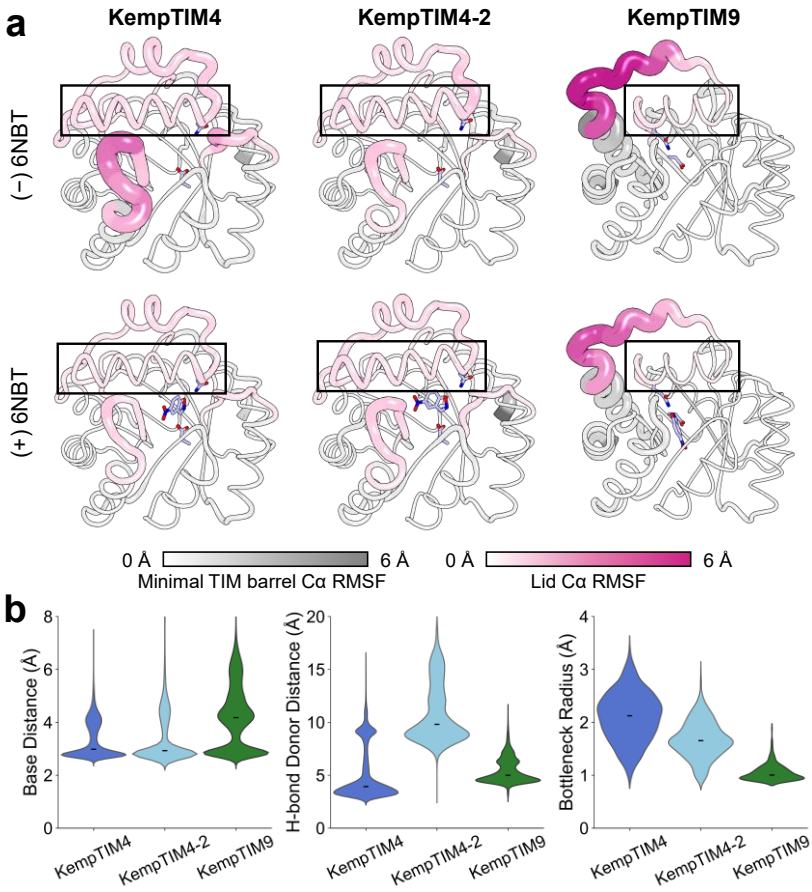


Figure 5. Molecular dynamics in the presence and absence of transition-state analogue 6NBT. (a) Root-mean-square fluctuations (RMSF) of the protein backbone, represented as putty plots, reveal that the lid α-helix anchoring the catalytic H-bond donor (boxed) is rigid, whereas loops exhibit greater flexibility. 6NBT and catalytic residues are depicted as sticks. (b) Distributions of catalytic contact distances between 6NBT and catalytic base or H-bond donor are shown alongside the active-site entrance bottleneck radius, presented as violin plots, with the median value indicated by a black line.

A key difference between the active designs and the inactive KempTIM9 was the significantly narrower entrance to the active site in KempTIM9 (Figure 5b). Previous studies have demonstrated that widening the active-site entrance enhances Kemp eliminase activity by lowering the energy barriers associated with substrate entry and product release[24]. Thus, we postulate that the inactivity of KempTIM9 is due to an insufficiently open active site, which likely impedes substrate binding. By contrast, the lids of KempTIM4 and KempTIM4-2, which incorporate five fragments rather than the single fragment found in KempTIM9 (Supplementary Figure 3), may facilitate substrate entry. The additional inserted loops in these designs likely contribute to active-site accessibility by providing greater flexibility to the lid in the unbound state (Figure 5a), allowing the active site to open more readily. While other factors, such as incorrect positioning of catalytic residues and lid misorientation, cannot be excluded, these findings suggest that reducing barriers to substrate entry during lid generation could improve catalytic performance in future designs.

Opportunities for CANVAS design

Natural TIM barrels use lids to bind substrates and form an active site conducive to catalysis[25]. By contrast, minimal *de novo* TIM barrels lack such lids and are inherently functionless. Using our CANVAS approach, we successfully designed a custom lid to form an active site for a target reaction, resulting in the first enzyme constructed from a *de novo* TIM barrel. While KempTIM4 exhibits modest activity, it matches the performance of many first-round designs generated by traditional methods[17,19], despite the added complexity of engineering a tailored lid comprising multiple loops to create an active site cavity. Further activity improvements could be achieved by using the crystal structure reported here as a template for ensemble-based design[24], a method that leverages crystallographic ensembles to design highly preorganized active sites, leading to catalytic efficiency enhancements of several orders of magnitude[18]. Unlike traditional enzyme design methods that rely on pre-existing scaffolds, CANVAS offers the flexibility to customize lids for specific reaction requirements. It can also be readily adapted to scaffold more than two catalytic residues, a task that becomes increasingly challenging with pre-existing scaffolds as the number of catalytic groups grows. With CANVAS, we anticipate that any minimal TIM barrel, including AI-generated virtual ones, could serve as a blank canvas for the creation of a large diversity of *de novo* enzymes, unlocking the full potential of this versatile fold for catalysis of a wide range of reactions.

Methods

Reagents and solutions. All experiments were conducted using analytical-grade chemicals, with solutions prepared using double-distilled water.

Circular permutation of DeNovoTIM6-SB with inpainting. Circular permutation of DeNovoTIM6-SB (PDB ID: 7OSV) was performed using the inpainting method described by Wang et al.[21] The first residue of either the first or third β-strand was selected as the new N-terminus, with the corresponding βα-loop deleted (Supplementary Figure 2). A new connection for the corresponding βα-loop was modeled by inpainting 5 to 10 residues. Unresolved residues in the crystal structure were remodeled, keeping intact the positions and identities of the remaining residues. For each new connection, 100 designs were modeled, followed by relaxation and scoring using the Rosetta protein design software[26]. Structures of each design were predicted using AlphaFold2[16] with model 4 weights. Designs with a pLDDT value above 90 and a Rosetta score below 3 Rosetta energy units per residue were kept. One design from each connection was selected for experimental characterization (Supplementary Table 1). The proteins were named NT6-CP1 (N-terminus on the first β-strand) and NT6-CP2 (N-terminus on the third β-strand).

Computational Enzyme Design. All calculations were performed with the Triad protein design software (Protabit, Pasadena, CA, USA). Rotamer optimization was performed using a Monte Carlo with simulated annealing search algorithm. Input structures were prepared for Triad calculations via the *addH.py* application within Triad. The Kemp elimination transition state (TS) structure was built using parameters from Privett *et al.*[17]. To provide sidechain conformations, a backbone-independent conformer library (bda-bbind_1.0)[13] was used for theozyme placement, and a backbone-independent rotamer library (bbind02.May.e2)[27] with expansions of ±1 standard deviation around χ_1 and χ_2 was used for active-site repacking. Energies were calculated using a modified version of the Phoenix energy function[12] consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field[28] with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal mol^{-1} and an equilibrium donor–acceptor distance of 2.8 Å,[29] and an electrostatic energy term modeled using Coulomb's law with a distance-dependent dielectric of 10. An energy benefit of −100 kcal mol^{-1} was applied when TS-side-chain interactions satisfied catalytic contact geometries (Supplementary Table 5), as described by Lassila *et al.*[13]

Theozyme placement. AlphaFold2 predictions of NT6-CP1 and NT6-CP2, along with the 7MCD crystal structure, were used as backbone templates for theozyme placement. Amino-acid residues with a C_α–C_β vector oriented towards the β-barrel interior and located at the C-terminus of each β-strand were selected as positions for introduction of the catalytic base (Asp). Neighboring residues were mutated to alanine to avoid steric clashes with the TS (Supplementary Table 6). TS poses were built using the contact geometries listed in Supplementary Table 7, and those where the isoxazolic oxygen pointed toward the catalytic face of the TIM barrel and remained coplanar with the catalytic base were selected (Supplementary Figure 1). In parallel, the catalytic H-bond donor (Asn or Gln) was placed on a Gly-Asn/Gln-Gly tripeptide scaffold, and TS poses were built using H-bond donor side-chain contacts specified in Supplementary Table 7. The

resulting TS was superimposed with the TS from the catalytic base calculation to create the full theozyme, positioning the catalytic base within the minimal TIM barrel and the H-bond donor in the empty space above the catalytic face. The Gly residues from the tripeptide were then deleted, preserving only the Asn/Gln residue, whose alpha carbon atom served as an anchor for the designed lid.

Lid design. For all theozyme placements, peptide connections between the catalytic H-bond donor residue and the minimal TIM barrel structures were designed using RFdiffusion[14]. For each position of the H-bond donor alpha carbon, various βα-loops were selected as insertion points for a primary peptide fragment, and multiple fragment lengths were sampled to connect to the catalytic residue while preserving the TIM-barrel topology. Additional βα-loops were sometimes elongated to improve interactions with the primary fragment (Supplementary Figure 3). For each inserted primary fragment, 100–200 structures were sampled. Primary fragments lacking secondary structure, causing steric clashes with the TS, or displacing the H-bond donor from its theozyme position, were discarded (Supplementary Figure 1). The filtered fragments were sequence-optimized using ProteinMPNN[15] using a temperature factor of 0.1, ensuring the minimal TIM barrel template sequence remained unchanged. Cysteine and methionine residues were excluded during sequence design. Structures of designed sequences were predicted using AlphaFold2[16] or ColabFold v1.3.0[30] with all five model weights, and those displaying an average pLDDT value >90 and backbone RMSD < 3 Å relative to the RFdiffusion model were selected for active-site repacking.

Active-site repacking. AlphaFold2 models generated above were used for a theozyme placement step, as described previously, to verify that contacts between catalytic residues and the TS could be formed on these scaffolds. Active-site repacking calculations were then performed on the structures with this new theozyme placement. In these calculations, the TS was translated by ±0.4 Å along each Cartesian coordinate in 0.2-Å steps and rotated about all three axes (origin at the TS geometric center) in 5° increments over a 10° range (clockwise and counterclockwise). This resulted in a total combinatorial search space of 15,625 possible poses. Residues near the catalytic amino acids and TS were designated as design positions. During repacking, these positions were allowed to sample rotamers of various amino acids, favoring hydrophobic and aromatic residues, while the identities of the catalytic residues remained fixed (Supplementary Table 8).

Computational library design. After repacking, the CLEARSS computational library design algorithm[31] was used to generate a combinatorial library comprising the most favorable amino acids predicted by Triad at each designed active-site position. In this method, libraries of a specific size configuration are generated from a pre-scored list of sequences using the highest probability set of amino acids at each position based on the sum of their Boltzmann weights. Using the list of energy-ranked sequences from active-site repacking as input, libraries of 192 sequences were generated. Rotamer configurations for each sequence in the library were optimized using the *cleanSequences.py* application within Triad to find the lowest-energy conformation of each sequence on its respective backbone, generating “cleaned” structures. To compare energies, the energy difference between each “cleaned” structure and a corresponding all-Gly

structure in the absence of the TS was calculated. These energies are reported in Supplementary Table 9. Structures were cleaned with and without the TS to evaluate preorganization, as described previously[18].

Design filtering. Energies of the TS-bound and unbound "cleaned" structures were calculated using Triad. Dihedral angles of catalytic residues relative to the TS, as well as the solvent-accessible surface area (SASA) of the TS, were analyzed using PyMOL (v2.3.0, Schrödinger, LLC). Bottleneck radii of active-site entrances were evaluated using CAVER v3.0[32], and the number of residues mutated to glycine or methionine was counted. Designs were filtered based on the following criteria (Supplementary Table 9): energy difference between bound and unbound structures below 0 kcal mol^{-1}, catalytic residue dihedral angles outside 50° to 130° and –50° to –130°, SASA of TS between 80 and 200 $Å^2$, < 6 Met and < 4 Gly, bottleneck radius of active-site entrance > 0.9 Å, and < 7 non-preorganized residues. Designs passing all criteria were used for structure prediction with AlphaFold2 or ColabFold (v1.3.0), where designs with a backbone RMSD < 3 Å relative to the structure cleaned by Triad, overall pLDDT > 90, and lid pLDDT > 85 were chosen for experimental characterization (Supplementary Table 1).

KempTIM4 redesign. Sequence optimization of KempTIM4 was performed using ProteinMPNN with LigandMPNN weights[33]. Amino-acid identities of all residues designed during active-site repacking of KempTIM4 were retained (Supplementary Table 8). Four sequences were generated, omitting cysteine. The generated sequences were used for structure prediction with ColabFold v1.3.0[30] using all five model weights. Predictions were filtered based on their all-atom RMSD to the input KempTIM4 AlphaFold2 model. Theozyme placement and active-site repacking for each sequence were performed as described earlier using the backbones of AlphaFold2 models that passed filtering. Following energy calculations with Triad, a single design, KempTIM4-2, was selected for experimental characterization (Supplementary Table 1).

Cloning and generation of constructs. Genes for all proteins were ordered as codon-optimized fragments flanked by restriction sites for *Nde*I and *Xho*I from Twist Bioscience (South San Francisco, CA, USA). After digestion with *Nde*I and *Xho*I, the gene fragment was ligated into pET21b(+) for KempTIM1–5, or pET29b(+) for KempTIM6–9. Catalytic knockout mutations were introduced by a modified QuikChange PCR protocol utilizing KAPA Polymerase followed by *Dpn*I digestion. *E. coli* Top10 or BL21 (DE3) cells were transformed with ligated vector or reaction mixture and plated on Lysogeny Broth (LB) agar plates containing 100 µg mL^{-1} ampicillin or 50 µg mL^{-1} kanamycin as selection markers. Single colonies were used to inoculate 5 mL LB supplemented with 100 µg mL^{-1} ampicillin or 50 µg mL^{-1} kanamycin. After overnight growth (37 °C, 250 rpm), cells were harvested by centrifugation and DNA was isolated using NucleoSpin Plasmid EasyPure-Kit (Machery & Nagel) according to the manufacturer's protocol. Vector assembly and introduction of mutations were validated by sequencing (Eurofins Genomics) using standard T7 primers.

Protein expression and purification. *E.coli* BL21 (DE3) cells (Novagen) were transformed with plasmid, plated on agar plates containing antibiotic, and incubated overnight at 37 °C. Single colonies were picked to inoculate precultures and incubated at 30–37 °C overnight. On the next day, 1 L LB supplemented with antibiotic was inoculated with 10 mL of the preculture and incubated at 37 °C until OD600 reached a value between 0.6 and 0.8. Overexpression was induced by adding isopropyl-β-thiogalactoside (IPTG) to a final concentration of 0.1 mM for KempTIM1–5 or 1 mM for KempTIM6–9 followed by overnight incubation at 16–20 °C with shaking. Cells were harvested by centrifugation and pellets were either frozen at –20 °C or used directly for purification.

Two different purification schemes were employed. In the first scheme, cell pellets were resuspended in 35 mL buffer A (50 mM sodium phosphate pH 7.0, 100 mM NaCl, 10 mM imidazole). Resuspended cells were lysed by sonication (Branson Ultrasonic Sonifier 250, output 4, duty cycle 40%, 3 × 3 min) and centrifuged. The supernatant was loaded onto a HisTrapHP column (5 mL, Cytiva Life Science) equilibrated with buffer A and coupled to an ÄKTApure system (Cytiva Life Science). After washing with 10 column volumes (CV) of buffer A, the protein was eluted with a linear gradient over 20 CV to 60 % buffer B (50 mM sodium phosphate pH 7.0, 100 mM NaCl, 500 mM imidazole). For KempTIM4-2L, which contains an additional GSG tripeptide linker between its His-tag and first TIM barrel residue (Supplementary Table 1), an additional TEV-cleavage over night was performed during a dialysis against buffer C (50 mM of sodium phosphate pH 7.0, 100 mM NaCl). The dialyzed protein was loaded onto a HisTrapHP column (5 mL, Cytiva Life Science) equilibrated with buffer C and coupled to an ÄKTApure system (Cytiva Life Science) and the flowthrough was collected. Fractions containing the protein were pooled, concentrated with a centrifugal concentrator, and loaded onto a HiLoad 26/600 Superdex 75 preparative grade column (Cytiva Life Sciences) preequilibrated in buffer C and eluted with 1 CV buffer C.

In the second scheme, cell pellets were resuspended in 8 mL buffer D (5 mM imidazole in 100 mM potassium phosphate buffer at pH 8.0) supplemented with 1 mg mL^{-1} lyophilized lysozyme (MP Biomedicals) and 1 U mL^{-1} benzonase nuclease (Merck Millipore). Cells were homogenized using an Avestin EmulsiFlex-B15 cell disruptor. Proteins were purified from the lysate by immobilized metal affinity chromatography using Ni-NTA agarose (Qiagen) pre-equilibrated with buffer D in individual Econo-Pac gravity-flow columns (Bio-Rad). Contaminants were washed away using buffer E (10 mM imidazole in 100 mM potassium phosphate buffer pH 8.0) and then 20 mM imidazole in the same buffer. Proteins were eluted with 5 mL of buffer F (250 mM imidazole in 100 mM potassium phosphate buffer pH 8.0). Proteins were further subjected to gel filtration in buffer G (25 mM HEPES pH 7.25 and 100 mM sodium chloride) using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad). Purification was checked by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and fractions containing the protein of interest were pooled. Protein concentration was determined spectrophotometrically using the absorption at 280 nm and applying Beer-Lambert's law using calculated extinction coefficients obtained from the ExPAsy ProtParam tool (https://web.expasy.org/protparam/).

Far-UV circular dichroism (CD) spectroscopy. CD spectra were collected with a Jasco J-710 or J-815 spectrometer. Spectra were recorded using a 1-mm quartz cuvette at 20 °C with scanning speed 10–100 nm min^{-1}, bandwidth 1 nm, response time 1 s. Experiments were performed in 20 mM sodium phosphate buffer pH 7.0 or 50 mM sodium phosphate buffer pH 8.0 supplemented with 100 mM sodium chloride, using a protein concentration of approximately 0.2 mg mL^{-1}. For each protein, 3 to 10 spectra were accumulated and averaged. For data normalization, a buffer spectrum was subtracted and the signal was converted to mean residue molar ellipticity using $[\theta_{MRE}] = \theta/(n \times d \times c)$, where n is the number of residues in the protein, θ the collected ellipticity in mdeg, d the path length in mm, and c the protein concentration in M. Thermal denaturation assays were performed by heating the samples from 20 to 95 °C at a rate of 1 °C per minute and ellipticity at 222 nm was measured every 1 °C. Melting temperature (T_m) was determined following the protocol by Greenfield[34].

Size-exclusion chromatography-multi angle light scattering. SEC-MALS measurements were performed as described in Beck *et al*.[8] Each protein was measured at 2 mg mL^{-1} except KempTIM4 N46A (1 mg mL^{-1}). For all measurements, buffer C and an injection volume of 50 µL were used.

Steady-state kinetics. All assays were performed in buffer C or G at 27 °C in a Spark (Tecan) or Synergy H1 (Biotek) plate reader. Enzyme concentrations varied from 5 to 80 µM. Reactions (200 µL final volume) were initiated by addition of varying concentrations (0.03–2 mM) of 5-nitrobenzisoxazole (abcr or AAblocks) dissolved in methanol (final methanol concentration 10 %). Product formation was monitored at 380 nm (ε = 15,800 M^{-1} cm^{-1})[17] in individual wells of a 96-well plate (Nunc or Greiner Bio-One). Path lengths for each well were calculated ratiometrically using the difference in absorbance at 900 and 975 nm. Control measurements without protein were conducted for each substrate concentration and subtracted from the enzyme measurements. Linear phases of the kinetic traces were used to measure initial reaction rates. If the enzyme showed saturation within the substrate concentration used, the data were fitted to the Michealis-Menten model using ($v_0 = (v_{max} \times [S]) / (K_M + [S])$ where v_0 is the initial velocity rate, v_{max} is the maximal velocity rate, [S] is the substrate concentration, K_M is the Michaelis constant. The catalytic constant k_{cat} was calculated with $k_{cat} = v_{max} / [E]$ where [E] is the enzyme concentration. If saturation was not achieved within the substrate's solubility limit, a linear equation ($v_0 = k_{cat}/K_M \times [S]$) was used.

Crystallization and structure determination. KempTIM4-2L (Supplementary Table 1) was purified using the first purification scheme, with the final size-exclusion chromatography step performed in 20 mM HEPES buffer pH 8.0 supplemented with 20 mM NaCl. Crystallization screens were set up using the sitting-drop vapor diffusion method with a Phoenix pipetting robot (Art Robbins Instruments) and commercially available sparse-matrix screens (NeXtal) in 96-well sitting-drop plates (3-drop Intelli-Plates, Art Robbins Instruments). Protein and reservoir solutions were mixed in ratios of 1:1, 2:1, and 1:2. The protein concentration was 12 mg mL^{-1}, supplemented with 5 mM 6NBT (5% DMSO). Crystals were grown at 293 K, and diffracting crystals were obtained after 40 days in 4 M sodium formate. Cryoprotection was achieved by adding glycerol to a final concentration of 25%, and crystals were mounted using cryo-loops

on SPINE standard bases and flash-cooled in liquid nitrogen. Diffraction data were collected at 100 K at beamline BL 14.1 with a Pilatus3 S 6M detector at the BESSY II synchrotron (Helmholtz-Zentrum Berlin). Data processing was performed using X-ray Detector Software APP3 (XDSAPP3)[35] with XDS[36], and data quality was assessed using phenix.xtriage[37]. Phases were solved by molecular replacement using the AlphaFold2 prediction as search model with Phaser[38]. The resulting model was manually rebuilt using Coot[39] and refined with phenix.refine[37] in an iterative process. Refinement statistics and crystallographic data are shown in Supplementary Table 4. Coordinates and structure factors were validated and deposited in the PDB under accession code 9QKX.

Size-Exclusion Chromatography Small-angle X-ray Scattering (SEC-SAXS). SEC-SAXS measurements were conducted at the BioSAXS beamline BM29 of the ESRF in Grenoble, France. A Superdex 75 Increase 10/300 GL column (Cytiva Life Sciences) was used with a flow rate of 0.8 mL min^{-1} in buffer C. KempTIM4-2L (Supplementary Table 1) was measured at a concentration of 5 mg mL^{-1}, with an injection volume of 100 µL. Data processing and analysis were performed using ATSAS 3.2.1[40,41]. For analysis, AlphaFold2 models of KempTIM4-2 and NT6-CP2, as well as an AlphaFlow[42] model of KempTIM4-2 with a displaced lid obtained using the base molecular dynamics weights were used. For each of these models a theoretical scattering curve were generated and fitted to the experimental data using CRYSOL[43].

Molecular dynamics (MD) simulations. The Amber 2020 software (http://ambermd.org/) with the ff19SB protein force field[44], gaff2 ligand force field[45], and OPC water force field[46] was used for all simulations. A cutoff of 10 Å was applied to electrostatics modeled using the particle mesh Ewald method[47]. All MD trajectories were run with a time step of 2 fs. Models of each design generated using Triad as described above were prepared for MD. MD trajectories were run for each enzyme in the presence and absence of bound transition-state analogue 6-nitrobenzotriazole (6NBT). Parameters for 6NBT were generated using the Antechamber package[48]. Hydrogen atoms were added using Reduce[49], and the prepared structures were solvated with OPC water in a truncated octahedral box with periodic boundary conditions where the distance between the protein surface and the box edges was set to 10 Å. The addions2 algorithm in Amber was used to place counterions to neutralize the system. The system was first minimized through the method of steepest descent using a force constant of 500 kcal mol^{-1} $Å^{-2}$. The system was then heated to a temperature of 300 K over 240 ps while restraints were gradually removed, followed by equilibration under an NPT ensemble at 300 K and 1 bar for 10 ns and subsequent equilibration under an NVT ensemble at 300 K for another 10 ns. Constant temperature and pressure were achieved using the Langevin thermostat[50] and Berendsen barostat[51], respectively. Following equilibration, triplicate 1-µs production simulations were run. For analyzing bottleneck radii, 1000 snapshots separated by 1 ns each were extracted from the production trajectory. A minimum cutoff of 0.9 Å was used to identify active-site entrance bottlenecks in the unbound MD trajectories using CAVER 3.0[32]. To analyze catalytic interactions and RMSF, 5×10^4 snapshots separated by 20 ps each were extracted from the production trajectory. Catalytic interactions and RMSF values were extracted using CPPTRAJ and PYTRAJ, respectively[52].

Acknowledgments

J.B. and B.H. acknowledge support from the Elite Network of Bavaria and its study program "Biological Physics". B.H. acknowledges support from European Union's Horizon 2020 research and innovation program under grant agreement No 951375 (ArtMotor). R.A.C. acknowledges grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-03484 and RGPAS-2021-00017) and the Canada Foundation for Innovation (26503). B.J.S. was supported by an NSERC CREATE scholarship. This research was enabled in part by support provided by Compute Ontario (www.computeontario.ca) and the Digital Research Alliance of Canada (alliancecan.ca). We acknowledge financial support and allocation of synchrotron beamtime by HZB and thank the beamline staff at BESSY for support. We thank Sabrina Wischt for technical support, Janosch Hennig for help with SAXS data analysis and the University of Bayreuth Centre of International Excellence "Alexander von Humboldt" for facilitating this collaboration through a short-term grant to R.A.C.

Competing interests

The authors declare no competing financial interest.

ORCID identifiers

Julian Beck: 0009-0007-3555-9890

Roberto A. Chica: 0000-0003-3789-9841

Emily Freund: 0009-0006-2876-3319

Birte Höcker: 0000-0002-8250-9462

Benjamin J. Smith: 0009-0003-5953-4998

Niayesh Zarifi: 0000-0003-4748-7082

References

1 Nagano, N., Orengo, C. A. & Thornton, J. M. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *Journal of Molecular Biology* **321**, 741-765, doi:https://doi.org/10.1016/S0022-2836(02)00649-6 (2002).

2 Höcker, B., Jürgens, C., Wilmanns, M. & Sterner, R. Stability, catalytic versatility and evolution of the (βα)8-barrel fold. *Current Opinion in Biotechnology* **12**, 376-381, doi:https://doi.org/10.1016/S0958-1669(00)00230-5 (2001).

3 Wierenga, R. K., Kapetaniou, E. G. & Venkatesan, R. Triosephosphate isomerase: a highly evolved biocatalyst. *Cellular and Molecular Life Sciences* **67**, 3961-3982, doi:10.1007/s00018-010-0473-9 (2010).

4 Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chemical Biology* **12**, 29-34, doi:10.1038/nchembio.1966 (2016).

5 Kordes, S., Romero-Romero, S., Lutz, L. & Höcker, B. A newly introduced salt bridge cluster improves structural and biophysical properties of de novo TIM barrels. *Protein Science* **31**, 513-527, doi:https://doi.org/10.1002/pro.4249 (2022).

6 Romero-Romero, S. *et al.* The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach. *Journal of Molecular Biology* **433**, 167153, doi:https://doi.org/10.1016/j.jmb.2021.167153 (2021).

7 Chu, A. E., Fernandez, D., Liu, J., Eguchi, R. R. & Huang, P.-S. De Novo Design of a Highly Stable Ovoid TIM Barrel: Unlocking Pocket Shape towards Functional Design. *BioDesign Research* **2022**, 9842315, doi:https://doi.org/10.34133/2022/9842315 (2022).

8 Beck, J., Shanmugaratnam, S. & Höcker, B. Diversifying de novo TIM barrels by hallucination. *Protein Science* **33**, e5001, doi:https://doi.org/10.1002/pro.5001 (2024).

9 Kordes, S., Beck, J., Shanmugaratnam, S., Flecks, M. & Höcker, B. Physics-based approach to extend a de novo TIM barrel with rationally designed helix-loop-helix motifs. *Protein Engineering, Design and Selection* **36**, gzad012, doi:10.1093/protein/gzad012 (2023).

10 Wiese, J. G., Shanmugaratnam, S. & Höcker, B. Extension of a de novo TIM barrel with a rationally designed secondary structure element. *Protein Science* **30**, 982-989, doi:https://doi.org/10.1002/pro.4064 (2021).

11 Tantillo, D. J., Jiangang, C. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Current Opinion in Chemical Biology* **2**, 743-750, doi:https://doi.org/10.1016/S1367-5931(98)80112-9 (1998).

12 Lee, F. S., Anderson, A. G. & Olafson, B. D. Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Engineering, Design and Selection* **36**, gzad013, doi:10.1093/protein/gzad013 (2023).

13 Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 16710-16715, doi:10.1073/pnas.0607691103 (2006).

14 Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100, doi:10.1038/s41586-023-06415-8 (2023).

15 Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56, doi:10.1126/science.add2187 (2022).

16 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

17 Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790-3795, doi:10.1073/pnas.1118082108 (2012).

18 Rakotoharisoa, R. V. *et al.* Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling. *J Am Chem Soc* **146**, 10001-10013, doi:10.1021/jacs.4c00677 (2024).

19 Rothlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-195 (2008).

20 Anand, N. *et al.* Protein sequence design with a learned potential. *Nature Communications* **13**, 746, doi:10.1038/s41467-022-28313-9 (2022).

21 Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387-394, doi:10.1126/science.abn2100 (2022).
22 Zhang, S., Zhang, J., Luo, W., Wang, P. & Zhu, Y. A preorganization oriented computational method for de novo design of Kemp elimination enzymes. *Enzyme and Microbial Technology* **160**, 110093, doi:https://doi.org/10.1016/j.enzmictec.2022.110093 (2022).
23 Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *Journal of the American Chemical Society* **146**, 2054-2061, doi:10.1021/jacs.3c10941 (2024).
24 Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nature Communications* **11**, 4808, doi:10.1038/s41467-020-18619-x (2020).
25 Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. *Chemical Reviews* **105**, 4038-4055, doi:10.1021/cr030191z (2005).
26 Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, doi:10.1016/B978-0-12-381270-4.00019-6 (2011).
27 Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Current Opinion in Structural Biology* **12**, 431-440, doi:https://doi.org/10.1016/S0959-440X(02)00344-5 (2002).
28 Mayo, S. L., Olafson, B. D. & Goddard, W. A. Dreiding - a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **94**, 8897-8909, doi:Doi 10.1021/J100389a010 (1990).
29 Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10172-10177 (1997).
30 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682, doi:10.1038/s41592-022-01488-1 (2022).
31 Allen, B. D., Nisthal, A. & Mayo, S. L. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 19838-19843, doi:10.1073/pnas.1012985107 (2010).
32 Chovancova, E. *et al.* CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLOS Computational Biology* **8**, e1002708, doi:10.1371/journal.pcbi.1002708 (2012).
33 Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv*, 2023.2012.2022.573103, doi:10.1101/2023.12.22.573103 (2023).
34 Greenfield, N. J. Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. *Nature Protocols* **1**, 2733-2741, doi:10.1038/nprot.2006.229 (2006).
35 Sparta, K. M., Krug, M., Heinemann, U., Mueller, U. & Weiss, M. S. XDSAPP2.0. *Journal of applied crystallography* **49**, 1085-1092, doi:doi:10.1107/S1600576716004416 (2016).
36 Kabsch, W. XDS. *Acta Crystallographica Section D* **66**, 125-132, doi:doi:10.1107/S0907444909047337 (2010).
37 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta crystallographica. Section D, Structural biology* **75**, 861-877, doi:10.1107/S2059798319011471 (2019).
38 McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of applied crystallography* **40**, 658-674, doi:10.1107/S0021889807021206 (2007).
39 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).
40 Hopkins, J. B., Gillilan, R. E. & Skou, S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *Journal of applied crystallography* **50**, 1545-1553, doi:doi:10.1107/S1600576717011438 (2017).
41 Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *Journal of applied crystallography* **54**, 343-355, doi:doi:10.1107/S1600576720013412 (2021).
42 Jing, B., Berger, B., Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. *arXiv* **2402**, 04845 (2024).
43 Franke, D. *et al.* ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of applied crystallography* **50**, 1212-1225, doi:doi:10.1107/S1600576717007786 (2017).

44 Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950-1958, doi:https://doi.org/10.1002/prot.22711 (2010).
45 He, X., Man, V. H., Yang, W., Lee, T.-S. & Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics* **153**, 114502, doi:10.1063/5.0019056 (2020).
46 Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters* **5**, 3863-3871, doi:10.1021/jz501780a (2014).
47 Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092, doi:10.1063/1.464397 (1993).
48 Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **25**, 247-260, doi:https://doi.org/10.1016/j.jmgm.2005.12.005 (2006).
49 Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* **285**, 1735-1747, doi:10.1006/jmbi.1998.2401 (1999).
50 Izaguirre, J. A., Catarello, D. P., Wozniak, J. M. & Skeel, R. D. Langevin stabilization of molecular dynamics. *The Journal of Chemical Physics* **114**, 2090-2098, doi:10.1063/1.1332996 (2001).
51 Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684-3690, doi:10.1063/1.448118 (1984).
52 Roe, D. R. & Cheatham, T. E., III. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **9**, 3084-3095, doi:10.1021/ct400341p (2013).

Supplementary Materials for

Customizing the Structure of a Minimal TIM Barrel to Craft a *De Novo* Enzyme

Julian Beck,[1,†] Benjamin J. Smith,[2,3,†] Niayesh Zarifi,[2,3] Emily Freund,[1] Roberto A. Chica,[2,3,*] Birte Höcker[1,*]

[1] Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.

[2] Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[3] Center for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[†] These authors contributed equally.

*** Corresponding authors**

Birte Höcker, E-mail: birte.hoecker@uni-bayreuth.de

Roberto A. Chica, E-mail: rchica@uottawa.ca

This file includes:

Tables S1–S8

Figures S1–S8

Table S1. Amino-acid sequences

Enzyme	# residues	MW [a] (kDa)	Sequence [b]
NT6-CP1	188	22.1	RIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLGAKEIAYRSDDWRDLQEALKKGAD ILIVDATDPEEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRR LGAKEIAYRSDDWRDLQEALKKGADILIVDASDAERALKQVEILRRLEHHHHHH
NT6-CP2	188	22.1	EIAYRSDDWRDLQEALKKGADILIVDATDPEEAWKQVEILRRLGAKRIAYRSDDWRDLQEALKKGAD ILIVDATDKDEAWKQVEILRRLGAKEIAYRSDDWRDLQEALKKGADILIVDADDPEEAEKQVEILRR LGAKRIAYRSDDWRDLQEALKKGADILIVDATDKDEAWKQVEILRRLEHHHHHH
7MCD	191	21.4	DILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAEELKKAEKAGADILIVNPDDFEKGVEEVKELK RHGAKIIAYISKSAEELKKAEKAGADILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAEELKKAE KAGADILIVNPDDFEKGVEEVKELKRHGAKIIAYISKSAEELKKAEKAGLEHHHHHH
KempTIM1	225	26.3	EIAYASDDWRDLQEALKKGADILD**VVNVNP**EEAWKQVEILRRLGAKRILYASLRWRDLQEALKKGAD ILA**VPLVDK**DEAWKQVEILRRLGAKEIAYGSDDWRDLQEALKKGADILGV**ASNTPRVRRLVLR**N**LET LYSREEAERIVEEKIKLNDPE**EAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADILLV**AADDES EESK**DEAWKQVEILRRLEHHHHHH
KempTIM2	225	26.3	EIAYASDDWRDLQEALKKGADILD**VVNVNP**EEAWKQVEILRRLGAKRIAYASLRWRDLQEALKKGAD ILA**VPLVDK**DEAWKQVEILRRLGAKEIAYGSDDWRDLQEALKKGADILGV**ASNTPRVRRLVLR**N**LET LYSREEAERIVEEKIKLNDPE**EAEKQVEILRRLGAKRIMYASDDWRDLQEALKKGADILLV**VADDES EESK**DEAWKQVEILRRLEHHHHHH
KempTIM3	227	25.6	RIAYA**SLASRW**RDLQEALKKGADILMV**ALARDAASVRALAEGAAG**N**GDPRSVAELLALLRPTDKD**EA WKQVEILRRLGAKEIAYMSDDWRDLQEALKKGADILMV**LASSALSPEE**AWKQVEILRRLGAKRIAYG SGDWRDLQEALKKGADILD**VLGAKGKD**EAWKQVEILRRLGAKEIAYASFDWRDLQEALKKGADILMV **FAVLDA**ERALKQVEILRRLEHHHHHH
KempTIM4	227	25.6	RIAYM**SLASRW**RDLQEALKKGADILMV**ALARDAASVRALAEGAAG**N**GDPRSVAELLALLRPTDKD**EA WKQVEILRRLGAKEIAYMSDDWRDLQEALKKGADILMV**LASSALSPEE**AWKQVEILRRLGAKRIAYG SGDWRDLQEALKKGADILD**VLGAKGKD**EAWKQVEILRRLGAKEIAYASFDWRDLQEALKKGADILVV **FAVLDA**ERALKQVEILRRLEHHHHHH
KempTIM5	227	25.6	RIAYA**SLASRW**RDLQEALKKGADILMV**ALARDAASVRALAEAAAG**N**GDPRSVAELLALLRPTDKD**EA WKQVEILRRLGAKEIAYMSDDWRDLQEALKKGADILMV**LASSALSPEE**AWKQVEILRRLGAKRIAYG SGDWRDLQEALKKGADILD**VLGAKGKD**EAWKQVEILRRLGAKEIAYASFDWRDLQEALKKGADILMV **FAVLDA**ERALKQVEILRRLEHHHHHH
KempTIM6	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEKAGADILLVAPDDFEKGVEEVKELK RHGAKIIAYMSKSAEELKKAEKAGADILMVYPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAE KAGADILD**V**A**NAE**Q**EKIASKFLGRKTKVKIEENDF**EKGVEEVKELKRHGAKIIAYGSKSAEELKKAE KAGWHHHHHH
KempTIM7	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEKAGADILLVAPDDFEKGVEEVKELK RHGAKIIAYMSKSAEELKKAEKAGADILMVFPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAE KAGADILD**V**A**NAE**Q**EKIASKFMGRKTKVKIEENDF**EKGVEEVKELKRHGAKIIAYGSKSAEELKKAE KAGWHHHHHH
KempTIM8	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEKAGADILMVAPDDFEKGVEEVKELK RHGAKIIAYMSKSAEELKKAEKAGADILAVFPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAE KAGADILD**V**A**NAE**Q**EKIASKLMGRKTKVKIEENDF**EKGVEEVKELKRHGAKIIAYGSKSAEELKKAE KAGWHHHHHH
KempTIM9	211	23.5	DILGVAPDDFEKGVEEVKELKRHGAKIIGYMSKSAEELKKAEKAGADILLVAPDDFEKGVEEVKELK RHGAKIIAYMSKSAEELKKAEKAGADILMVYPDDFEKGVEEVKELKRHGAKIIAYLSKSAEELKKAE KAGADILD**V**A**NAE**Q**EKIASKLLGRKTKVKIEENDF**EKGVEEVKELKRHGAKIIAYGSKSAEELKKAE KAGWHHHHHH
KempTIM4-2	233	25.0	HHHHHHENLYFQSRIAYM**ALASDL**DSLVEALKLGADILMV**ALMADAAAVRAMAEGLHA**N**GDPRSVAE LEALLRPTDLD**DALAAVRELKALGAKEIAFMSHDVDHLIRAMEAGADILMV**LESSATSVEA**ALAQVR RLKAAGAKRISFGSGDVAHLKAAMEAGADILD**VLERHGLD**VALAQIRELKAAGAKEIAFASLDPDHL LRAREEGADILVV**FGATDP**ARALATVRYLRAW
KempTIM4-2L	236	25.1	HHHHHHENLYFQSGSGRIAYM**ALASD**LDSLVEALKLGADILMV**ALMADAAAVRAMAEGLHA**N**GDPRS VAELEALLRPTDLD**DALAAVRELKALGAKEIAFMSHDVDHLIRAMEAGADILMV**LESSATSVEA**ALA QVRRLKAAGAKRISFGSGDVAHLKAAMEAGADILD**VLERHGLD**VALAQIRELKAAGAKEIAFASLDP DHLLRAREEGADILVV**FGATDP**ARALATVRYLRAW

[a] Molecular weight

[b] Catalytic residues and residues generated by RFdiffusion are underlined and bolded, respectively.

Table S2. Purification yields

Enzyme	Yield (mg L^{-1}) [a]
NT6-CP1	–
NT6-CP2	70 ± 10
7MCD	31 ± 3
KempTIM1	–
KempTIM2	–
KempTIM3	–
KempTIM4	3 ± 1
KempTIM5	–
KempTIM6	11 ± 2
KempTIM7	9 ± 5
KempTIM8	7 ± 4
KempTIM9	16 ± 7
KempTIM4-2	40 ± 20

[a] Data represent the average of 6 to 51 individual replicate measurements from 2 to 17 independent protein batches, with error bars indicating the standard deviation. NT6-CP1, KempTIM1–3 and KempTIM5 could not be expressed or purified to homogeneity.

Table S3. Theoretical and experimentally determined molecular weight (MW) using SEC-MALS

Protein	Theoretical MW (kDa)	Experimental MW (kDa)
NT6-CP2	22.1	21.2 ± 0.02
KempTIM4	25.6	24.9 ± 0.12 / 50.3 ± 1.36
KempTIM4 D153A	25.6	24.7 ± 0.10
KempTIM4 N46A	25.6	25.7 ± 0.05
KempTIM4-2	25.0	23.5 ± 0.02
KempTIM4-2 D153A	25.0	23.5 ± 0.02
KempTIM4-2 N46A	25.0	23.5 ± 0.02

Table S4. Geometric constraints used to define catalytic contacts during computational design

Contact	Residue	Type	Atom 1 [a]	Atom 2 [a]	Atom 3 [a]	Atom 4 [a]	Min [b]	Max [b]
Base	**Asp**	Distance	OD1 or OD2	**H3**			1.0	1.6
		Angle	CG	OD1 or OD2	**H3**		109	131
		Angle	OD1 or OD2	**H3**	**C3**		159	180
		Torsion	CB	CG	OD1 or OD2	**H3**	–21, 159	21, 201
H-bond donor	**Asn**	Distance	1HD2 or 2HD2	**O1**			1.2	2.3
		Angle	ND2	1HD2 or 2HD2	**O1**		145	157
		Angle	1HD2 or 2HD2	**O1**	**N2**		120	140
		Torsion	1HD2 or 2HD2	**O1**	**N2**	**C3**	–20, 160	20, 200
	Gln	Distance	1HE2 or 2HE2	**O1**			1.2	2.3
		Angle	NE2	1HE2 or 2HE2	**O1**		145	157
		Angle	1HE2 or 2HE2	**O1**	**N2**		120	140
		Torsion	1HE2 or 2HE2	**O1**	**N2**	**C3**	–20, 160	20, 200

[a] Atoms in bold are from the transition state. All other atoms are from the catalytic residues.
[b] Distance measurements given in Å, all others in degrees.

Table S5. Amino-acid positions for theozyme placement

Input structure	Residues for catalytic base placement [a]	Residues mutated to alanine [a]
7MCD	I4, I31, I50, I77, I96, I123, I142, I169	N6, I31, N52, I77, N98, I123, N144, I169
NT6-CP1	R5, I24, R51, I70, R97, I116, R143, I162	A3, R5, I24, D26, A49, R51, I70, D72, A95, R97, I116, D118, A141, R143, I162, D164
NT6-CP2	R5, I24, R51, I70, R97, I116, R143, I162	A3, R5, I24, D26, A49, R51, I70, D72, A95, R97, I116, D118, A141, R143, I162, D164

[a] Residue numbering based on input structure

Table S6. Geometric definitions for generation of transition-state poses off the side chain of catalytic base and H-bond donor

Catalytic residue	Type	Atom 1 [a]	Atom 2 [a]	Atom 3 [a]	Atom 4 [a]	Values [b]
Asp	Distance	OD1 or OD2	**H3**			1.0, 1.2, 1.5
	Angle	CG	OD1 or OD2	**H3**		112, 117, 122
	Angle	OD1 or OD2	**H3**	**C3**		159, 164, 169, 174, 179
	Torsion	CB	CG	OD1 or OD2	**H3**	0, 5, 10, 170, 175, 180
	Torsion	CG	OD1 or OD2	**H3**	**C3**	170, 175, 180, 185, 190
	Torsion	OD1 or OD2	**H3**	**C3**	**N2**	0, 5, 170, 175, 180
Asn	Distance	1HD2 or 2HD2	**O1**			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	1HD2 or 2HD2	ND2	**O1**		145, 151, 157
	Angle	1HD2 or 1HD2	**O1**	**N2**		120, 135, 140
	Torsion	CG	ND2	1HD1 or 2HD2	**O1**	120, 135, 140
	Torsion	ND2	1HD2 or 2HD2	**O1**	**N2**	–121, –116, –111
	Torsion	1HE2 or 2HE2	**O1**	**N2**	**C3**	160, 180, 200
Gln	Distance	1HE2 or 2HE2	**O1**			1.2, 1.5, 1.8, 2.1, 2.3
	Angle	1HE2 or 2HE2	NE2	**O1**		145, 151, 157
	Angle	1HE2 or 2HE2	**O1**	**N2**		120, 135, 140
	Torsion	CD	NE2	1HE2 or 2HE2	**O1**	120, 135, 140
	Torsion	NE2	1HE2 or 2HE2	**O1**	**N2**	–121, –116, –111
	Torsion	1HE2 or 2HE2	**O1**	**N2**	**C3**	160, 180, 200

[a] Atoms in bold are from the transition state. All other atoms are from the catalytic residue.
[b] Distance measurements given in Å, all others are in degrees.

Table S7. Amino-acid positions optimized during active-site repacking of KempTIM4.

Designed positions	Allowed amino acids [a]
3	**ALA**, ILE, LEU, VAL, MET
5	ALA, ILE, LEU, VAL, **MET**
26	ALA, ILE, LEU, VAL, **MET**
28	**ALA**, ILE, LEU, VAL, MET
38	MET, LEU, TYR, PHE, **ALA**, VAL, ILE
42	**GLY**
45	**GLY**
83	**ALA**, ILE, LEU, VAL, MET
85	ALA, ILE, LEU, VAL, **MET**
104	ALA, ILE, LEU, VAL, **MET**
106	ALA, **LEU**, ILE
134	**GLY**
155	ALA, **LEU**, ILE
156	**GLY**
181	**ALA**, LEU, ILE
200	ALA, ILE, LEU, **VAL**, MET
202	ALA, ILE, LEU, VAL, MET, (**PHE**)

[a] Residues in bold are mutations found at those positions in the KempTIM4 sequence. Residues in parentheses were introduced during library design.

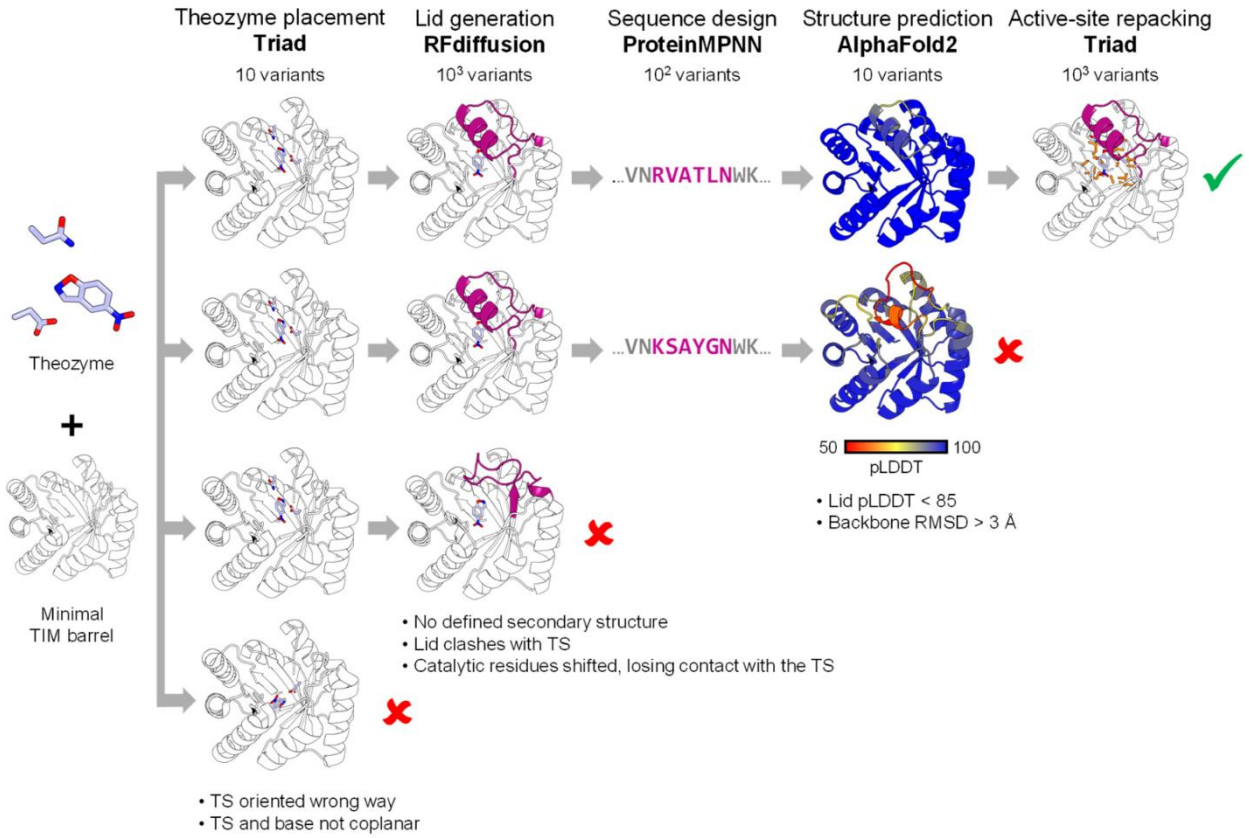
Table S8. Design filtering

Enzyme	Base dihedral (°)	H-bond dihedral (°)	Energy (kcal mol^{-1}) [a]	SASA (Å^2) [b]	# residues not preorganized [c]	Tunnel present [d]
KempTIM1	–42.5	25.7	–39.4	174	5	Yes
KempTIM2	–7.3	25.7	–36.3	188	6	Yes
KempTIM3	27.6	–156.2	–10.4	107	6	Yes
KempTIM4	27.6	–156.2	–2.3	107	5	Yes
KempTIM5	27.6	–156.2	–10.4	107	6	Yes
KempTIM6	–1	28.8	–115.54	87	2	Yes
KempTIM7	–1	28.8	–111.84	87	2	Yes
KempTIM8	–0.7	44.2	–108.87	100	3	Yes
KempTIM9	–13.3	29.6	–107.41	91	2	Yes

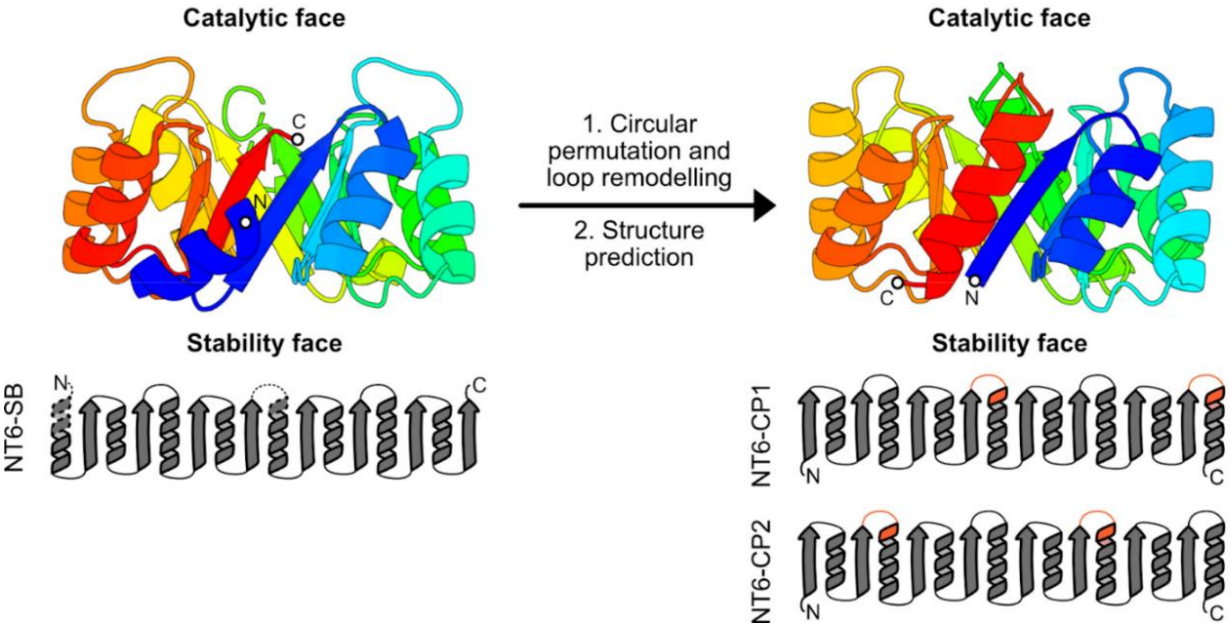
[a] Energies are calculated by subtracting the energy of an all-Gly structure as described in the methods under *Active-site repacking*.
[b] Solvent-accessible surface area of the transition state in the design model.
[c] Preorganized residues are those predicted to adopt the same rotamer in the presence and absence of the transition state. KempTIM1–2, KempTIM3–5, and KempTIM6–9 had 14, 17, and 18 residues that were optimized during active-site repacking, respectively.
[d] The presence of a tunnel was determined by Caver 3.0 using a minimum cutoff bottleneck radius of 0.9 Å.

Table S9. Crystallographic data collection and refinement statistics. Statistics for the highest resolution shell are shown in parentheses.

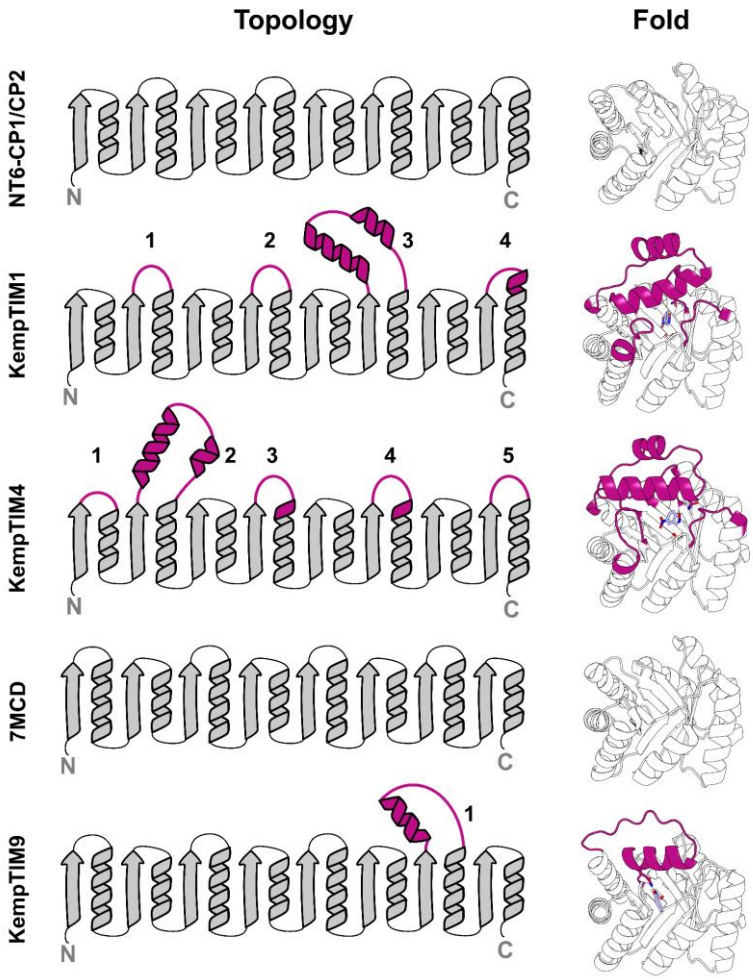
Data collection	
Beamline	BESSY BL14.1
Wavelength [Å]	0.9184
Space group	P 3_1 2 1
Unit cell [Å, °]	a = b = 59.8 c = 120.69 α = β = 90 γ = 90
Resolution range [Å]	47.59 - 2.302 (2.53 - 2.3)
Unique reflections	11469 (2772)
Multiplicity	19.5 (20.4)
Completeness [%]	98.65 (98.30)
R-meas [%]	0.4812 (2.946)
<I/σI>	8.24 (1.58)
$CC_{1/2}$	0.995 (0.404)
*CC**	0.999 (0.759)
Wilson *B*-factor [Å^2]	33.87
Refinement	
R_{work}/R_{free} [%]	25.47/ 29.96
No. of atoms (non-H)	
macromolecules	1640
ligands	42
solvent	52
RMSD from ideal geometry	
bonds [Å]	0.001
angles [°]	0.31
Ramachandran statistics	
favored [%]	98.63
outliers [%]	0.00
Clashscore	4.96
Average *B* [Å^2]	42.19
macromolecules	41.79
ligands	59.28
solvent	41.22



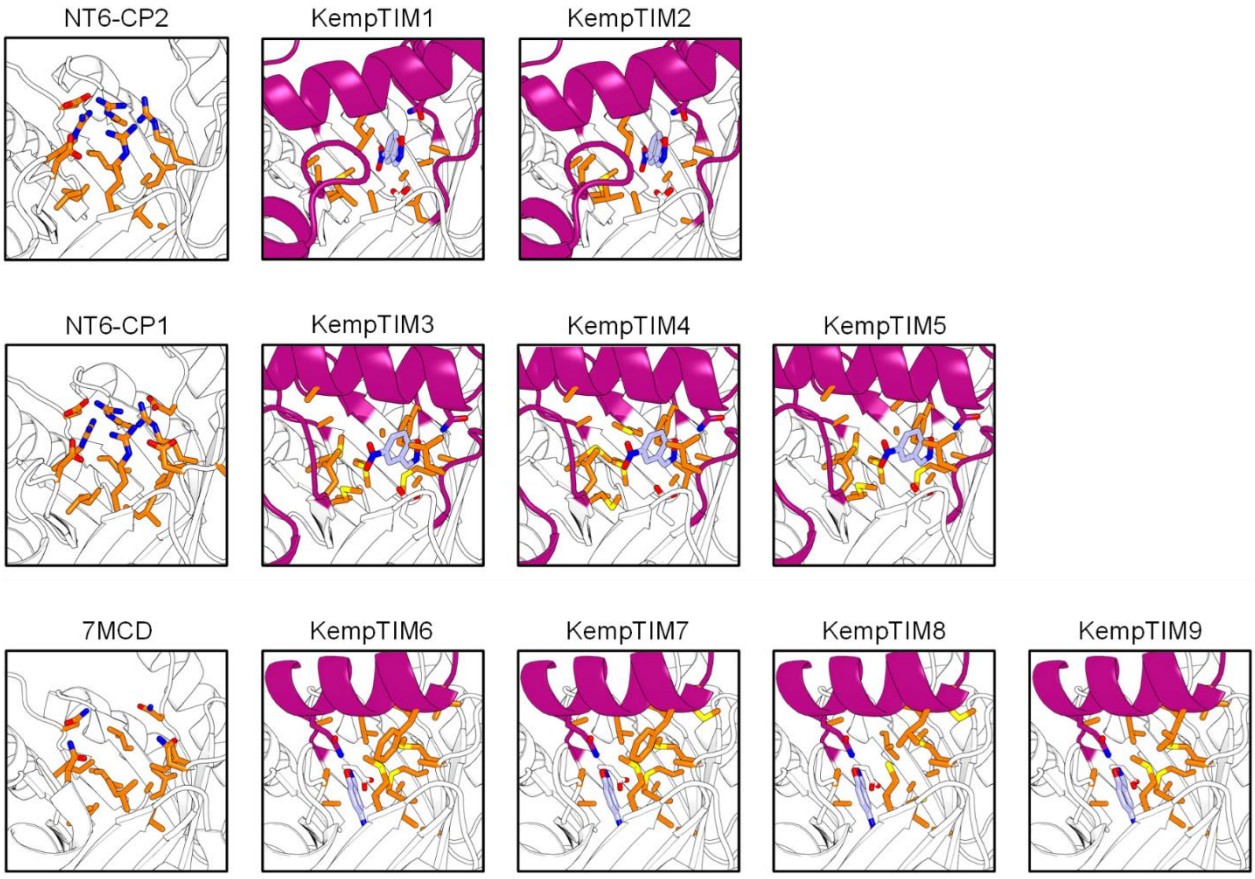
Supplementary Figure 1. Detailed CANVAS workflow. Step 1: A catalytic residue from the theozyme is placed onto a minimal TIM barrel scaffold, and the transition state (TS) is constructed from its side chain using Triad. The second catalytic residue is built from the TS, positioning its α-carbon in the empty space above the TIM barrel catalytic face. Additional catalytic residues can be added inside the barrel or the empty space using similar steps. Step 2: A lid composed of protein fragments of desired length and secondary structure is generated with RFdiffusion to anchor catalytic residues located above the TIM barrel face. Step 3: Lid sequences are designed with ProteinMPNN while maintaining catalytic residue identities. Step 4: Designed structures are predicted with AlphaFold2. Step 5: The amino acid sequence and rotameric configuration of the active site is optimized with Triad using the AlphaFold2 model as template to maximize TS packing and catalytic contact geometry. Resulting sequences are then filtered using key enzyme design criteria. At each step, structures failing filtering criteria are rejected. Approximate number of structures/sequences generated at each step are indicated as variants.



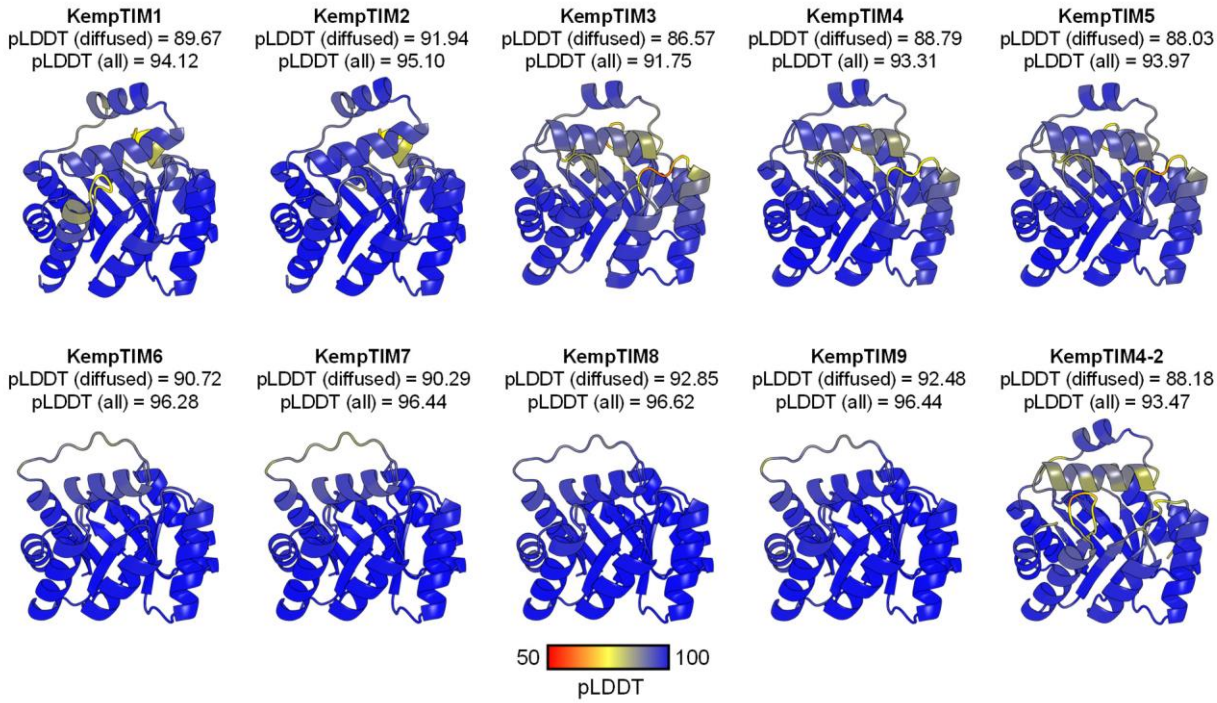
Supplementary Figure 2. Circular permutation of *de novo* TIM barrel NT6-SB (PDB ID: 7OSV). Circularpermutation and loop remodeling were used to relocate the N- and C-termini of NT6-SB (left) from the catalytic face to the stability face. After structure prediction with AlphaFold2, variants NT6-CP1 and NT6-CP2 (right) were generated. Protein structures are colored from blue (N-terminus) to red (C-terminus). Topology diagrams show the secondary structural elements of TIM barrels. Dashed parts of NT6-SB indicate all non-resolved regions of the crystal structure. Orange regions in the circularly permuted TIM barrels represent the remodelled or inserted loops required for circular permutation.



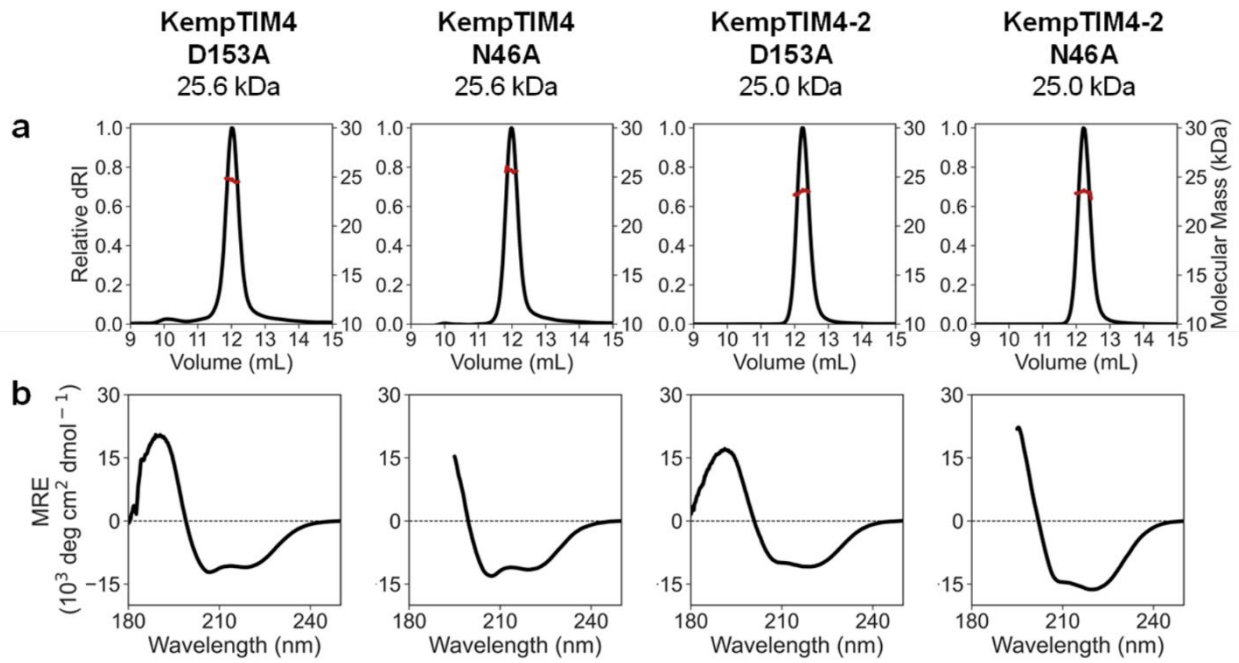
Supplementary Figure 3. Structure of TIM barrels. The topology and fold of various TIM barrels are shown. The designed lids (magenta) of KempTIM1 and KempTIM4 consist of four or five inserted fragments, including a long helix-turn-helix and three or four extended loops. The lid of KempTIM9 contains a single inserted helix-loop fragment.



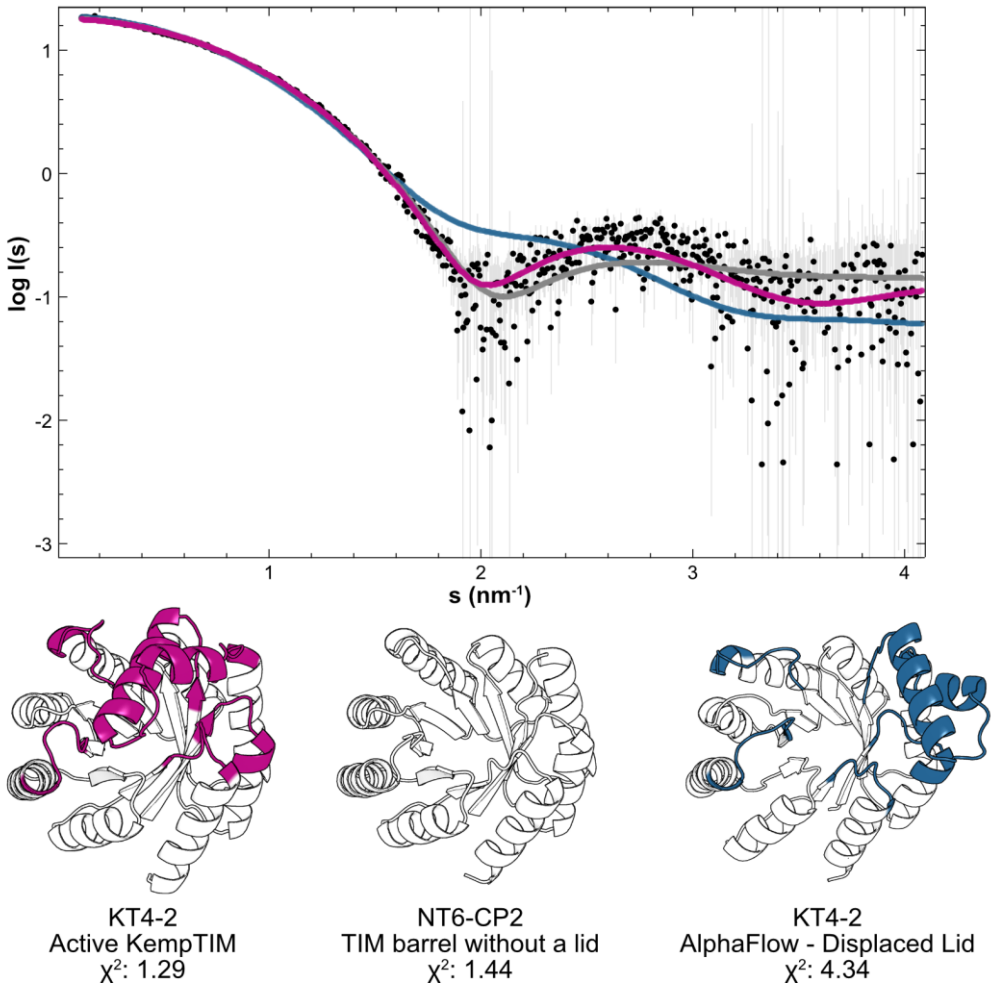
Supplementary Figure 4. Active site configurations of designed enzymes. Residues optimized by Triad during active-site repacking of KempTIM variants, along with the corresponding residues on the parent minimal TIM barrel, are shown in orange. Lids introduced by RFdiffusion are shown in magenta. The catalytic base, catalytic hydrogenbond donor, and transition state are depicted in white, magenta, and light blue sticks, respectively.



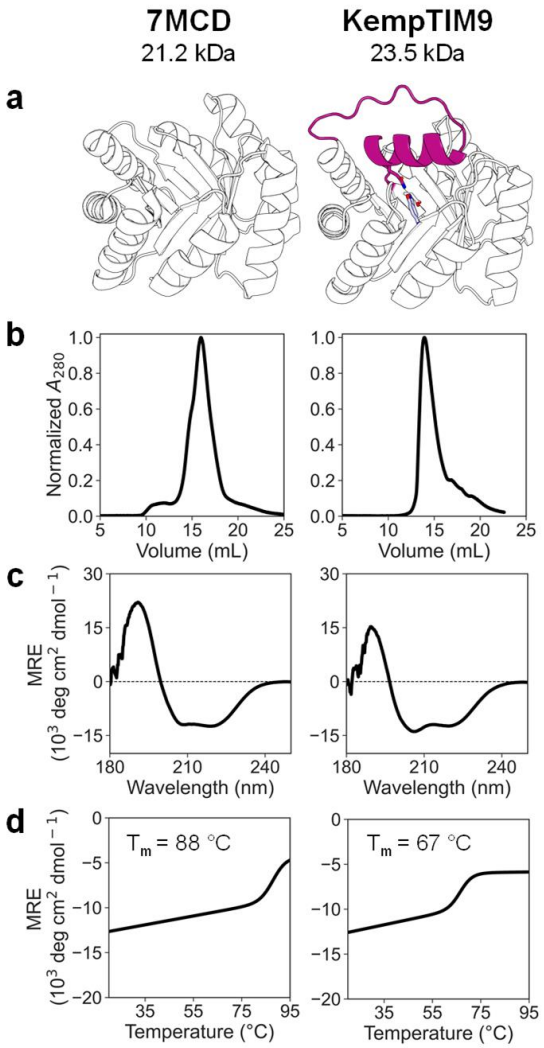
Supplementary Figure 5. AlphaFold2 models. Structures of KempTIM variants are coloured by pLDDT, with the average pLDDT of diffused residues and entire structure listed above each variant.



Supplementary Figure 6. Structural characterization of catalytic knockout mutants. (a) SEC-MALS chromatograms of KempTIM4 and KempTIM4-2 catalytic knockouts indicate that these proteins are predominantly monomeric, with molecular weights matching their expected values. (b) CD spectra demonstrate that mutation of catalytic residues does not change the structure. MRE: mean residue ellipticity.



Supplementary Figure 7. Comparison of KempTIM4-2 experimental SAXS data with scattering calculated from various similar size TIM barrels. Experimental scattering data for KempTIM4-2 (black) align more closely with the calculated scattering curve from the KempTIM4-2 AlphaFold2 model (magenta) than with models of the minimal TIM barrel NT6-CP2 (grey) or a KempTIM4-2 model with improperly positioned lid generated using AlphaFlow (blue).



Supplementary Figure 8. Structural characterization of inactive design KempTIM9. (a) Computational models of KempTIM9 and its parent minimal TIM-barrel 7MCD, with the designed lid shown in magenta and the theozyme as sticks. (b) Preparative SEC show primarily single peaks that might indicate predominantly monomeric proteins. (c) CD spectra reveal a mixed αβ signal characteristic of TIM barrels. MRE: mean residue ellipticity. (d) Melting curves demonstrate that KempTIM9 is substantially destabilized compared to parent 7MCD.

List of Publications

1. S. Kordes*, J. Beck*, S. Shanmugaratnam, M. Flecks, B. Höcker, **Physics-based approach to extend a *de novo* TIM barrel with rationally designed helix-loop-helix motifs.** *Protein Eng. Des. Sel.* **36**, 1–8 (2023).

2. J. Beck, S. Shanmugaratnam, B. Höcker, **Diversifying *de novo* TIM barrels by hallucination.** *Protein Sci.* **33**, e5001 (2024).

3. J. Beck*, B. J. Smith*, N. Zarifi, E. Freund, R. A. Chica, B. Höcker, **Customizing the Structure of a Minimal TIM Barrel to Craft a *De Novo* Enzyme.** *bioRxiv*, 2025.01.28.635154 (2025).

4. ** B. Stüven, R. Stabel, R. Ohlendorf, J. Beck, R. Schubert, A. Möglich, **Characterization and engineering of photoactivated adenylyl cyclases.** *Biol. Chem.* **400**, 429–441 (2019).

* equal contribution

** not part of this thesis

References

1. A. M. Lau, N. Bordin, S. M. Kandathil, I. Sillitoe, V. P. Waman, J. Wells, C. A. Orengo, D. T. Jones, Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science (80-.).* **386** (2024).

2. R. Aharoni, D. Tobi, Dynamical comparison between myoglobin and hemoglobin. *Proteins* **86**, 1176–1183 (2018).

3. H. Saibil, Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* **14**, 630 (2013).

4. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).

5. T. Hamamsy, M. Barot, J. T. Morton, M. Steinegger, R. Bonneau, K. Cho, Learning sequence, structure, and function representations of proteins with language models. *bioRxiv*, 2023.11.26.568742 (2023).

6. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).

7. R. Zwanzig, A. Szabo, B. Bagchi, Levinthal's paradox. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 20 (1992).

8. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science (80-.).* **181**, 223–230 (1973).

9. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

10. C. M. Dobson, Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.* **15**, 3–16 (2004).

11. L. L. Porter, L. L. Looger, Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5968–5973 (2018).

12. T. L. Solomon, Y. He, N. Sari, Y. Chen, D. T. Gallagher, P. N. Bryan, J. Orban, Reversible switching between two common protein folds in a designed system using only temperature. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2215418120 (2023).

13. K. Pauwels, I. Van Molle, J. Tommassen, P. Van Gelder, Chaperoning Anfinsen: the steric foldases. *Mol. Microbiol.* **64**, 917–922 (2007).

14. P. Koehl, M. Levitt, De novo protein design. I. in search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181 (1999).

15. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nat. 2016 5377620* **537**, 320–327 (2016).

16. S. Romero-Romero, S. Kordes, F. Michel, B. Höcker, Evolution, folding, and design of TIM barrels and related proteins. *Curr. Opin. Struct. Biol.* **68**, 94 (2021).

17. R. Geyer, J. R. Jambeck, K. L. Law, Production, use, and fate of all plastics ever made. *Sci. Adv.* **3** (2017).

18. S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura, K. Oda, A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science*

(80-.). **351**, 1196–1199 (2016).

19. H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, Machine learning-aided engineering of hydrolases for PET depolymerization. *Nat. 2022 6047907* **604**, 662–667 (2022).

20. E. L. Bell, R. Smithson, S. Kilbride, J. Foster, F. J. Hardy, S. Ramachandran, A. A. Tedstone, S. J. Haigh, A. A. Garforth, P. J. R. Day, C. Levy, M. P. Shaver, A. P. Green, Directed evolution of an efficient and thermostable PET depolymerase. *Nat. Catal. 2022 58* **5**, 673–681 (2022).

21. V. Pirillo, M. Orlando, C. Battaglia, L. Pollegioni, G. Molla, Efficient polyethylene terephthalate degradation at moderate temperature: a protein engineering study of LC-cutinase highlights the key role of residue 243. *FEBS J.* **290**, 3185–3202 (2023).

22. S. Weigert, P. Perez-Garcia, F. J. Gisdon, A. Gagsteiger, K. Schweinshaut, G. M. Ullmann, J. Chow, W. R. Streit, B. Höcker, Investigation of the halophilic PET hydrolase PET6 from Vibrio gazogenes. *Protein Sci.* **31**, e4500 (2022).

23. O. Turak, A. Gagsteiger, A. Upadhyay, M. Kriegel, P. Salein, S. Agarwal, E. Borchert, B. Höcker, A third type of PETase from the marine Halopseudomonas lineage. *bioRxiv*, 2024.12.31.630877 (2025).

24. E. Bevacqua, C. F. Schleussner, J. Zscheischler, A year above 1.5 °C signals that Earth is most probably within the 20-year period that will reach the Paris Agreement limit. *Nat. Clim. Chang.* **15**, 262–265 (2025).

25. M. Zhao, L. Lei, Y. Jiang, Y. Tian, Y. Huang, M. Yang, Unveiling the Threat of Disease X: Preparing for the Next Global Pandemic. *J. Med. Virol.* **97**, e70227 (2025).

26. Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, V. A. Bohr, Ageing as a risk factor for neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 565–581 (2019).

27. T. Kortemme, De novo protein design—From new structures to programmable functions. *Cell* **187**, 526–544 (2024).

28. A. E. Chu, T. Lu, P. S. Huang, Sparks of function by de novo protein design. *Nat. Biotechnol. 2024 422* **42**, 203–215 (2024).

29. A. Lauko, S. J. Pellock, K. H. Sumida, I. Anishchenko, D. Juergens, W. Ahern, J. Jeung, A. Shida, A. Hunt, I. Kalvet, C. Norn, I. R. Humphreys, C. Jamieson, R. Krishna, Y. Kipnis, A. Kang, E. Brackenbrough, A. K. Bera, B. Sankaran, K. N. Houk, D. Baker, Computational design of serine hydrolases. *Science (80-.).*, doi: 10.1126/SCIENCE.ADU2454 (2025).

30. L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls, Y. J. Park, E. M. Strauch, L. Stewart, M. S. Diamond, D. Veesler, D. Baker, De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science (80-.).* **370** (2020).

31. E. Callaway, Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures. *Nature* **634**, 525–526 (2024).

32. K. I. Albanese, S. Barbe, S. Tagami, D. N. Woolfson, T. Schiex, Computational protein design. *Nat. Rev. Methods Prim. 2025 51* **5**, 1–28 (2025).

33. A. N. Lupas, What I cannot create, I do not understand. *Science (80-.).* **346**, 1455–1456 (2014).

34. I. V. Korendovych, W. F. DeGrado, De novo protein design, a retrospective. *Q. Rev. Biophys.* **53**, e3 (2020).

35. T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, D. C. Richardson, Betadoublet: De novo design, synthesis, and characterization of a β-sandwich protein. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 8747–8751 (1994).

36. J. S. Richardson, D. C. Richardson, The de novo design of protein structures. *Trends Biochem. Sci.*

14, 304–309 (1989).

37. L. Regan, W. F. Degrado, Characterization of a Helical Protein Designed from First Principles. *Science (80-.).* **241**, 976–978 (1988).

38. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science (80-.).* **302**, 1364–1368 (2003).

39. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

40. G. A. Lazar, J. R. Desjarlais, T. M. Handel, De novo design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167–1178 (1997).

41. B. I. Dahiyat, C. A. Sarisky, S. L. Mayo, De Novo protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789–796 (1997).

42. B. I. Dahiyat, S. L. Mayo, De Novo Protein Design: Fully Automated Sequence Selection. *Science (80-.).* **278**, 82–87 (1997).

43. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

44. H. Khakzad, I. Igashov, A. Schneuing, C. Goverde, M. Bronstein, B. Correia, A new age in protein design empowered by deep learning. *Cell Syst.* **14**, 925–939 (2023).

45. P. Gainza, K. E. Roberts, I. Georgiev, R. H. Lilien, D. A. Keedy, C. Y. Chen, F. Reza, A. C. Anderson, D. C. Richardson, J. S. Richardson, B. R. Donald, OSPREY: protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol.* **523**, 87–107 (2013).

46. A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

47. F. S. Lee, A. G. Anderson, B. D. Olafson, Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Eng. Des. Sel.* **36** (2023).

48. K. Maksymenko, A. Maurer, N. Aghaallaei, C. Barry, N. Borbarán-Bravo, T. Ullrich, T. M. H. Dijkstra, B. Hernandez Alvarez, P. Müller, A. N. Lupas, J. Skokowa, M. ElGamacy, The design of functional proteins using tensorized energy calculations. *Cell Reports Methods* **3**, 100560 (2023).

49. A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, S. J. Fleishman, Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346 (2016).

50. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P. S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O.

Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D. A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y. R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665 (2020).

51. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031 (2017).

52. C. Bystroff, D. Baker, Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281**, 565–577 (1998).

53. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).

54. B. Kuhlman, Designing protein structures and complexes with the molecular modeling program Rosetta. *J. Biol. Chem.* **294**, 19436 (2019).

55. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Struct. Funct. Genet.* **37**, 171–176 (1999).

56. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, De novo computational design of retro-aldol enzymes. *Science (80-.).* **319**, 1387–1391 (2008).

57. P. M. Murphy, J. M. Bolduc, J. L. Gallaher, B. L. Stoddard, D. Baker, Alteration of enzyme specificity by computational loop remodeling and design. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9215 (2009).

58. A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Ro¨thlisberger, R. Ro¨thlisberger, D. Baker, New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).

59. R. K. Jha, A. Leaver-Fay, S. Yin, Y. Wu, G. L. Butterfoss, T. Szyperski, N. V. Dokholyan, B. Kuhlman, Computational Design of a PAK1 Binding Protein. *J. Mol. Biol.* **400**, 257–270 (2010).

60. J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, D. Baker, Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J. Mol. Biol.* **331**, 281–299 (2003).

61. F. Dimaio, N. Echols, J. J. Headd, T. C. Terwilliger, P. D. Adams, D. Baker, Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods 2013 1011* **10**, 1102–1104 (2013).

62. P. S. Huang, Y. E. A. Ban, F. Richter, I. Andre, R. Vernon, W. R. Schief, D. Baker, RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **6**, e24109 (2011).

63. T. Van Montfort, M. Melchers, G. Isik, S. Menis, P. S. Huang, K. Matthews, E. Michael, B. Berkhout, W. R. Schief, J. P. Moore, R. W. Sanders, A chimeric HIV-1 envelope glycoprotein trimer with an embedded Granulocyte-Macrophage Colony-stimulating Factor (GM-CSF) domain induces enhanced antibody and T cell responses. *J. Biol. Chem.* **286**, 22250–22261 (2011).

64. B. E. Correia, M. A. Holmes, P. S. Huang, R. K. Strong, W. R. Schief, High-resolution structure prediction of a circular permutation loop. *Protein Sci.* **20**, 1929–1934 (2011).

65. J. Thiyagalingam, M. Shankar, G. Fox, T. Hey, Scientific machine learning benchmarks. *Nat. Rev. Phys. 2022 46* **4**, 413–420 (2022).

66. A. Holzinger, A. Saranti, A. Angerschmid, B. Finzel, U. Schmid, H. Mueller, Toward human-level

concept learning: Pattern benchmarking for AI algorithms. *Patterns* **4**, 100788 (2023).

67. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science (80-.).* **378**, 49–56 (2022).

68. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, D. Baker, De novo protein design by deep network hallucination. *Nat. 2021 6007889* **600**, 547–552 (2021).

69. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nat. 2023 6207976* **620**, 1089–1100 (2023).

70. N. Bordin, C. Dallago, M. Heinzinger, S. Kim, M. Littmann, C. Rauer, M. Steinegger, B. Rost, C. Orengo, Novel machine learning approaches revolutionize protein knowledge. *Trends Biochem. Sci.* **48**, 345–359 (2023).

71. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins Struct. Funct. Bioinforma.* **91**, 1539–1549 (2023).

72. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nat. 2020 5777792* **577**, 706–710 (2020).

73. M. Alquraishi, AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).

74. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinforma.* **89**, 1711–1721 (2021).

75. D. T. Jones, J. M. Thornton, The impact of AlphaFold2 one year on. *Nat. Methods 2022 191* **19**, 15–20 (2022).

76. M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).

77. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034 (2022).

78. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C. C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions

with AlphaFold 3. *Nat. 2024 6308016* **630**, 493–500 (2024).

79. J. Hennig, Structural Biology of RNA and Protein-RNA Complexes after AlphaFold3. *ChemBioChem*, e202401047 (2025).

80. Y. Du, J. Meier, J. Ma, R. Fergus, A. Rives, Energy-based models for atomic-resolution protein conformations. *8th Int. Conf. Learn. Represent. ICLR 2020* (2020).

81. R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **9**, 2927–2935 (2020).

82. J. Wang, H. Cao, J. Z. H. Zhang, Y. Qi, Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **8** (2018).

83. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. Dustin Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science (80-.).* **373**, 871–876 (2021).

84. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1496–1503 (2020).

85. N. Ferruz, B. Höcker, Dreaming ideal protein structures. *Nat. Biotechnol.* **40**, 171–172 (2022).

86. J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J. H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, D. Baker, Scaffolding protein functional sites using deep learning. *Science (80-.).* **377**, 387–394 (2022).

87. N. Ferruz, S. Schmidt, B. Höcker, ProteinTools: a toolkit to analyze protein structures. *Nucleic Acids Res.* **49**, W559–W566 (2021).

88. D. Akpinaroglu, K. Seki, A. Guo, E. Zhu, M. J. S. Kelly, T. Kortemme, Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space. *bioRxiv*, 2023.12.15.571823 (2023).

89. C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, F. Players, D. Baker, S. Ovchinnikov, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2017228118 (2021).

90. B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, T. Jaakkola, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. *11th Int. Conf. Learn. Represent. ICLR 2023* (2022).

91. C. Frank, A. Khoshouei, L. Fuβ, D. Schiwietz, D. Putz, L. Weber, Z. Zhao, M. Hattori, S. Feng, Y. de Stigter, S. Ovchinnikov, H. Dietz, Scalable protein design using optimization in a relaxed sequence space. *Science* **386**, 439–445 (2024).

92. N. Anand, R. Eguchi, I. I. Mathews, C. P. Perez, A. Derry, R. B. Altman, P. S. Huang, Protein sequence design with a learned potential. *Nat. Commun. 2022 131* **13**, 1–11 (2022).

93. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assemblies. *Science (80-.).* **378**, 2025 (2022).

94. N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. De Munck, S. N. Savvides, D. Baker, Improving de novo protein

binder design with deep learning. *Nat. Commun. 2023 141* **14**, 1–9 (2023).

95. K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. Wicky, L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson, A. Kang, J. De La Cruz, B. Sankaran, A. K. Bera, G. Jiménez-Osés, D. Baker, Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).

96. R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. Neil Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science (80-.).* **384** (2024).

97. J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, D. Baker, Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods 2025*, 1–7 (2025).

98. D. Medina-Ortiz, A. Khalifeh, H. Anvari-Kazemabad, M. D. Davari, Interpretable and explainable predictive machine learning models for data-driven protein engineering. *Biotechnol. Adv.* **79**, 108495 (2025).

99. W. Samek, K. R. Müller, Towards Explainable Artificial Intelligence. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **11700 LNCS**, 5–22 (2019).

100. B. Huang, B. Coventry, M. T. Borowska, D. C. Arhontoulis, M. Exposit, M. Abedi, K. M. Jude, S. F. Halabiya, A. Allen, C. Cordray, I. Goreshnik, M. Ahlrichs, S. Chan, H. Tunggal, M. DeWitt, N. Hyams, L. Carter, L. Stewart, D. H. Fuller, Y. Mei, K. C. Garcia, D. Baker, De novo design of miniprotein antagonists of cytokine storm inducers. *Nat. Commun. 2024 151* **15**, 1–11 (2024).

101. S. Vázquez Torres, M. Benard Valle, S. P. Mackessy, S. K. Menzies, N. R. Casewell, S. Ahmadi, N. J. Burlet, E. Muratspahić, I. Sappington, M. D. Overath, E. Rivera-de-Torre, J. Ledergerber, A. H. Laustsen, K. Boddum, A. K. Bera, A. Kang, E. Brackenbrough, I. A. Cardoso, E. P. Crittenden, R. J. Edge, J. Decarreau, R. J. Ragotte, A. S. Pillai, M. Abedi, H. L. Han, S. R. Gerben, A. Murray, R. Skotheim, L. Stuart, L. Stewart, T. J. A Fryer, T. P. Jenkins, D. Baker, De novo designed proteins neutralize lethal snake venom toxins. *Nature* **639** (2025).

102. W. Yang, D. R. Hicks, A. Ghosh, T. A. Schwartze, B. Conventry, I. Goreshnik, A. Allen, S. F. Halabiya, C. J. Kim, C. S. Hinck, D. S. Lee, A. K. Bera, Z. Li, Y. Wang, T. Schlichthaerle, L. Cao, B. Huang, S. Garrett, S. R. Gerben, S. Rettie, P. Heine, A. Murray, N. Edman, L. Carter, L. Stewart, S. C. Almo, A. P. Hinck, D. Baker, Design of high-affinity binders to immune modulating receptors for cancer immunotherapy. *Nat. Commun.* **16** (2025).

103. M. Braun, A. Tripp, M. Chakatok, S. Kaltenbrunner, M. G. Totaro, D. Stoll, A. Bijelic, W. Elaily, S. Y. Y. Hoch, M. Aleotti, M. Hall, G. Oberdorfer, Computational design of highly active de novo enzymes. *bioRxiv*, 2024.08.02.606416 (2024).

104. I. Anishchenko, Y. Kipnis, I. Kalvet, G. Zhou, R. Krishna, S. J. Pellock, A. Lauko, G. R. Lee, L. An, J. Dauparas, F. DiMaio, D. Baker, Modeling protein-small molecule conformational ensembles with ChemNet. *bioRxiv*, 2024.09.25.614868 (2024).

105. S. L. Lisanza, J. M. Gershon, S. W. K. Tipps, J. N. Sims, L. Arnoldt, S. J. Hendel, M. K. Simma, G. Liu, M. Yase, H. Wu, C. D. Tharp, X. Li, A. Kang, E. Brackenbrough, A. K. Bera, S. Gerben, B. J. Wittmann, A. C. McShan, D. Baker, Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nat. Biotechnol. 2024*, 1–11 (2024).

106. W. Ahern, J. Yim, D. Tischer, S. Salike, S. M. Woodbury, D. Kim, I. Kalvet, Y. Kipnis, B. Coventry, H. R. Altae-Tran, M. Bauer, R. Barzilay, T. S. Jaakkola, R. Krishna, D. Baker, Atom level enzyme active site scaffolding using RFdiffusion2. *bioRxiv*, 2025.04.09.648075 (2025).

107. S. Romero-Romero, M. Costas, D. A. Silva Manzano, S. Kordes, E. Rojas-Ortega, C. Tapia, Y. Guerra, S. Shanmugaratnam, A. Rodríguez-Romero, D. Baker, B. Höcker, D. A. Fernández-Velasco, The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach.

J. Mol. Biol. **433** (2021).

108. R. Sterner, B. Höcker, Catalytic versatility, stability, and evolution of the (βα)8-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055 (2005).

109. A. Faljoni, G. Cilento, Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem. Biophys. Res. Commun.* **72**, 146–155 (1976).

110. R. K. Wierenga, The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).

111. P. S. Huang, K. Feldmeier, F. Parmeggiani, D. F. Velasco, B. Hocker, D. Baker, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).

112. N. Nagano, C. A. Orengo, J. M. Thornton, One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *J. Mol. Biol.* **321**, 741–765 (2002).

113. M. S. Vijayabaskar, S. Vishveshwara, Insights into the Fold Organization of TIM Barrel from Interaction Energy Based Structure Networks. *PLOS Comput. Biol.* **8**, e1002505 (2012).

114. R. K. Wierenga, E. G. Kapetaniou, R. Venkatesan, Triosephosphate isomerase: a highly evolved biocatalyst. *Cell. Mol. Life Sci. 2010 6723* **67**, 3961–3982 (2010).

115. T. D. Myers, M. J. Palladino, Newly discovered roles of triosephosphate isomerase including functions within the nucleus. *Mol. Med.* **29**, 1–6 (2023).

116. W. A. Lim, R. T. Raines, J. R. Knowles, Triosephosphate Isomerase Catalysis Is Diffusion Controlled. *Biochemistry* **27**, 1158–1165 (1988).

117. V. Olivares-Illana, H. Riveros-Rosas, N. Cabrera, M. Tuena de Gómez-Puyou, R. Pérez-Montfort, M. Costas, A. Gómez-Puyou, A guide to the effects of a large portion of the residues of triosephosphate isomerase on catalysis, stability, druggability, and human disease. *Proteins Struct. Funct. Bioinforma.* **85**, 1190–1211 (2017).

118. T. Tanaka, Y. Kuroda, H. Kimura, S. ichi Kidokoro, H. Nakamura, Cooperative deformation of a de novo designed protein. *Protein Eng. Des. Sel.* **7**, 969–976 (1994).

119. T. Tanaka, M. Hayashi, H. Kimura, M. Oobatake, H. Nakamura, De novo design and creation of a stable artificial protein. *Biophys. Chem.* **50**, 47–61 (1994).

120. M. Beauregard, K. Goraj, V. Goffin, K. Heremans, E. Goormaghtigh, J. M. Ruysschaert, J. A. Martial, Spectroscopic investigation of structure in octarellin (a de novo protein designed to adopt the alpha/beta-barrel packing). *Protein Eng.* **4**, 745–749 (1991).

121. K. Goraj, A. Renard, J. A. Martial, Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the α/β-barrel Proteins. *Protein Eng. Des. Sel.* **3**, 259–266 (1990).

122. M. Figueroa, N. Oliveira, A. Lejeune, K. W. Kaufmann, B. M. Dorr, A. Matagne, J. A. Martial, J. Meiler, C. Van de Weerdt, Octarellin VI: using rosetta to design a putative artificial (β/α)8 protein. *PLoS One* **8** (2013).

123. F. Offredi, F. Dubail, P. Kischel, K. Sarinski, A. S. Stern, C. Van de Weerdt, J. C. Hoch, C. Prosperi, J. M. François, S. L. Mayo, J. A. Martial, De novo Backbone and Sequence Design of an Idealized α/β-barrel Protein: Evidence of Stable Tertiary Structure. *J. Mol. Biol.* **325**, 163–174 (2003).

124. M. Figueroa, M. Sleutel, M. Vandevenne, G. Parvizi, S. Attout, O. Jacquin, J. Vandenameele, A. W. Fischer, C. Damblon, E. Goormaghtigh, M. Valerio-Lepiniec, A. Urvoas, D. Durand, E. Pardon, J. Steyaert, P. Minard, D. Maes, J. Meiler, A. Matagne, J. A. Martial, C. Van de Weerdt, The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure

prediction tools. *J. Struct. Biol.* **195**, 19–30 (2016).

125. D. Nagarajan, G. Deka, M. Rao, Design of symmetric TIM barrel proteins from first principles. *BMC Biochem.* **16**, 1–22 (2015).

126. N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, Principles for designing ideal protein structures. *Nat. 2012 4917423* **491**, 222–227 (2012).

127. S. Kordes, S. Romero-Romero, L. Lutz, B. Höcker, A newly introduced salt bridge cluster improves structural and biophysical properties of de novo TIM barrels. *Protein Sci.* **31**, 513–527 (2022).

128. A. Losi, E. Polverini, B. Quest, W. Gärtner, First evidence for phototropin-related blue-light receptors in prokaryotes. *Biophys. J.* **82**, 2627–2634 (2002).

129. and D. S. H. Xiwei Zheng, Cong Bi, Marissa Brooks, *HHS Public Access* (2015)vol. 25.

130. A. E. Chu, D. Fernandez, J. Liu, R. R. Eguchi, P. S. Huang, De Novo Design of a Highly Stable Ovoid TIM Barrel: Unlocking Pocket Shape towards Functional Design. *BioDesign Res.* **2022** (2022).

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

132. S. J. Caldwell, I. C. Haydon, N. Piperidou, P. S. Huang, M. J. Bick, H. Sebastian Sjöström, D. Hilvert, D. Baker, C. Zeymer, Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 30362–30369 (2020).

133. A. S. Klein, F. Leiss-Maier, R. Mühlhofer, B. Boesen, G. Mustafa, H. Kugler, C. Zeymer, A De Novo Metalloenzyme for Cerium Photoredox Catalysis. *J. Am. Chem. Soc.* **146** (2024).

134. S. Kordes, J. Beck, S. Shanmugaratnam, M. Flecks, B. Höcker, Physics-based approach to extend a de novo TIM barrel with rationally designed helix-loop-helix motifs. *Protein Eng. Des. Sel.* **36**, 1–8 (2023).

135. J. Beck, S. Shanmugaratnam, B. Höcker, Diversifying de novo TIM barrels by hallucination. *Protein Sci.* **33**, e5001 (2024).

136. W. M. Dawson, G. G. Rhys, D. N. Woolfson, Towards functional de novo designed proteins. *Curr. Opin. Chem. Biol.* **52**, 102–111 (2019).

137. J. Beck, B. J. Smith, N. Zarifi, E. Freund, R. A. Chica, B. Höcker, Customizing the Structure of a Minimal TIM Barrel to Craft a De Novo Enzyme. *bioRxiv*, 2025.01.28.635154 (2025).

138. D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, Kemp elimination catalysts by computational enzyme design. *Nat. 2008 4537192* **453**, 190–195 (2008).

Acknowledgements

In the first place, I would like to express my sincere gratitude to my supervisor, Birte, for giving me the opportunity to pursue my Ph.D. under her guidance. I am thankful for her support, encouragement, and advice throughout the years, as well as for the opportunity to work on fascinating projects and attend numerous conferences.

I would also like to thank my past and present colleagues for their support and for creating the best working atmosphere – from playing soccer to BBQ events.

Many thanks to all the amazing students I had the pleasure of supervising and working with over the years. I hope you learned as much from me as I learned from you.

A big thanks to Andreas for his friendship and almost unstoppable motivation throughout the years – from our first bachelor events to the completion of our doctoral journey. University life, and our shared flat, would not have been nearly as much fun without you.

I would like to thank Christina for her love and support. Thank you for putting up with me over the past few years. I hope we have not driven each other too crazy – especially while writing this thesis – and continue like that for years to come.

I would like to extend my thanks to my family and friends, whose unwavering support has accompanied me throughout my life.

Eidesstattliche Versicherungen und Erklärungen

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit versichere ich eidesstattlich, dass ich die Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe (vgl. Art. 97 Abs. 1 Satz 8 BayHIG).

(§ 8 Satz 2 Nr. 3 PromO Fakultät)

Hiermit erkläre ich, dass ich die Dissertation nicht bereits zur Erlangung eines akademischen Grades eingereicht habe und dass ich nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden habe.

(§ 8 Satz 2 Nr. 4 PromO Fakultät)

Hiermit erkläre ich, dass ich Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe noch künftig in Anspruch nehmen werde.

(§ 8 Satz 2 Nr. 7 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass die elektronische Fassung der Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung unterzogen werden kann.

(§ 8 Satz 2 Nr. 8 PromO Fakultät)

Hiermit erkläre ich mein Einverständnis, dass bei Verdacht wissenschaftlichen Fehlverhaltens Ermittlungen durch universitätsinterne Organe der wissenschaftlichen Selbstkontrolle stattfinden können.

(Ort, Datum) Unterschrift