Enzyme optimization and design with deep learning

Kiera Harumi Sumida

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

David Baker, Chair

Dustin Maly

Jorge Marchand

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Enzyme optimization and design with deep learning

Kiera Harumi Sumida

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

Enzymes are valuable tools for the catalysis of reactions under mild conditions. However, they often misfold when applied outside of their natural contexts. Thus, methods to optimize natural enzymes and design completely bespoke enzymes would be highly valuable. Recent advances in deep learning-guided protein design have made tractable what have historically been grand challenges in the field. Here, I apply these tools to enzyme design, both in the optimization of natural enzymes and the generation of completely novel hydrolases and plastic-degrading enzymes. In the former, I developed a simple and accessible method of sequence redesign to optimize the physical properties of valuable natural proteins. As a proof of concept, I applied this method to the protease from Tobacco etch virus (TEV protease). The designed variants showed increased expression, stability, and even function. In the latter project, I applied *de novo* protein design and modeling tools to create bespoke hydrolases for the model reaction of simple ester hydrolysis, and achieved catalytic efficiencies of up to 10^5 M^{-1} s^{-1}. Applying these tools to the hydrolysis of polyethylene terephthalate (PET), I created, to our knowledge, the first completely *de novo* designed plastic-degrading enzyme. These data represent significant advances in the field of enzyme design and establish methods for the development of new catalysts.

Table of Contents

Acknowledgements

There are innumerable people whom I would like to thank for their support of this work. Firstly, my family – Saskia, James, Mum, Dad, and Auntie Janne – who believed in me and held me up through all the highs and lows of a doctoral program. I thank Sam Pellock, who has been both my mentor and close friend, sharing my passion for PETases and 70's funk. Anna Lauko, who also started as my mentor and quickly became my best friend, whose level-headedness grounded me in both science and life. My friends and roommates, Daniel Brush and Noel Jameson, who have been my family in Seattle and brought so much levity and fun to my life. I would like to thank momi for always being there for me at the end of the day and never trying to talk to me about work. Finally, I would like to thank Eden Tzanetopoulos, my best friend, who has inspired me with her intelligence and kindness, made me laugh more than anyone else, and supported me endlessly throughout these last five years.

Thank you all.

Dedication

This work is dedicated to scientists of all underrepresented groups. By being here, you are changing who the future generation believes a scientist can be.

Chapter 1: Introduction

The central dogma of biology outlines the flow of genetic information from DNA to protein to function. While many are familiar with DNA's role as the storage facility for our genetic information, fewer understand the process in which it is constantly being transcribed and translated in the body to generate proteins. Most proteins serve primarily as structural support for life's architectures, but a specific subset known as enzymes fulfill a critical role by performing chemistries which drive physiological processes. For example, acetylcholine esterase (AChE) catalyzes the degradation of acetylcholine, a neurotransmitter which signals the nervous system through transient interactions at neuromuscular junctions. Deficiency of AChE would result in overstimulation of the nervous system and neurological dysfunction (*1*). Enzymes like AChE catalyze an astounding diversity of reactions, many of which have been co-opted for application in industry, making them critical to countless natural and unnatural human processes.

1.1 – Biocatalysis for chemical transformation

Enzymes are proteins that accelerate chemical transformations by decreasing the activation energy of the catalyzed reaction. This is achieved through some combination of the following mechanisms: substrate localization, transition state stabilization, covalent catalysis, and general acid-base catalysis. The resulting rate enhancements are specific and substantial in their magnitude. Urease, which catalyzes the hydrolysis of urea, increases the rate by a factor of $\sim 10^{14}$ over the uncatalyzed reaction (*2*). Traditional chemical catalysts often necessitate harsh conditions, toxic precursors, and multiple purification steps. Enzymes offer an attractive alternative due to their ability to catalyze reactions specifically and efficiently under mild, physiological conditions.

The global market for industrial enzymes is projected to be $25.88 billion by 2029, with applications in therapeutics, food science, chemical synthesis, and more. In the household product industry, enzymes have long been used as additives for the breakdown of oils, proteins, and starches

which contaminate food- and dish-waste while posing less threat to the environment than their chemical counterparts. Enzymes have also revolutionized chemical synthesis processes across industries. Where traditional chemical syntheses often necessitate hazardous solvents at high temperatures and pressures, enzymes operate at ambient temperature and pressure in aqueous solvent, facilitating greener and cheaper chemical processes. Additionally, the inherent chiral environment of an enzyme active site often enables synthesis with high enantiomeric excess of the target product. This feature has caused them to be critical to the pharmaceutical industry, wherein approximately 50% of drugs are chiral, and enantiomers are known to possess significant differences in *in vivo* behavior (*3*). One notable example is the thalidomide disaster, in which the drug thalidomide was prescribed as a racemic mixture for nausea in pregnant women (*4*). It was later discovered that, while the (*R*)-enantiomer does possess therapeutic effect, the (*S*)-enantiomer is teratogenic, resulting in thousands of instances of birth defects amongst children exposed to the drug *in utero* (*5*).

1.2 – Limitations and traditional optimization

Despite their utility, many natural enzymes possess properties which render their application challenging. Most have evolved function to be possible solely within their natural contexts, and thus have limited tolerance to the conditions used for industrial protein production. Consequently, considerable effort has been made to improve the stability and soluble expression of technologically important enzymes. One prominent technique utilizes ancestral sequence reconstruction to revive extinct proteins predicted to possess similar activity with enhanced thermostability, and thus serve as more effective scaffolds for stability engineering (*6*). This method is based upon the observation that many pre-Cambrian ancestral proteins are significantly more stable than their extant forms. Additional efforts utilize rational design to increase stability. For example, disulfide bridges which stabilize the folded state (*7*) have in some cases been successfully introduced to increase protein stability (*8*), although they are not applicable to all targets (*9*). More recently, computational methods have been developed to increase stability and

soluble expression while preserving function; PROSS (Protein Repair One Stop Shop) utilizes physics-based calculations and evolutionary information to model sequence variants and predict stability through related *in silico* metrics such as overall energy (*10*).

Enhancing the activity, rather than stability, of natural enzymes has historically been more challenging due to the limited understanding of physical features which correlate to function. Laboratory evolution, in which random single-point mutants are made in the active site, evaluated for the desired property, and propagated onto subsequent mutants (*11*), has proven to be robust in generating variants with increased activities and improved stability (*12*). However, it is time- and labor-intensive, requiring multiple rounds of evolution to accumulate beneficial mutations. It is also heavily dependent on the existence of a medium- to high-throughput functional screen and some measurable basis of activity in the starting point, which does not exist for many chemical transformations.

1.3 – Expanding functionality via *de novo* design

Given the limitations of natural enzymes, the ability to design bespoke proteins for a given chemical reaction would have an incredible impact on various industries. Natural enzymes, through the slow evolution process, sample a very limited region within the sequence and structure landscape of proteins. *De novo* protein design seeks to find novel solutions to functional problems by accessing these unsampled spaces, resulting in a protein customized for the desired functionality with high stability and soluble expression. This is accomplished through rational design driven by first principles (*13*). Historically, the field was centered around the so-called “structure prediction problem”, which describes the relationship between protein sequence and structure. Although elusive, the prediction of protein structure from amino acid sequence is considered to be critical to effective protein design. As machine learning-based models showed increasing success in analogous tasks such as image denoising, developers began to apply similar models to protein structure prediction. Critically, training of deep learning models requires a large amount of data which hypothetically encodes the information being predicted. For

proteins, this came in the form of the Protein Data Bank (PDB) (*14*), a database of experimentally determined protein crystal structures and their corresponding sequences. In 2021, two tools utilizing three-track neural networks trained on the PDB, AlphaFold2 and RoseTTAfold, showed unprecedented accuracy on the structure prediction task (*15*, *16*). Not only did these advances expand the field's understanding of proteins, but they also provided a critical *in silico* metric for evaluation of the quality of *de novo* designed proteins. Ensuring self-consistency between designed and predicted structures *in silico* is a large indicator of experimental success.

Inspired by this success, developers sought to use the same deep learning-based models to perform the inverse task – not of protein structure *prediction*, but *generation*. The two key elements in this process, both of which affect function and physical properties, are the 3-D structure and amino acid sequence. ProteinMPNN, a graph neural network which designs protein sequences for a given structure using context-dependent amino acid probability distributions at each position, showed great success in creating sequences that fold to the desired structures and express with high soluble yields and stability *in vitro* (*17*). RFdiffusion, released in 2024, accomplished structure generation by initializing with random Gaussian noise and iteratively denoising to a complete protein structure (*18*). Thus, RFdiffusion could be used to generate protein structures scaffolding a desired motif, and ProteinMPNN could then generate amino acid sequences which fold to these structures.[1] This strategy has proven to be highly successful in the design of protein and peptide binders (*19–21*). While these tools utilize different architectures, they are both trained on information from the PDB.

Despite these advances, enzyme design has remained an outstanding challenge in the field of *de novo* protein design because reactivity often requires extremely precise positioning of catalytic elements, and thus, atomic accuracy in both the structure generation and prediction steps, which has historically been out of reach. Traditionally, enzyme design has involved the definition of a "theozyme" containing catalytic residues around a target substrate, which was then matched into a native protein scaffold

[1] While simultaneous generation of protein sequence and structure is the more elegant solution, current methods for backbone design suffer from low-quality amino acid sequences. Efforts to improve these are ongoing.

(*22–27*). This strategy was significantly limited by the compatibility (or lack thereof) between the theozyme and scaffold library. Since its creation, RFdiffusion has enabled the generation of completely novel proteins which scaffold theozymes, and significantly increased the success rates of such design campaigns.(*28*) However, due to the high degree of accuracy required, many reactions – particularly ones utilizing complex catalytic machinery – remain undesignable.

1.4 – Preface

In this work I will present three projects centered around computational enzyme design: deep learning-based sequence redesign of industrially-relevant natural enzymes to enhance physical properties, *de novo* design of serine hydrolases for model esterase reactions, and *de novo* design of plastic-degrading enzymes. Each of these utilize some combination of the tools described above with chemical intuition to expand the protein universe beyond what evolution has made. I also hope to convince you of the potential of protein design as a solution for many pressing environmental concerns, which has been the constant motivation through my academic career.

Chapter 2: Improving protein expression, stability, and function with ProteinMPNN

Note: The majority of this chapter is borrowed directly from the manuscript of the same name published in *J. Am. Chem. Soc.* in 2024. I, Kiera H. Sumida, am the lead author of the study.

Kiera H. Sumida[1,2], Reyes Núñez-Franco[3], Indrek Kalvet[2,4,5], Samuel J. Pellock[2,4], Basile I. M. Wicky[2,4], Lukas F. Milles[2,4], Justas Dauparas[2,4], Jue Wang[2,4], Yakov Kipnis[2,4,5], Noel Jameson[1], Alex Kang[2], Joshmyn De La Cruz[2], Banumathi Sankaran[6], Asim K. Bera[2,4], Gonzalo Jiménez-Osés[3,7], David Baker[2,4,5]*

[1]Department of Chemistry, University of Washington, Seattle, WA, USA.

[2]Institute for Protein Design, University of Washington, Seattle, WA, USA.

[3]Center for Cooperative Research in Biosciences, Basque Research and Technology Alliance, Derio, Spain.

[4]Department of Biochemistry, University of Washington, Seattle, WA, USA.

[5]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

[6]Berkeley Center for Structural Biology, Molecular Biophysics, and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA, USA.

[7]Ikerbasque, Basque Foundation for Science, 48013 Bilbao, Spain.

*Corresponding author. E-mail: dabaker@uw.edu

2.1 – Introduction

Evolution has optimized function over stability in many natural proteins (*29*); as a result, they often exhibit poor solubility, thermostability, and expression in heterologous systems, all of which reduce the yield of functional protein (*30*, *31*). Many protein-based therapeutics and catalysts are limited in their industrial application by low stability, making protein stabilization a research area of increasing interest (*6*, *32*). Experimental methods such as directed evolution have been extensively used to optimize desirable features in proteins, but are often prohibitively resource- and labor-intensive (*11*, *33*). Computational tools have been developed to achieve the benefits of directed evolution while minimizing experimental screening (*10*, *34–36*). PROSS (protein repair one-stop shop), for example, utilizes evolutionary information and Rosetta physics-based energy calculations to perform sequence redesign using a three-dimensional (3D) structure as input, and has been shown to increase soluble expression and thermostability of several natural proteins (*10*). More recently, advances in deep learning-based modeling of proteins have been applied to generate new variants of natural proteins, including language models that generate sequences for a given enzyme family or function (*36*), convolutional neural networks that leverage structural information for prediction of gain-of-function mutations (*35*), and shallow neural networks for guiding combinatorial directed evolution (*37*).

Deep learning-based tools for protein sequence design have shown success in the generation of novel proteins with excellent expression, solubility, and sub-angstrom accuracy to design models (*17*, *36*, *38*). ProteinMPNN generates highly stable sequences for designed backbones, and for native backbones, generates sequences that are predicted to fold to the intended structures more confidently than their native sequences (*17*). We reasoned that ProteinMPNN could be applied to protein stability optimization, and set out to develop a strategy for applying ProteinMPNN to natural proteins to increase solubility and stability. We chose as model systems one of the first proteins whose structure was solved, the oxygen storage protein myoglobin, and the widely used protease from tobacco etch virus (TEV).

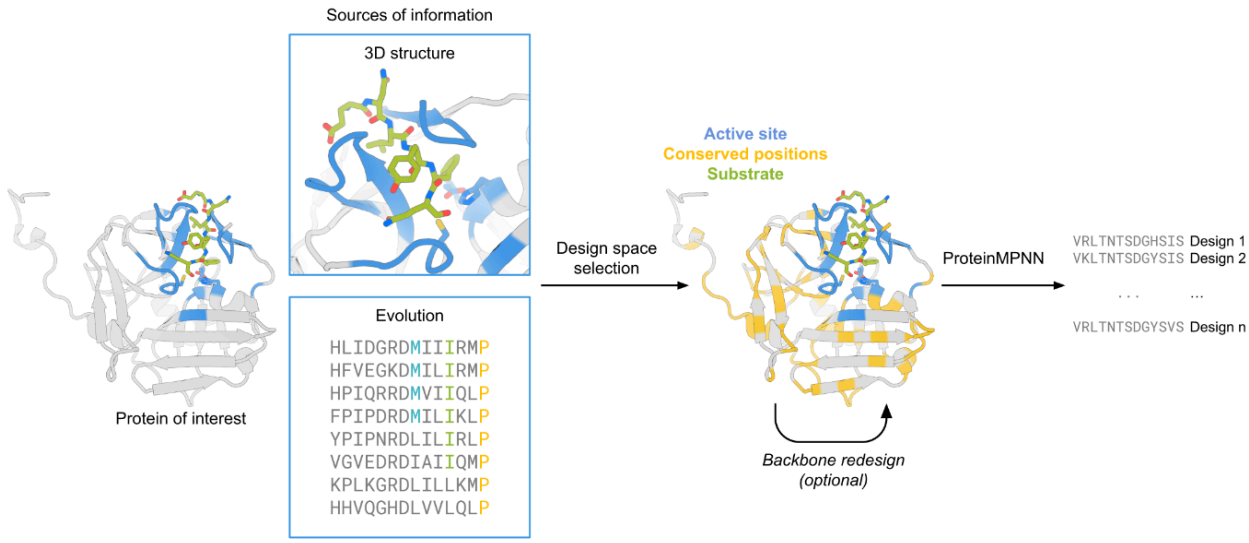

Figure 1. Design strategy for optimization of protein expression and stability using ProteinMPNN. The design space is chosen to preserve native protein function by fixing the amino acid identities of residues close to the ligand/substrate and those that are highly conserved in multiple sequence alignments. The protein backbone structure and fixed position information of amino acids are input into ProteinMPNN, which generates new amino acid sequences likely to fold to the input structure. The backbone structure in loop regions can optionally be remodeled using RoseTTAfold joint inpainting to further idealize the input protein.

2.2 – Protein stabilization with ProteinMPNN

ProteinMPNN generates amino acid sequences that are predicted to fold to a given 3D structure. The method is purely structure-based, and does not have access to functional information. Therefore, to retain protein function during sequence design, additional information must be provided to the network. We experimented with a range of approaches to retain functionality during the design process. In all targets, to preserve the catalytic machinery and substrate-binding site, we fixed the amino acid identities of the first shell functional positions – defined as those within 7 Å of the substrate in a ligand-bound crystal structure complex. In TEV protease, we used evolutionary information to further identify residues critical to activity. In myoglobin, we performed limited backbone redesign to further stabilize the structure. With the design space selected, we performed sequence design with ProteinMPNN, predicted the structures with AlphaFold2 (*15*), and filtered by predicted local distance difference test score (pLDDT) and Cα root mean square deviation (RMSD) to the input structure (Fig. 1).

2.3 – Design of myoglobin variants with increased stability

We first applied our design strategy to the model protein myoglobin. Myoglobin binds heme to carry oxygen in mammalian muscle tissue (*39*) and has relevance in clinical applications as a biomarker (*40*), as a versatile platform for biocatalytic applications (*41–43*), and in food science as an ingredient in artificial meat products (*44*, *45*). Current efforts to create more stable variants of myoglobin have focused on the stabilization of the globin fold through stapling with cysteine-reactive noncanonical amino acids (*46*, *47*).

We applied our ProteinMPNN design protocol described above starting from a crystal structure of human myoglobin, nMb (PDB: 3RGK) (*48*). To preserve the oxygen storage function, we fixed the identities of 17 positions located around the heme ligand in the heme-bound structure (Fig. 2A). 60 sequences were generated with ProteinMPNN and evaluated for their likelihood to recapitulate the myoglobin backbone coordinates using AlphaFold2 single-sequence predictions (see methods). Eight of the designs did so with high confidence (pLDDT > 85.0 and Cα RMSD < 1.0 Å; analogous single-sequence prediction of the native sequence yielded pLDDT = 50.6 and Cα RMSD = 7.5 Å). Four designs with close structural agreement in the heme-binding region were selected for experimental testing.

We also explored limited backbone redesign of poorly ordered regions to attempt to further stabilize the protein. The globin superfamily, of which myoglobin is a member, has a fold made up of eight alpha helical regions, with diversity in the termini and two loop regions flanking the heme-binding pocket (*49–51*) (fig. S1). We selected these less-conserved loop regions for backbone remodeling with RoseTTAFold joint inpainting (Fig. 2A) (*52*). We generated two distinct sets of designs with structural remodeling: one with the region joining helices E and F redesigned, and one additionally including the CD-loop region (Fig. 2A). From these remodeled backbones, we again performed sequence design with ProteinMPNN with the heme-binding site kept fixed as described above. Following filtering on structure prediction metrics (fig. S2), additional 16 sequences were selected for experimental testing.

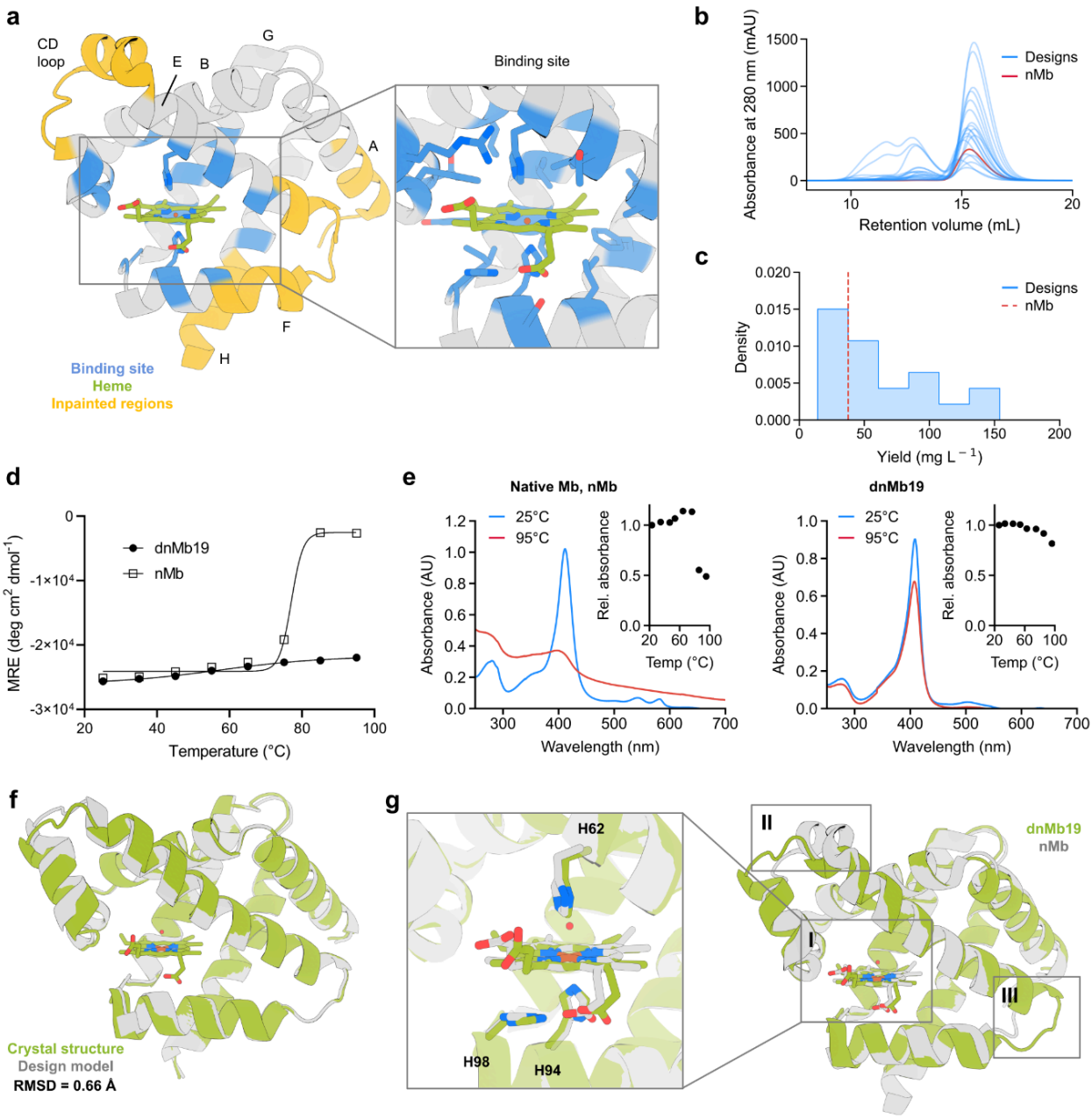

Figure 2. ProteinMPNN design improves myoglobin expression and thermostability. (**A**) Positions adjacent to the heme were kept fixed during sequence design (shown in blue). Non-conserved regions (in yellow) were subjected to backbone remodeling. Inset shows the heme-binding site. (**B**) SEC traces of 20144 designed myoglobin variants. (**C**) Soluble yield of myoglobin designs and native myoglobin nMb (represented as a black dashed line). (**D**) CD melting temperature plots of dnMb19 compared to native myoglobin (signal reported in molar residue ellipticity (MRE)). (**E**) Absorbance plots of dnMb19 and native myoglobin (inset shows temperature scan). (**F**) Structural alignment of the crystal structure (green) and AlphaFold2 (AF2) prediction (gray) of dnMb19. (**G**) Overlay of the crystal structure of native myoglobin (gray) and the crystal structure of dnMb19 (green, PDB: 8U5A). Non-conserved regions displayed in insets **II** and **III** were subjected to backbone redesign.

All 20 tested myoglobin designs have 41-55% sequence identity to the most similar protein (a myoglobin in all cases) in the UniRef100 database (Table S1) (*53*).

Synthetic genes encoding the designs and the parent sequence, nMb, were expressed in E. coli. The heme-loaded holo-proteins were purified via immobilized metal affinity chromatography (IMAC) and size exclusion chromatography (SEC). All designs were solubly expressed and monomeric by SEC (Fig. 2B). 13 of the 20 designs had higher levels (up to 4.1-fold increase) of total soluble protein yield (Fig. 2C). All 20 designs had similar heme-binding spectra to native myoglobin, with agreement in Soret maximum (407-413 nm vs 409 nm in native) and Q band features (500, 537, 582 and 630 nm), suggesting preservation of the native heme-binding mechanism (fig. S3).

The thermal stabilities of eight highly-expressing designs (six and two designed with and without backbone remodeling, respectively) were evaluated using circular dichroism (CD) spectroscopy. All eight designs had higher melting temperatures than native myoglobin, with six remaining fully folded at 95 °C (native myoglobin melts at 80 °C; Fig. 2D and fig. S4). Heme binding was also evaluated over a temperature gradient to determine functional thermal stability. All designs preserved heme binding at higher temperatures than native myoglobin (as monitored by changes to Soret band wavelength and intensity in the UV/Vis spectrum), with five designs maintaining significant heme-binding at 95 °C (fig. S5). One of the five designs, **dnMb19**, generated with the more aggressive backbone remodeling strategy, showed much higher thermal stability of heme binding compared to native myoglobin (Fig. 2E). Overall, remodeling regions of the myoglobin backbone with inpainting increased the success rate for retaining heme-binding at elevated temperatures.

To understand the structural basis of these improvements in stability, we solved the crystal structure of **dnMb19** (2.0 Å resolution, PDB: 8U5A). We found that it closely agreed with the AlphaFold2 prediction (0.66 Å Cα RMSD, Fig. 2F), including the regions remodeled with inpainting. Native sidechain contacts with the heme group are largely preserved in **dnMb19** (Fig. 2G, inset I). Outside of the heme-binding site, the crystal structure confirms the structural changes introduced by inpainting: the C and E helices were elongated as designed and connected by a new loop (Fig. 2G, inset

II); the loop connecting the E and F helices has a new conformation, and the F helix was straightened through the replacement of PRO88 with GLU89 (Fig. 2G, inset III). The Cα RMSD over the inpainted regions between the crystal structure and the AlphaFold2 model is 0.88 Å, with the largest deviation being in the CD-loop region (1.51 Å). These results illustrate the power of RoseTTAFold joint inpainting and ProteinMPNN to accurately remodel native protein backbones while increasing solubility, thermostability, and functional stability.

2.4 – Design of TEV protease variants with improved stability and catalytic activity

To explore the utility of ProteinMPNN sequence design for stabilizing enzymes, we next applied our design strategy to the cysteine protease from tobacco etch virus (TEV). TEV protease is widely used in biotechnological applications to specifically cleave between glutamine and serine in its recognition sequence (ENLYFQ/S) to remove purification tags from recombinant proteins. However, it is often difficult to use TEV protease due to its minimal soluble yield, low thermostability, and poor catalytic activity. These properties often necessitate long incubation times and result in incomplete cleavage (*54*).

We applied our sequence design strategy to TEV protease starting from an autolysis-resistant S219D variant, TEVd (PDB: 1LVM) (*55*). We defined the active site residues as described above to fix during redesign. We additionally fixed the amino acid identities of residues that are most conserved within the protein family (determined from a sequence alignment generated against UniRef30 (*53*)), as residues distant from the active site can contribute significantly to function (*56*). We ranked each amino acid identity at each position by degree of conservation in the sequence alignment and varied the percentage of these most highly conserved residues to fix during sequence redesign between 30-70%. We generated four distinct sets of designs that fixed the amino acid identities of just the active site residues, or the active site residues and 30%, 50%, and 70% of the most conserved residues in the TEV family (Fig. 3A, see Methods). 144 sequences were generated with ProteinMPNN which were all predicted with high

confidence to fold to the TEV structure by AlphaFold2 (pLDDT > 87.5; native TEV is predicted with pLDDT = 90) and possess 55% to 85% sequence identity to the parent sequence. All 144 designs were selected for experimental testing.

Synthetic genes encoding the designs, the parent sequence, TEVd, and several previously reported TEV variants were expressed in *E. coli*, and the resultant proteins were purified via IMAC and SEC. 134 of 144 designs expressed solubly and were monomeric by SEC (Fig. 3B). 129 of 144 designs exhibited higher levels of soluble expression than TEVd (TEVd average yield = 1 mg / L culture, designs average yield = 20.1 mg / L culture (Fig. 3F)).

We evaluated catalytic activity using a previously described (*57*) coumarin derivative with 7-amino-4-trifluoromethylcoumarin conjugated to the C-terminus of the substrate peptide Ac-ENLYFQ (fig. S7A). Purified protein was incubated with the peptide-coumarin substrate and 64 designs displayed progress curves with fluorescence above background, indicating substrate turnover (fig. S7B and S7C). Designs made with no evolutionary constraints had improved soluble expression over the parent but were not active on the peptide substrate, while designs with the highest activities were designed with the top 50% most conserved residues fixed (Fig. 3F and 3G). We performed detailed kinetic analysis of three highly active designs from the 50% method – **hyperTEV56**, **hyperTEV60**, and **hyperTEV89** – and the parent sequence TEVd. The designs displayed improved catalytic efficiencies (k_{cat}/K_m) compared to TEVd, with up to 26-fold improvements (Table 1 and fig. S8).

Next, we tested the most active designs with a fusion protein substrate to assess performance on the target application of tag removal. The designs and a set of previously engineered TEV proteases (*8*, *54*, *55*, *58*, *59*) were incubated at 30 °C with the fusion protein substrate MBP-TEVcs-FKBP-EGFP, where MBP is maltose-binding protein, TEVcs is the TEV peptide cleavage site (ENLYFQS), FKBP is FK506-binding protein, and EGFP is enhanced green fluorescent protein. The extent of proteolysis was evaluated by monitoring the accumulation of cleaved product via sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (fig. S9). Two designs, **hyperTEV56** and **hyperTEV60**, exhibited significantly higher rates of cleavage of protein substrate compared to the parent TEVd, yielding 50%

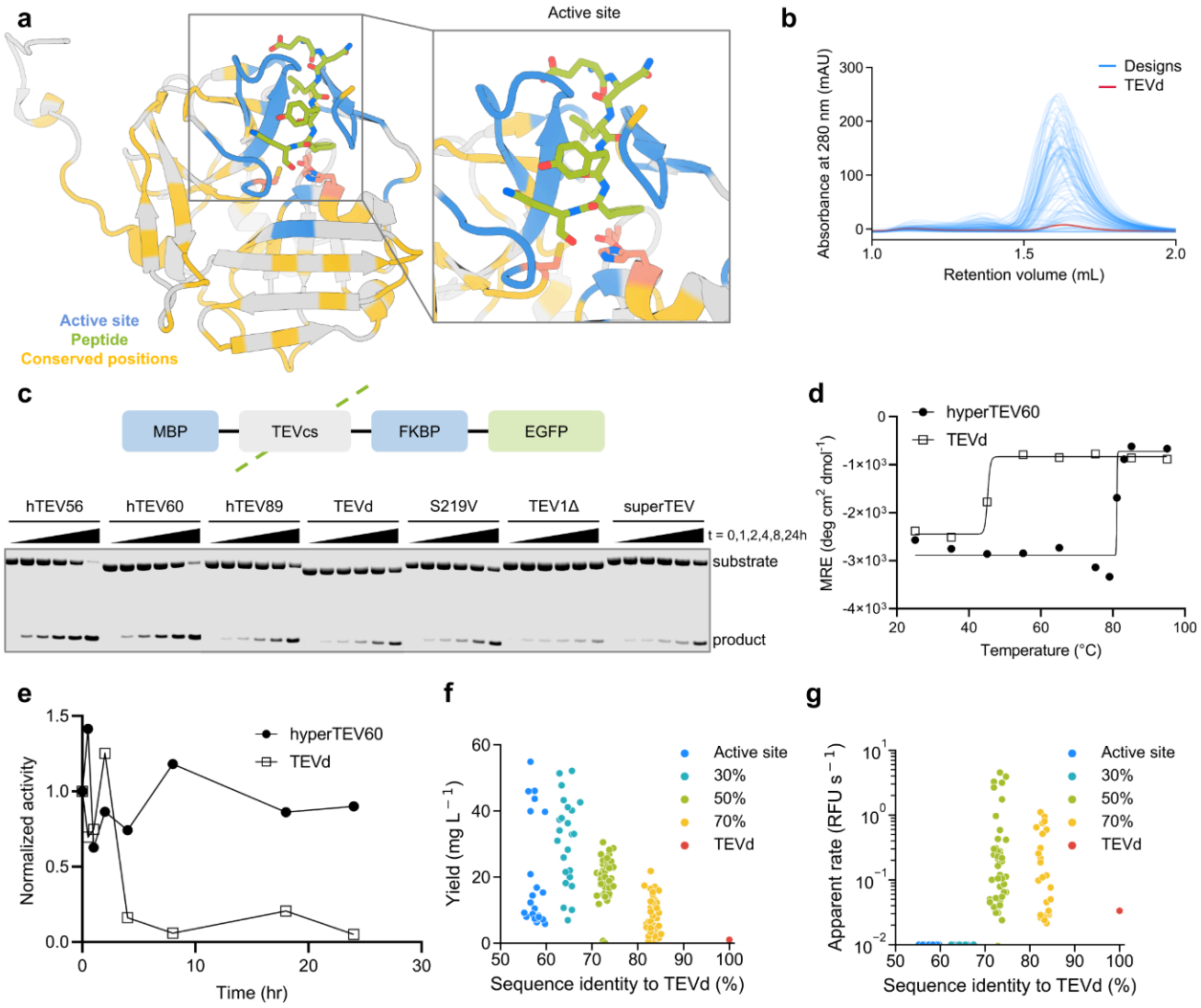

Figure 3. ProteinMPNN sequence design improves TEV protease expression, thermostability, and catalytic efficiency. (**A**) TEVd (PDB: 1LVM) input structure with positions fixed during redesign highlighted. Active site residues surrounding the substrate (blue), and 50% most highly conserved residues (yellow), and catalytic residues (pink) are highlighted. Inset shows zoom-in of the active site region. (**B**) SEC traces of designed TEV variants. (**C**) Diagram of TEV substrate (top) and fluorescent gel image of TEV cleavage reactions at various time points (bottom). (**D**) CD melting temperature plots of designed and native TEV (signal reported in molar residue ellipticity (MRE)). (**E**) Benchtop stability comparison of native TEVd and designed variant assessed as activity measured over time incubated at 30 °C before inclusion in assay. (**F**) Decreased evolutionary constraints correlate with higher soluble expression levels. Legend indicates regions fixed during design (all designs have active site fixed). (**G**) Designs made with the active site and 50% most conserved residues fixed during design exhibited highest catalytic activity. Raw apparent rate reported in relative fluorescence units (RFU) per second.

---

cleaved product at ~ 4 hours of incubation while TEVd required 24 hours to reach an equivalent yield.

The designs also outperformed other published TEV variants, with 30% turnover for superTEV, 15%

turnover for TEV1Δ, and 50% turnover for S219V at 24 hours of incubation (Fig. 3C and fig. S10A). Straight-line fit of product accumulation and substrate depletion reveal catalytic efficiencies that corroborate those determined in the peptide assay (fig. S10B). The gains in catalytic efficiency are primarily due to increases in k_{cat}, which could reflect a higher fraction of enzyme in a catalytically competent state (see below).

Analysis by CD spectroscopy of TEVd and the most active design, **hyperTEV60**, indicated an approximate melting temperature of 84 °C for **hyperTEV60**, 40 °C higher than that of TEVd (Fig. 3D and fig. S11), and to our knowledge, higher than any previously described TEV variant. To further probe stability of the designed variant, TEVd and **hyperTEV60** were incubated at 30 °C for various times and then used in the peptide-coumarin cleavage assay. After 4 hours of incubation, **hyperTEV60** retained 90% of its original cleavage activity while TEVd was reduced to 15% of its original activity (Fig. 3E), a significant improvement in benchtop stability.

Given that catalytic and substrate-binding residues were kept fixed during design with ProteinMPNN, it is notable that significant improvements in k_{cat} were observed with both the peptide and protein substrates. Mutations distal to the active site can influence catalytic activity through stabilization of catalytically productive conformational states (*60*, *61*) or global conformational changes (*62*). To investigate if stabilization of functional conformational states may be involved in activity enhancement, we performed microsecond molecular dynamics (MD) simulations on TEV-peptide complexes to probe the impact of the introduced mutations on overall protein dynamics. A general rigidification of loop regions distributed across the structure was observed in designs as compared to TEVd (fig. S12A). This backbone rigidification in distal regions not directly involved in substrate binding may be related to allosteric improvement of substrate binding as reflected by the 2- to 3-fold lower K_m values measured for the designed variants (Table 1). Rigidification in the region spanning residues 115 to 124 appeared to correlate with activity; the highest activity design, **hyperTEV60**, was most rigid, while TEVd and a design with no activity on the peptide substrate were most flexible in this region (fig. S12B). These trends were also observed in per-residue pLDDT analysis of AlphaFold2 ensemble predictions (fig. S12C). In all

Variant	k_{cat} (min^{-1})	K_m (μM)	k_{cat}/K_m (μM^{-1} min^{-1})	Fold-improvement in k_{cat}/K_m over parent
hyperTEV56	0.0106 ± 0.0005	1.4 ± 0.2	0.0077	20
hyperTEV60	0.014 ± 0.002	1.4 ± 0.4	0.01	26
hyperTEV89	0.0050 ± 0.0001	2 ± 1	0.0024	6.2
TEVd	0.0023 ± 0.0003	6 ± 3	0.00039	

Table 1. Kinetic parameters for TEV redesigns and parent TEV variant. Kinetic parameters derived from the cleavage assay with the fluorescent peptide-coumarin substrate in figure S7A. Uncertainties are standard deviations of values calculated from fitting three technical replicates.

designs, we observed a decrease in the population of catalytically competent conformations of the Cys-His dyad (dN-SH) compared to TEVd, but this shift was least significant in **hyperTEV60**, in agreement with its higher relative k_{cat} (fig. S13). These notable differences may begin to explain how ProteinMPNN enables substantial activity enhancements without explicit design elements to improve function. It is also possible that the major contribution to the increase in k_{cat} is from an increase in the fraction of the protein in the catalytically competent state more globally.

2.5 – Conclusion

We show that the expression, stability, and function of native proteins can be improved using ProteinMPNN guided by available sequence and structural information. For both TEV protease and human myoglobin, multiple variants were identified which showed higher soluble yield and thermostability than the native protein starting point. The best of the TEV protease designs have higher apparent catalytic efficiency on peptide and protein substrates than the parent enzyme and previously reported variants. While the optimal number of residues to conserve to maintain (and perhaps enhance) function may have to be determined empirically for each case, the simplicity of our procedure and the compute efficiency and ease of use of ProteinMPNN make this straightforward, and the number of variants that need to be tested is far smaller than in typical experimental screens. We expect that our

approach should be widely useful for improving the expression, stability, and function of biotechnologically important proteins.

2.6 – Methods

Fixed residue selection for TEV protease. Active site positions were defined as residues containing backbone atoms within 7 Å of the substrate or sidechain atoms within 6 Å of the substrate in the ligand-bound crystal structure of autolysis resistant S219D (PDB: 1LVM). For enzyme targets, highly conserved residues were also fixed during sequence redesign. Highly conserved residues were determined with multiple sequence alignments (MSA). To generate the MSA, four iterative HHblits searches (*63*) were performed against the UniRef30 database (accessed June 30, 2020) at E-value cutoffs of 1e-50, 1e-30, 1e-10, and 1e-4, and the final result was filtered for 90% identity redundancy, 50% coverage, and 30% minimum query identity. Within the sequence alignment, we identified the frequency of each amino acid at each position and found the most highly conserved amino acid identity at each position. We then ranked each position by how highly conserved the most frequent amino acid identity was, and selected the top 30%, 50%, and 70% most conserved positions to fix during sequence design.

ProteinMPNN design of myoglobin. For fixed-backbone sequence redesign, the crystal structure of human myoglobin (PDB: 3RGK) was used as input to ProteinMPNN, and 17 positions located around the heme were excluded from design. Three temperatures (0.1, 0.2, and 0.3) were sampled, with 20 sequences generated per temperature. Cysteine and methionine were excluded from the amino acid identities that could be installed during design. A model of ProteinMPNN trained with 0.2 Å noise applied to training set protein backbones was used to perform sequence generation. For combined sequence and backbone redesign, two strategies were employed. First, the sequence and structure from the crystal structure of human myoglobin (PDB: 3RGK) were input to RFjoint, with the N- and C-termini and loop region between helices 5 and 6 masked, to generate new secondary structure in these regions (RoseTTAFold joint inpainting). Ten backbones were generated with this strategy. In a more aggressive strategy, helix 4

and its adjoining loops, as well as both termini and the loop joining helices 5 and 6, were masked. Twenty backbones were generated with this strategy. Following backbone redesign, 60 sequences were generated per backbone with ProteinMPNN, keeping heme-binding positions fixed as described above.

Sequences generated with ProteinMPNN were predicted with AlphaFold2, using model 4 with 10 recycling steps. Structural templating with MSAs was not used for prediction. Designs with only sequence redesign were filtered to Cα RMSD < 1.0 Å and pLDDT > 85.0. Designs with sequence redesign and backbone redesign on the termini and the loop connecting helices 5 and 6 were filtered to Cα RMSD < 0.8 Å and pLDDT > 90.0. Designs with sequence redesign and backbone redesign on the termini, helix 4, and the loop connecting helices 5 and 6 were filtered to Cα RMSD < 0.6 Å and pLDDT > 90.0 (see fig. S2 for details). Predicted models passing these criteria were finally evaluated by eye and those recapitulating finer structural details of the heme binding pocket (low backbone deviation after global alignment to the structure of 3RGK; close agreement with the placement of heme-coordinating histidine side chain) were selected for experimental testing. Four designs generated with only sequence design, and 16 designs with sequence and backbone design were selected for experimental testing (10 with both loops remodeled, and 6 with one loop).

ProteinMPNN design of TEV protease. The crystal structure of TEVd (PDB: 1LVM) was used as structural input to ProteinMPNN, and active site and conserved residues were excluded from design. Cysteine was excluded from the amino acid identities that could be installed during design. Three temperatures (0.1, 0.2, and 0.3) were sampled during design. A model of ProteinMPNN trained with 0.2 Å noise applied to training set protein backbones was used to perform sequence generation. 24 sequences were generated with only the active site residues fixed, 24 sequences were generated with the active site and the 30% most highly conserved positions fixed, 48 sequences were generated with the active site and the 50% most highly conserved positions fixed, and 48 sequences were generated with the active site and the 70% most highly conserved positions fixed.

Sequences generated with ProteinMPNN were predicted with AlphaFold2, using model 3 with 6 recycling steps. Both designs and native TEV predicted with low confidence if given only the single sequence and minimal recycling steps. We found that structural templating with MSAs was necessary for accurate prediction. To generate MSAs of each design for structure prediction, the MSA of the parent sequence was used, and the parent sequence was swapped for the design sequence. All sequences generated were predicted with Cα RMSD < 2.0 Å and pLDDT > 85.0 and were predicted to maintain critical structural features in the active site. Thus, all were ordered for experimental characterization.

Expression and purification of myoglobin designs. Double-stranded DNA fragments encoding the designs (codon-optimized for bacterial expression) were purchased from Integrated DNA Technologies (IDT) as eBlocks™ Gene Fragments. Following the Golden Gate cloning protocol (*38*), the DNA fragments encoding design sequences and including overhangs suitable for a BsaI restriction digest were cloned into a custom pET29b(+) target vector containing lethal ccdb gene, and C-terminal SNAC (*64*) and hexahistidine tags (#191551, Addgene). This yielded final expressed sequences as: MSG<design>GSGSHHWGSTHHHHHH. Assembled plasmids containing the designs were transformed into E. coli BL21(DE3) by heat shock. DNA was incubated on ice with competent cells for 30 minutes, followed by 30 second heat shock at 42 °C, and 2 minute incubation on ice. 100 µL rich medium (super optimal broth with catabolite repression) was added to transformed cells and samples were incubated at 37 °C, 1050 r.p.m. on a Heidolph shaker for 1 hour. The cells were subsequently spread on LB-agar plates containing 100 µg/mL kanamycin and incubated at 37 °C under 220 r.p.m. shaking for 18 hours. Single colonies were picked, and the DNA fragments encoding the designs were amplified following a colonyPCR protocol using GoTaq® Green DNA polymerase master mix (#M7122; Promega) and T7 reverse and forward primers. The PCR products identified to contain DNA of appropriate size (~600 bp) based on agarose gel (1.2%) electrophoresis with SybrSafe dye were sent to Sanger sequencing (GeneWiz/Azenta) for sequence-verification. Single colonies containing the correct design sequences were grown up in 5 mL TB-II media containing 50 µg/mL kanamycin, over 16 hours at 37 °C. 2 mL of

the grown culture was used to inoculate 40 mL TB-II media containing 50 μg/mL kanamycin and the rest used for plasmid extraction following the Qiagen QIAprep MiniPrep protocol. The 40 mL cultures were grown at 37 °C for 4 hours, after which protein expression was induced with the addition of 1 mM IPTG, and the cultures were incubated at 18 °C for 20 hours. Pellets were harvested by centrifugation at 4,198 g for 8 minutes and resuspended in a lysis buffer containing 25 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, 0.01 mg/mL DNAse, 0.1 mg/mL lysozyme, and a Pierce protease inhibitor tablet. Lysis was performed by ultrasonication (13 mm probe, 2.5 mins, 10s on, 10s off, 65% amplitude). Lysate was collected by centrifugation at 15,000 xg for 20 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer (25 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0). The resin was washed with 50 column volumes (CV) of wash buffer. Protein was eluted with 1.2 CV of elution buffer (25 mM Tris-HCl, 300 mM NaCl, 300 mM imidazole, pH 8.0) and further purified via size exclusion chromatography (SEC) using a Superdex Increase 75 10/300 GL column (GE Healthcare) on ÄKTAxpress (GE Healthcare) instrument at 0.8 mL min-1 flow rate. The monomeric or smallest oligomeric fractions of each run (eluting at approximately 15 ml) were collected. The obtained chromatograms are presented in figure S6.

Yields of purified hemoproteins were determined based on the absorbance of the Soret maximum (407-413 nm). The corresponding extinction coefficients were measured for each protein using the hemochromagen assay, according to the method of Berry and Trumpower (*65*). A reported extinction coefficient of 188 $mM^{-1} \cdot cm^{-1}$ was used for native myoglobin (*66*).

Expression and purification of TEV designs. Double-stranded DNA fragments encoding the designs (codon-optimized for bacterial expression) were purchased from Integrated DNA Technologies (IDT) as eBlocks™ Gene Fragments. Following the Golden Gate cloning protocol,(*38*) the DNA fragments encoding design sequences and including overhangs suitable for a BsaI restriction digest were cloned into a custom pET29b(+) target vector containing lethal ccdb gene, and C-terminal SNAC(*64*) and hexahistidine tags (#191551, Addgene). This yielded final expressed sequences as:

MSHHHHHHSG<design>GS. Vectors containing TEV designs were transformed into E. coli BL21(DE3) by heat shock. DNA was incubated on ice with competent cells for 30 minutes, followed by 10 second heat shock at 42 °C, and 2 minute incubation on ice. 100 µL rich medium (super optimal broth with catabolite repression) was added to transformed cells and samples were incubated at 37 °C, 1050 rpm on a Heidolph shaker for 1 hour. Entire transformations were transferred to 900 µL of TBM-5052 autoinduction expression medium containing 50 µg/mL Kanamycin. Expression cultures were incubated at 37 °C, 1050 rpm for 20 hours. Pellets were harvested by centrifugation at 4,000 g for 10 minutes and lysed with BPER lysis reagent containing 6.25 Units/mL benzonase (4 uL / 40 mL at 250 U/µL), 0.1 mg/mL lysozyme, and 1 mM PMSF. Lysate was collected by centrifugation at 4,000 xg for 20 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0). The resin was washed with 25 column volumes (CV) of wash buffer. Protein was eluted with 250 µL of elution buffer (20 mM Tris-HCl, 300 mM NaCl, 540 mM imidazole, pH 8.0) and further purified via size exclusion chromatography (SEC) in an S75 5/150 GL increase column (GE Healthcare). Protein collected from SEC was normalized to 1 µM where possible.

In scale-up experiments, 50-mL cultures of TBM-5052 autoinduction media with 50 µg/mL Kanamycin were inoculated with a scrape of transformed competent cells from glycerol stock and grown at 37 °C, 200 rpm for 20 hours. Cells were harvested by centrifugation at 10,000 xg for 10 minutes, resuspended in 30 mL of wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0) containing 0.01 mg/mL DNase, 0.1 mg/mL lysozyme, and a protease inhibitor tablet (Thermo Scientific Pierce), and lysed by sonication. Lysate was collected via centrifugation at 18,000 xg for 40 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer. The resin was washed with 30 CV of wash buffer. Protein was eluted with 4 mL of elution buffer and concentrated to 1 mL in a 3K protein concentrator (Millipore Sigma). Concentrated protein was purified by SEC as described above.

Expression and purification of MBP-TEVcs-FKBP-EGFP construct. The protease substrate FKBP-EGFP was cloned into an E. coli expression vector containing an N-terminal maltose binding

protein (MBP), a TEV protease recognition site, and a C-terminal His-6 tag. The FKBP-EGFP coding sequence was obtained from Addgene #106924, with a 4X GGS linker between FKBP and EGFP. Vector containing the protease substrate was transformed into E. coli BL21(DE3) by heat shock. Cells were transferred to 4 0.5-L LB medium cultures with 10 μg/mL Carbenicillin and 10 μg/mL Chloramphenicol and incubated at 37 °C, 200 rpm until optical density reached 0.5 AU, at which point expression was induced with 1 mM IPTG. Temperature was reduced to 18 °C and cells were incubated for an additional 18 hours. Cells were harvested by centrifugation at 10,000 xg for 10 minutes, resuspended in 30 mL of wash buffer (20 mM Tris-HCl, 300 mM NaCl, 25 mM imidazole, pH 8.0) containing 0.01 mg/mL DNase, 0.1 mg/mL lysozyme, and a protease inhibitor tablet (Thermo Scientific Pierce), and lysed by sonication. Lysate was collected via centrifugation at 18,000 xg for 40 minutes and applied to Ni-NTA resin that was equilibrated with wash buffer. The resin was washed with 30 CV of wash buffer. Protein was eluted with elution buffer until resin no longer appeared yellow and concentrated to 1 mL in a 3K protein concentrator (Millipore Sigma). Concentrated protein was purified by SEC as described above.

Kinetic characterization of designed proteases. Designs were initially screened for activity on a peptide-coumarin conjugate substrate (WuXi) of the TEV recognition sequence (ENLYFQ) fused to a fluorescent coumarin derivative, 7-amino-4-trifluoromethylcoumarin. The N-terminus of the peptide bears an acetyl modification and the C-terminus is conjugated to the coumarin group via an amide bond. Initial activity screen was performed in 50 mM Tris-HCl, 50 mM NaCl, pH 8.0 buffer containing Freshly prepared 2 mM DTT. Reactions contained 500 nM protein and 10 μM substrate at a total volume of 30 μL. Protein and substrate were rapidly mixed and monitored for fluorescence at excitation 400 nm, emission 492 nm at room temperature (RT) for 5 hours in a BioTek Synergy Neo2 microplate reader.

For detailed kinetic characterization, reactions were performed in 50 mM Tris-HCl pH 8.0 containing 50 mM NaCl, 1mM EDTA, and freshly prepared 2 mM DTT. For TEV redesigns, reactions contained 50 nM protein and substrate concentration ranging from 0.1 μM to 10 μM at a total volume of 30 μL. Protein and substrate were rapidly mixed and monitored for fluorescence at excitation 400 nm,

emission 492 nm at RT for 2 hours in a BioTek Synergy Neo2 microplate reader. Fluorescent signal was converted to concentration of cleaved coumarin product using a calibration curve of 7-amino-4-trifluoromethylcoumarin. Reactions were performed in triplicate and each technical replicate was separately fitted to a Michaelis Menten model. Expressed uncertainty in k_{cat} and K_m is the standard deviation between technical replicates.

Screening of designed proteases on fusion protein MBP-TEVcs-FKBP-EGFP. Reactions were performed in 50 mM Tris-HCl, 50 mM NaCl, 1mM EDTA, pH 8.0 buffer containing freshly prepared 2 mM DTT. Reactions contained 60 nM protein and substrate concentrations ranging from 2 μM to 17 μM. Reactions were incubated at 30 °C and at 0, 1, 2, 4, 8, and 24 hours, 10 μL aliquots were quenched in 10 μL of 2X Laemmli loading buffer and subsequently frozen in liquid nitrogen. Samples were analyzed by SDS-PAGE and imaged for EGFP fluorescence at 488 nm on a LI-COR Odyssey M imager. Band intensities were quantified with ImageJ software and converted to concentration using a standard curve prepared of known amounts of cleaved substrate with fluorescence gel imaging. A straight-line fit was applied to the initial velocities using GraphPad Prism. Points represent the averages of 3 technical replicates and error bars represent the standard deviations.

Benchtop stability characterization of TEV redesigns. Samples of purified enzyme were incubated at 30 °C for 0.5, 1, 2, 4, 8, 18, or 24 hours before being used in the previously described peptide-coumarin cleavage assay. Activity of samples was defined as initial rate of turnover and normalized to initial rate at incubation of t = 0 hrs.

Spectrophotometric measurements. UV-Vis spectra of purified holo-proteins (myoglobin variants) in the 230-800 nm range were collected using the Jasco Spec V750 spectrophotometer and 10 mm pathlength cuvette. To observe changes in the spectral properties of bound heme at increasing temperatures, UV−Vis spectra were collected at every 10 °C intervals between 25 °C and 95 °C. Temperature was increased at the rate of 5 °C min-1, and spectra were acquired after the temperature had

stabilized to within 0.5 °C of target temperature for 5 seconds. Measurements were performed with 20 μM solutions of purified holoprotein in TBS buffer (25 mM Tris-HCl, 300 mM NaCl, pH 8).

Circular dichroism spectroscopy. To determine secondary structure and thermostability of the designs, far-ultraviolet circular dichroism (CD) measurements were carried out on a JASCO J-1500 instrument using a 1 mm pathlength cuvette. Samples of purified protein were prepared at 1.0 mg/mL in 20 mM sodium phosphate, 50 mM potassium fluoride, pH 8.0 (TEV protease) or at 0.4 mg/mL in 25 mM Tris, 20 mM NaCl, pH 8.0 (myoglobin). The temperature of the sample was scanned from 25 °C to 95 °C with full spectrum scans from 190 nm to 260 nm performed after each 10 degree increment. The signal at 216 nm was plotted over the temperature gradient and fitted to a Boltzmann sigmoidal curve with GraphPad Prism 9. T_m values were calculated from the inflection point.

Mass spectrometry analysis. MS data for dnHEM1 variants were acquired on an Agilent 1200series LC G6230B TOF LC-MS with an AdvanceBio RP-Desalting column (A: H2O with 0.1% Formic Acid, B: Acetonitrile with 0.1% Formic Acid). The final protein concentrations were adjusted to 1-2 mg/mL in 25 mM Tris-HCl, 300 mM NaCl, pH 8.2. Subsequent data deconvolution was performed in Bioconfirm using a total entropy algorithm. All data are presented in Supplementary Table 1.

Molecular dynamics simulations. Structures generated with AlphaFold2 (*15*) were used as starting geometries. For the protein-substrate complexes, substrate peptide was superimposed onto AlphaFold2 structures using the crystallographic structure of catalytically active TEV protease (PDB: 1LVM) as a template. Simulations were carried out with AMBER 20 (*67*) implemented with the ff14SB force field for the protein and substrate peptide, and the general Amber force field (GAFF2) (*68*) for the substrate peptide C-terminal fluorescent probe (7-amino-4-(trifluoromethyl)coumarin). Parameters were generated with the antechamber module of AMBER, combining ff14SB and GAFF2 force fields and with partial charges set to fit the electrostatic potential generated with HF/6-31G(d) using the RESP method (*69*). The charges were calculated according to the Merz-Singh-Kollman scheme using Gaussian 16 (*70*).

Binding-site histidine residue (H46) was modeled in its Nδ1-H tautomeric state (corresponding to residue name HID in Amber). Initial structures were neutralized with either Na^+ or Cl^- ions and set at the center of a cubic TIP3P (*71*) water box with a buffering distance between solute and box of 10 Å.

A two-stage geometry optimization approach was performed. The first stage minimizes only the positions of solvent molecules and ions, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The system was then heated by incrementing the temperature from 0 to 300 K under a constant pressure of 1 atm and periodic boundary conditions (PBC). Harmonic restraints of 10 kcal/mol were applied to the solute, and the Andersen temperature coupling scheme (*72*, *73*) was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Water molecules were treated with the SHAKE algorithm (*74*) such that the angle between the hydrogen atoms is kept fixed through the simulations. Long-range electrostatic effects were modeled using the particle mesh Ewald method (*75*). An 8 Å cut-off was applied to Lennard-Jones interactions. The system was equilibrated for 2 ns with a 2 fs time step at a constant volume and temperature of 300 K. Ten independent production trajectories were then run for additional 1000 ns under the same simulation conditions, leading to accumulated simulation times of 10 μs for each system. Root mean square (rms) fluctuations and interatomic distance analyses were carried out with the cpptraj module of AMBER.

Chapter 3: Computational design of serine hydrolases

Note: The majority of this chapter is borrowed directly from the manuscript of the same name which was published in *Science* in 2025. I, Kiera H. Sumida, am a co-lead author of the study.

Authors: Anna Lauko[1,2,3,†], Samuel J. Pellock[1,2,†,*], Kiera H. Sumida[1,2,5†], Ivan Anishchenko[1,2], David Juergens[1,2,4], Woody Ahern[1,2,7], Jihun Jeung[1,2], Alexander F. Shida[1,2], Andrew Hunt[1,2], Indrek Kalvet[1,2,6], Christoffer Norn[1,2], Ian R. Humphreys[1,2], Cooper Jamieson[8], Rohith Krishna[1,2], Yakov Kipnis[1,2], Alex Kang[1,2], Evans Brackenbrough[1,2], Asim K. Bera[1,2], Banumathi Sankaran[1,2], K. N. Houk[8], David Baker[1,2,6*]

Affiliations:

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA

[2]Institute for Protein Design, University of Washington, Seattle, WA, USA

[3]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA, USA

[4]Graduate Program in Molecular Engineering, University of Washington, Seattle, WA, USA

[5]Department of Chemistry, University of Washington, Seattle, WA, USA

[6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

[7]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA.

[8]Department of Chemistry and Biochemistry, University of California, Los Angeles, California, USA

†These authors contributed equally: Anna Lauko, Samuel J. Pellock, Kiera H. Sumida

*Corresponding author. Email: spellock@uw.edu and dabaker@uw.edu

3.1 – Introduction

Enzymes are powerful catalysts that dramatically accelerate reaction rates in mild aqueous conditions. The ability to construct enzymes catalyzing arbitrary chemical reactions would have enormous utility across a wide range of applications, and hence, enzyme design has been a long-standing goal of computational protein design (*76*). De novo enzyme design has generally started from a specification of arrangements of catalytic residues around the reaction transition state (a theozyme), and sought to identify placements of this active site in pre-existing scaffolds (*22–27*). Utilizing fixed backbones restricts how accurately the catalytic geometry can be realized and has likely limited the activities of many designed enzymes to date prior to optimization by laboratory evolution, as recent studies of designed Kemp eliminases demonstrate (*60*, *77*, *78*). A further challenge of enzyme design is the preorganization of the active site such that the catalytic functional groups are accurately positioned relative to the transition state. Achieving preorganization is especially difficult for multistep reaction mechanisms because the enzyme must preferentially stabilize multiple transition states and intermediates, and current methods to evaluate design preorganization in silico are limited by low accuracy or computational cost (*27*, *79–82*). To enable the accurate design of multistep enzymes, new methods are needed for both the generation of proteins housing a given active site, and the assessment of their structural compatibility with each step in the reaction.

Ester hydrolysis has served as a model reaction for computational enzyme design for decades (*83–88*), and justifiably so: numerous mechanisms can be utilized for ester hydrolysis, enabling a range of distinct design approaches to target this reaction, activity is easily monitored by absorbance and fluorescence with reporter substrates, and esterases are highly valuable in industrial processes, most recently for their application in plastic recycling (*12*, *89*, *90*). The textbook example of enzymatic ester hydrolysis is the double-displacement reaction mechanism employed by serine hydrolases, in which a serine nucleophile undergoes acylation to form the acyl-enzyme intermediate (AEI) that is subsequently hydrolyzed by an activated water. Despite extensive structural, mutational, and computational

characterization of the mechanism of serine hydrolases found in nature (*91–102*), de novo design efforts attempting to employ this machinery have been unsuccessful, and to our knowledge, no previous efforts have successfully constructed a serine hydrolase that extends beyond the fold space found in nature.

A major challenge in designing serine hydrolases is overcoming the stability of the AEI, the resolution of which is typically rate-limiting when activated esters are employed. Numerous previously designed enzymes and peptide-based systems inactivate or dramatically slow down after acylation (*26*, *83–85*). In addition to this chemical challenge, constructing the serine hydrolase active site combines some of the most difficult current challenges in protein design: 1) the catalytic site is very complex, requiring the scaffolding of at least four individual residues with atomic precision, a task that state-of-the-art design tools struggle to achieve (*18*), 2) the serine nucleophile requires activation by construction of intricate hydrogen bond networks, and 3) the active site must undergo subtle conformational changes throughout the multistep catalytic cycle, and while there is recent progress in multistate design (*103*, *104*), it remains challenging, particularly when the energetic difference between desired states are small.

Previous efforts to design esterases have circumvented the challenges presented by serine hydrolases by employing simpler, more easily designable active sites, leveraging nucleophiles more activated than serine, and by targeting reaction mechanisms that do not require the formation of stable covalent intermediates. For example, previously designed metallohydrolases skip the AEI by activating water to cleave esters in a single step (*86*, *105*), the non-canonical amino acid N_δ-methylhistidine has been employed to make the AEI less stable (*85*), and cysteine has been used in place of serine due to its greater nucleophilicity (*26*, *83*). Structural analysis of the resulting cysteine esterases indicated key interactions between the cysteine nucleophile and histidine base of the desired dyad or triad were not formed (*26*, *83*), suggesting that the inherent chemical reactivity of the residues employed, not their coordinated effort, may have been responsible for the observed steady-state rate enhancements. Even with these chemical interventions, the efficiency of the initial computational designs remain far below the range observed for natural enzymes.

One hypothesis for the lack of designed serine hydrolases to date is a potential geometric incompatibility between the complex hydrolase active site and the sets of fixed protein scaffold libraries previously employed (*26*). We investigated whether increasing scaffold diversity could help identify backbones that more accurately reconstruct the desired active site, and carried out a preliminary design campaign searching for placements of a serine hydrolase active site in a large library of scaffolds based on the Nuclear Transport Factor 2 (NTF2) fold (*28*) (fig. S1 and Computational methods, NTF2 design campaign). As in previous studies (*27*), experimental characterization of the resulting designs revealed activated serines but no catalytic turnover on ester substrates, despite a close match between the experimental and designed structures (fig. S2). We suspect that an inability to install key catalytic features into NTF2s, such as the backbone oxyanion hole contact common to all serine hydrolases, limited the function of these designs.

We reasoned that advances in deep learning for protein design could enable the design of proteins from scratch to directly scaffold the serine hydrolase active site and assess design compatibility for the entire multistep catalytic cycle. Recent advances in scaffolding functional sites with RFdiffusion have yielded improved *in silico* and experimental success rates across a range of design tasks (*18*, *106*, *107*); we aimed to use the same approach to generate serine hydrolases starting from geometric descriptions of an active site (Fig. 1A). To assess preorganization and functional interactions in each step of the catalytic cycle, we sought to leverage advances in deep learning-based prediction of protein-small molecule complexes by modeling structural ensembles of catalytic intermediates (Fig. 1B).

3.2 – Assessing reaction path compatibility with PLACER

We set out to understand why previously designed serine hydrolases failed to appreciably catalyze ester hydrolysis and hypothesized that modeling each step of the reaction could be critical for assessing the ability of a design to achieve catalytic turnover. To model the extent to which a designed enzyme can stabilize each of the key states along the reaction coordinate and to assess the preorganization of the

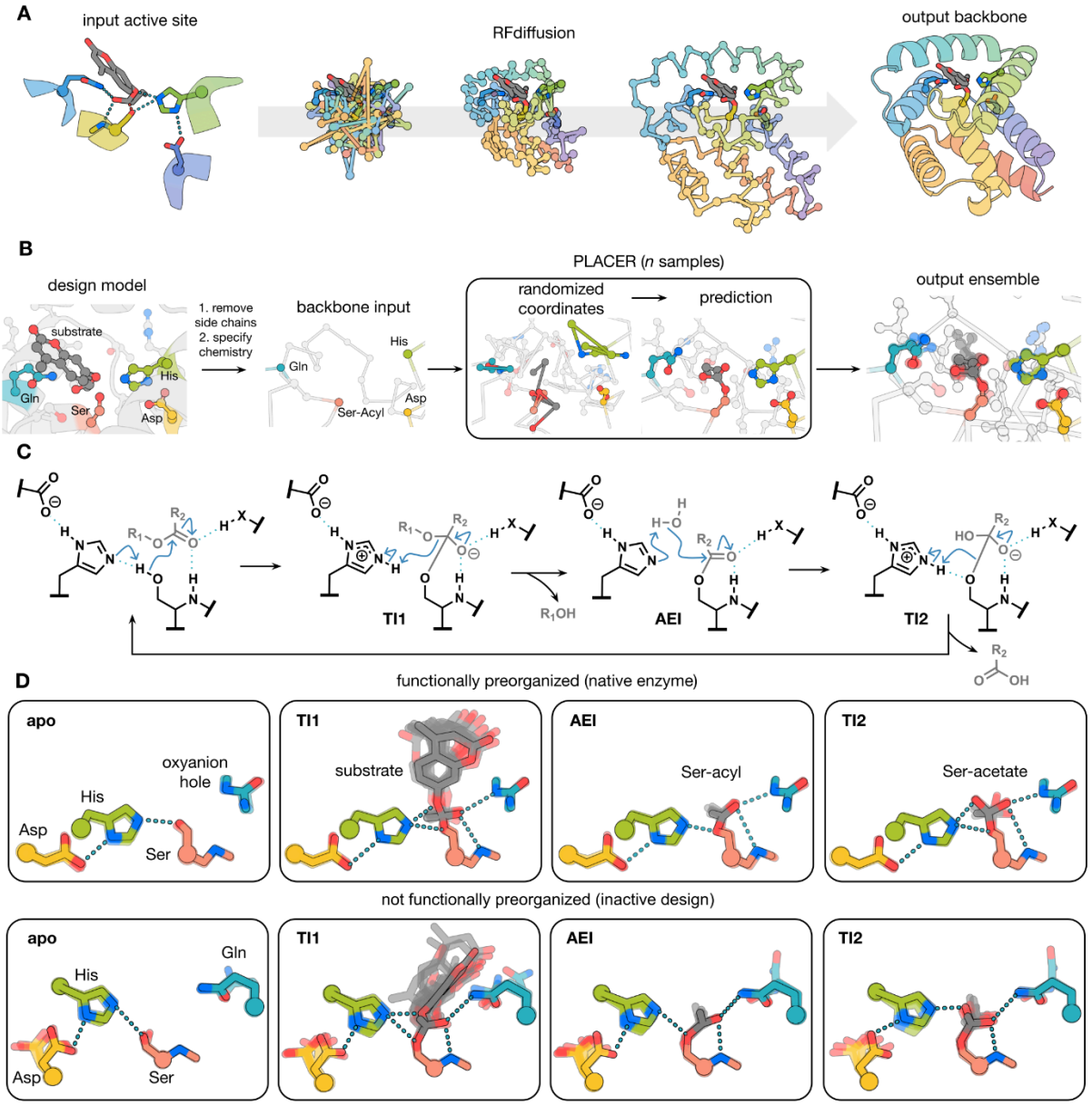

Figure 1. Design methods. (**A**) Active site-specific backbone generation with RFdiffusion. Given the geometry of a possible active site configuration, RFdiffusion denoising trajectories generate backbone coordinates which scaffold the site. (**B**) Generation of active site ensembles with PLACER. The coordinates of the sidechains around the active site and any bound small molecule for the step in the reaction being considered are randomized, and n samples are carried out to generate an ensemble of predictions. (**C**) Mechanism of ester hydrolysis by serine hydrolases. (**D**) PLACER ensembles for distinct states along the reaction coordinate for hydrolysis of 4MU-Ac for a native serine hydrolase (top, PDB: 1IVY) and an inactive designed serine hydrolase from round 3 (bottom, **josie**).

---

active site residues in the desired catalytic geometries, we developed a deep neural network that, given 1) the backbone coordinates of a small molecule binding pocket or active site, 2) the identities of the amino

acid residues at each position, and 3) the chemical structures of bound small molecules (but not their positions), generates the full atomic coordinates of the binding site, comprising both protein sidechains and small molecules. We trained this network, called PLACER (Protein-Ligand Atomistic Conformational Ensemble Reproduction) (*108*), on protein-small molecule complexes in the PDB by randomizing the atomic coordinates of sidechains and small molecules within spherical regions with up to 600 heavy atoms, and seeking to minimize a loss function assessing the recapitulation of the atomic coordinates within the region. In benchmark tests, PLACER predicted regions within native structures with an average RMSD of 1.1 Å. PLACER is stochastic, and repeated runs from different random seeds yield an ensemble of models for the predicted region (Fig. 1B).

We used PLACER to generate structural ensembles for each step of the catalytic cycle for a set of native and previously designed serine hydrolases. The catalytic cycle of serine hydrolases can be divided into four steps (Fig. 1C). First, the substrate binds to the apoenzyme (apo) and the catalytic serine, deprotonated by the catalytic histidine, attacks the carbonyl carbon of the ester to form the first tetrahedral intermediate (TI1). Second, the catalytic histidine protonates the leaving group oxygen promoting its departure, leaving the active site serine covalently linked to the acyl group of the substrate (the acyl-enzyme intermediate (AEI) mentioned above). Third, the histidine deprotonates a water molecule, which attacks the AEI to generate a second tetrahedral intermediate (TI2). Finally, this intermediate is resolved by histidine-mediated protonation of serine and release of the acyl group, reconstituting the free enzyme and completing the catalytic cycle. Throughout, negatively charged transition states and intermediates are stabilized by at least two hydrogen bond donors that constitute the oxyanion hole. Perturbation of the histidine pK_a, which tunes its acid/base function, is mediated by interaction with aspartate or glutamate, the final residue in the triad (*109–111*).

Modeling this catalytic cycle with PLACER showed that native serine hydrolases are more preorganized than previous designed systems (Fig. 1D, fig. S3). At each step in the reaction coordinate, the catalytic residues sample the key hydrogen bonds essential for catalysis more often in native than previously designed serine hydrolases (fig. S3). Since the reaction rate should be proportional to the

fraction of the enzyme in the active state, limited preorganization of the designed active sites is expected to compromise catalysis. To quantify the extent of active site formation in PLACER ensembles, we compute the frequency of formation of key interactions between the catalytic functional groups and reaction intermediates over each step of the reaction (see Computational methods, filtering section), and use this metric to assess new designs in the following sections.

3.3 – Design and characterization of serine hydrolases

We next set out to design proteins with active sites of increasing complexity, using RFdiffusion to scaffold serine hydrolase active site motifs and PLACER to assess their preorganization in each step of the reaction (Fig. 2A,B). We designed catalysts for the hydrolysis of 4-methylumbelliferone (4MU) esters (Fig. 2C) that fluoresce upon hydrolysis. To generate active site motifs, we sampled positions of the catalytic sidechains around a QM-optimized transition state (see Computational methods, motif generation) based on an analysis of natural hydrolases (*100*), and enumerated α-helix and β-strand backbone conformations for each catalytic residue, keeping the interactions with the transition state fixed in space. For each combination of the backbone N, Cα, and C atoms for each of the catalytic residues, we used RFdiffusion to build up backbones starting from random noise that have coordinates that nearly exactly match the input catalytic residue backbone positions (average all-atom RMSD ~ 0.1 Å) and form a binding pocket for the substrate (see Computational Methods, motif generation and backbone generation). To drive folding to the designed state, and to make favorable interactions with the substrate and active site residues, LigandMPNN (*112*) was used to design the sequence. Rosetta FastRelax (*113*) was used to refine the protein backbone and ligand pose, and sequence design with LigandMPNN was repeated with the new backbone as input (*114*). Following three cycles of LigandMPNN and FastRelax, the structures of the designs were predicted with AlphaFold2 (AF2) (*15*), and designs for which all catalytic residue Cα atoms were positioned within 1.0 Å of the design models were selected for experimental characterization (*15*) (see Computational methods, sequence design and filtering sections for

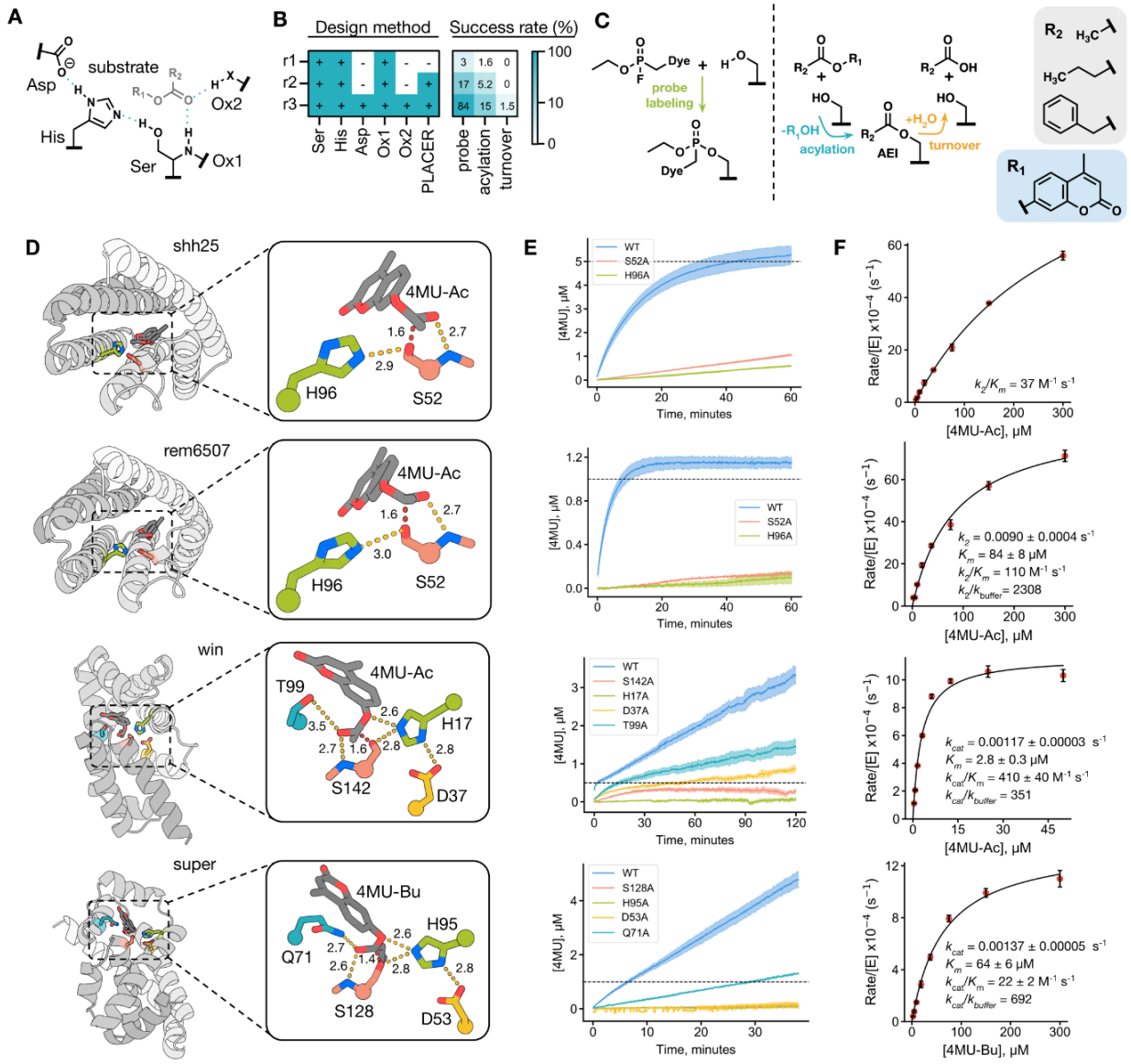

Figure 2. Functional characterization of designed serine hydrolases. (**A**) Chemical schematic of a serine hydrolase active site. (**B**) Summary of design method and experimental success rate for probe labeling, single turnover acylation, and catalytic turnover for each design round. (**C**) Chemical schematic depicting probe labeling, acylation, and catalytic turnover. (**D**) Fold (left) and active site (right) of serine hydrolase design models. (**E**) Reaction progress curves for the parent design and catalytic residue knockouts. Dashed line represents the enzyme concentration and shaded areas represent standard deviation of three technical replicates. (**F**) Michaelis-Menten plots derived from initial (shh25, rem6507) or steady state velocities (win,super). Error bars represent standard deviation of three technical replicates.

---

details).

In the first two rounds of design, we built relatively simple active sites consisting of Ser-His dyads with a single oxyanion hole contact from the backbone amide of the serine (Fig. 2A,B), and

explicitly evaluated the utility of PLACER to select designs for experimental characterization. Round 1 designs were filtered with AF2 alone, while round 2 designs that passed the AF2 filter were selected for experimental screening if PLACER ensembles of the apo state indicated the key Ser-His hydrogen bond was formed (see Computational Methods, filtering; only 1.6% of round 2 designs that passed the AF2 RMSD filter were predicted to be preorganized by PLACER). For experimental testing, we obtained synthetic genes encoding 129 and 192 designs for rounds 1 and 2, respectively, for *E. coli* overexpression and screening.

We used a fluorophosphonate (FP) activity-based probe and fluorescent 4MU-acetate (4MU-Ac) and 4MU-butyrate (4MU-Bu) ester substrates to identify designs with activated serines and esterase activity, respectively (Fig. 2C). The fraction of designs labeled by the FP probe in *E. coli* lysate increased 5-fold from 3% to 17% from round 1 to round 2 (Fig. 2B and fig. S4). Designs that reacted with the FP probe were purified and incubated with 4MU esters, and two round 1 designs (1.6%) and 10 round 2 designs (5.2%) showed catalytic activity. Retrospective PLACER analysis of the round 1 designs revealed that the Ser-His H-bonds in the two catalytically active designs were predicted to be among the most preorganized (fig. S5). PLACER filtering of round 2 designs on the extent of formation of the key Ser-His H-bond not only increased the fraction of designs exhibiting FP probe labeling and enzymatic activity, but also resulted in higher activities (Fig. 2E,F). The progress curves for these round 1 and 2 designs plateau after approximately one enzyme equivalent of fluorescent product is formed (Fig. 2E), suggesting the serine acylates but that the resulting AEI fails to hydrolyze, the rate-limiting step in the cleavage of activated esters (*99*). When incubated with substrate, mass spectra of these designs revealed a mass shift corresponding to acylation, further supporting protein inactivation following formation of the acylated intermediate (fig. S6).

We hypothesized that incorporating a histidine-stabilizing catalytic acid and a second oxyanion hole H-bond donor in a third round of designs (round 3) and filtering for PLACER preorganization in both the apo and AEI states could generate designs capable of catalytic turnover via hydrolysis of the AEI. For round 3 designs, we required all catalytic triad and oxyanion hole H-bonds to be highly

preorganized in PLACER ensembles of both the apo and AEI states. Of 132 round 3 designs, 111 (84%) displayed FP probe labeling, 20 hydrolyzed 4MU substrates (18%), and two designs (1.5%) displayed multiple turnover activity (Fig. 2B,E). Active designs from all three rounds showed significantly reduced activity upon mutation of any one of the catalytic residues (Ser, His, Asp/Glu, and oxyanion sidechain contact) (Fig. 2E), suggesting that the observed activities are dependent on the designed active site. To determine the kinetic parameters of the active designs, initial or steady-state rates were measured to determine k_2/K_m or k_{cat}/K_m for single-turnover and multiple-turnover designs, respectively (Fig. 2F and fig. S7). For the two designs that displayed catalytic turnover, called **'super'** and **'win**,' k_{cat}/K_m values were 22 M^{-1} s^{-1} (k_{cat} = 0.00137 ± 0.00005 s^{-1}, K_m = 64 ± 6 μM) and 410 M^{-1} s^{-1} (k_{cat} = 0.00117 ± 0.00003 s^{-1}, K_m = 2.8 ± 0.3 μM), respectively for the more preferred of the two 4MU substrates (**win** and **super** preferentially hydrolyzed 4MU-Ac and 4MU-Bu, respectively (fig. S8)). Despite the low K_m observed for **win**, we were unable to reach saturation of the initial burst phase of the reaction by increasing substrate concentration up to 100 μM (fig. S9), suggesting that $K_s >> K_m$ and that the low apparent K_m observed for **win** is a result of rapid acylation and not tight substrate binding.

3.4 – Structural characterization of designed serine hydrolases

We pursued x-ray crystallography to determine the accuracy with which **super** and **win** were designed. We were able to solve crystal structures of both **super** and **win**, and found that they had very low Cα RMSDs of 0.8 Å over 165 residues and 0.83 Å over 160 residues (Fig. 3A,D), respectively, to the design models. The design accuracy extends to the geometry of the active site: the sidechain conformations of the catalytic residues are in atomic agreement for **super** (all-atom RMSD = 0.38 Å over 22 atoms) and for **win** (all-atom RMSD = 0.86 Å over 20 atoms) except for a rotamer shift in the sidechain oxyanion contact, Thr99 (Fig. 3B,E). In the active site of **super**, a water molecule sits above the nucleophilic serine and forms hydrogen bonds with the oxyanion hole contacts, which likely mimics the positioning of the carbonyl oxygen of its ester substrate (Fig. 3B). Similarly, in **win**, an acetate molecule

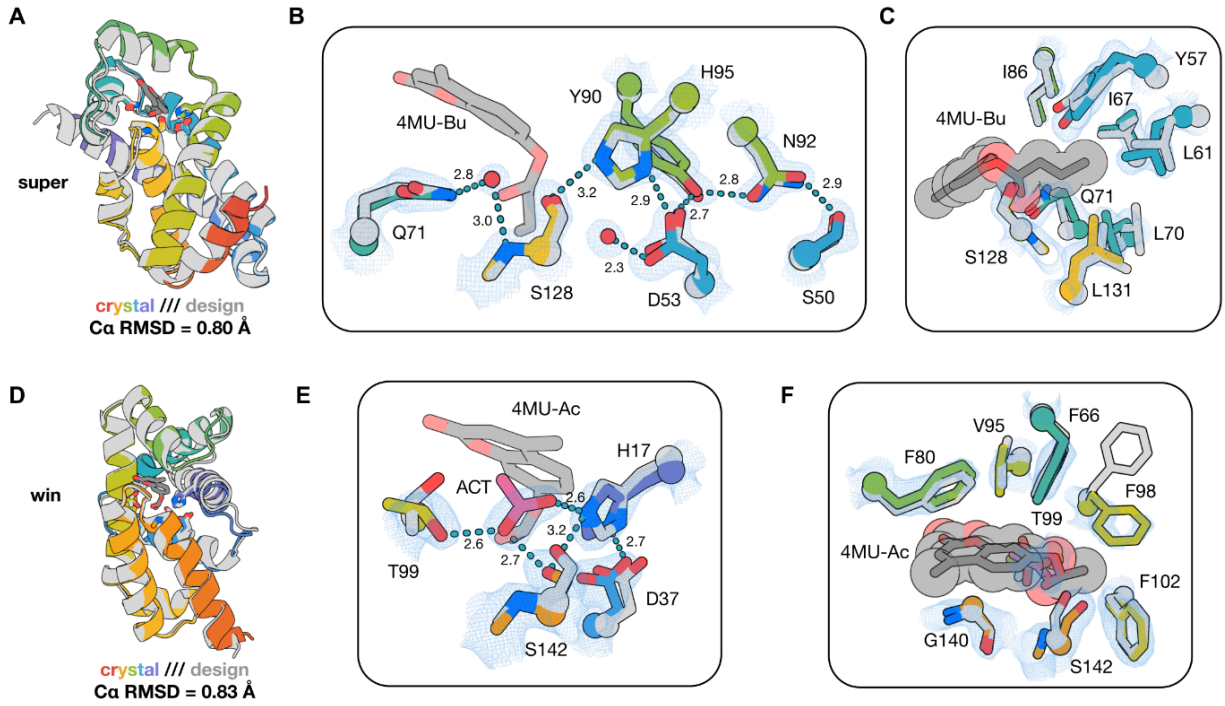

Figure 3. Structural characterization of designed serine hydrolases. (**A, D**) Structural superposition of design models (gray) and crystal structures (rainbow) for **super** (**A**) and **win** (**D**). (**B** and **E**) Active site overlays of design models (gray) and crystal structures (rainbow) of super (**B**) and win (**E**) with 2Fo-Fc map shown at 1σ (blue mesh). (**C and F**) Superposition of substrate binding sites of the design models (gray) and crystal structures (rainbow) of **super** (**C**) and **win** (**F**) with 2Fo-Fc map shown at 1 σ (blue mesh). Distances shown in Å.

---

is positioned at the catalytic center and hydrogen bonds to the catalytic serine (Ser142), the sidechain oxyanion hole (Thr99), and the histidine acid/base residue (His17) (Fig. 3E).

While the structures were solved in the absence of bound small molecule substrate or transition state analogue, overlay of the design model and crystal structure of **super** reveals high shape complementarity to the butyrate acyl group of its preferred substrate (Fig. 3C and fig. S8). At the same time, the 4MU moiety is largely exposed, corroborating the selectivity of **super** for 4MU-Bu over 4MU-Ac and suggesting that substrate binding, in this case, is largely driven by binding to the acyl group. For **win**, a rotamer shift in F98 in the crystal structure would clash with the butyrate moiety, and indeed, **win** is selective for the smaller substrate 4MU-Ac that avoids this clash (Fig. 3F and fig. S8).

The structures of **super** and **win** are very different from known structures; the closest matches found from Foldseek searches against all databases have TM-scores of 0.52 and 0.46 for **super** and **win**,

respectively (at or below the 0.5 cutoff below which structures are considered to have different topological folds), are proteins of unknown function, and have no similarity to known hydrolases at the fold or active site level (fig. S10A,B), demonstrating that the design method employed here yields structural solutions for serine hydrolase activity that extend well beyond those found in nature, expanding the structural space of this ancient enzyme family.

3.5 – Filtering for preorganization across the reaction coordinate improves catalysis

We next sought to generate and compare designs filtered explicitly with PLACER for preorganization over two states (apo and AEI) or over all four states of the reaction path by carrying out additional iterations of LigandMPNN and FastRelax starting from the active design **win** (fixing only the identities of the four catalytic residues) (Fig. 4A and fig. S1). We obtained genes encoding 45 two-state filtered designs for experimental characterization, all of which were diverse in sequence compared to the original designs (mean sequence identity to the parent design of 58% and 61% within the active site), and found 38 (84%) labeled with FP-probe (fig. S11A), and 9 (20%) displayed activity over background in a lysate screen (fig. S11C). Three of these, **win1**, **win11**, and **win31**, displayed higher catalytic turnover compared to the starting design: **win** has a k_{cat} of 0.00117 s^{-1}, which increases 15-fold in **win1** (0.018 s^{-1}), 17-fold in **win11** (0.0197 s^{-1}), and 9-fold in **win31** (0.0105 s^{-1}) (Fig. 4B and fig. S7). Of the 11 four-state filtered designs tested, 10 (91%) labeled with FP-probe (fig. S11B) and 8 (73%) displayed activity (fig. S11D). Two of these, **dadt1** and **wint4**, displayed higher catalytic efficiencies than **win**, with k_{cat}/K_m values of 3800 M^{-1} s^{-1} and 640 M^{-1} s^{-1}, driven by increases to k_{cat} and decreases in K_m relative to **win** (Fig. 4B,C,D and fig. S7). Catalytic triad residue knockouts for all designs showed significant reductions in activity, and for **win11** and **win31**, mutation of stabilizing residues in the second shell of the active site that H-bond to the catalytic aspartate also significantly reduced activity (fig. S12). The two redesigns with

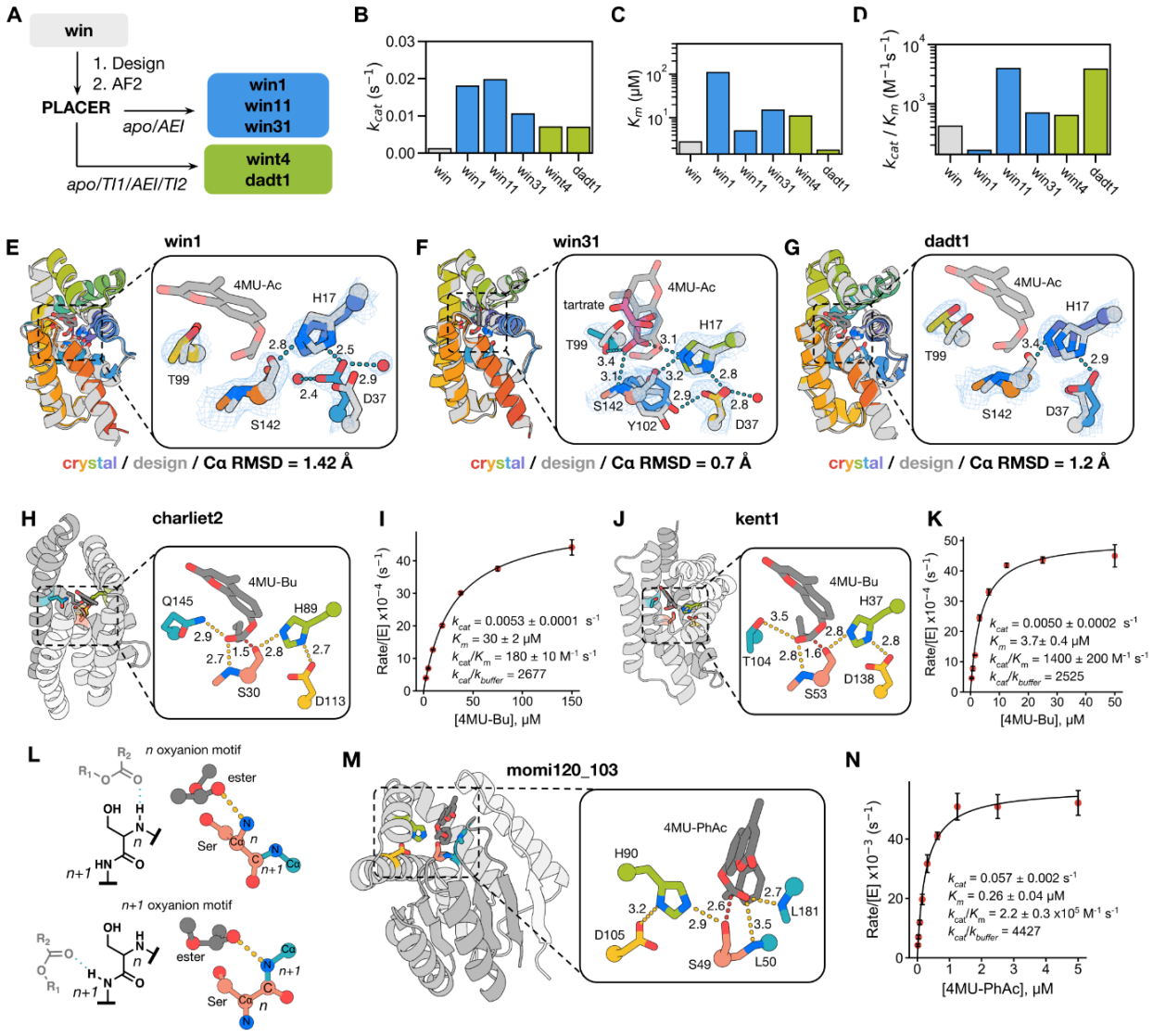

Figure 4. Computational redesign and more complex folds improve catalysis. (**A**) Computational pipeline for redesign of **win**. (**B,C,D**) k_{cat} (**B**), K_m (**C**), and k_{cat}/K_m (**D**) of parent **win** compared to computational redesigns. (**E,F,G**) Structural superposition of design model and crystal structure of win1 (**E**), **win31** (**F**), and (**G**) **dadt1** with 2Fo-Fc map shown at 1σ. (**H,I,J,K**) Design models (**H,J**) and Michaelis-Menten plots (**I,K**) for active designs with distinct folds. (**L**) Chemical and structural comparison of n and n+1 oxyanion hole motifs. (**M**) Chai-1 prediction of **momi120_103** in complex with 4MU-PhAc. (**N**) Michaelis-Menten plot for **momi120_103** with 4MU-PhAc. Error bars represent standard deviation of three technical replicates.

---

the highest k_{cat} values (**win1** and **win11**) do not display burst phase kinetics, suggesting that deacylation is no longer rate-limiting (fig. S7).

We determined the crystal structures of **win1**, **win31**, and **dadt1** and comparison to the design models revealed Cα RMSDs of 1.42 Å, 0.7 Å, and 1.2 Å, respectively (Fig. 4E,F,G). For **win1**, the active

site closely matches the designed architecture (mean all-atom RMSD = 0.54 Å) (Fig. 4E), and T99, the oxyanion hole contact, occupies the designed rotamer, which may account for the 15-fold increase in k_{cat} compared to **win**, in which T99 is rotated relative to the designed rotamer (Fig. 3E). In chain B of the **win1** structure, the catalytic serine partially occupies a second conformer with an occupancy of 0.23 (fig. S13A). For **win31**, five chains are present in the asymmetric unit, all of which closely match the design model (average Cα RMSD = 0.7 Å) at the backbone level (Fig. 4F and fig. S13B). Analysis of the active site across all chains in the asymmetric unit revealed mobility in the catalytic serine, sidechain oxyanion threonine, and a second shell tyrosine (fig. S13C), but overall a very close match to the design model active site with a mean all-atom RMSD of 0.7 Å. Tartrate, derived from the crystallization solution, fit the electron density present in the active site of all five chains, and forms hydrogen bonds with the serine, histidine, and oxyanion hole contacts (Fig. 4F), likely mimicking key contacts employed throughout the catalytic cycle. For **dadt1**, the active site closely matches the design model with a mean all-atom RMSD of 0.95 Å, and the T99 sidechain oxyanion residue occupies the designed conformation.

We next explored whether stringent PLACER filtering for optimal catalytic geometry and preorganization across the reaction coordinate could generate active esterases with novel backbone topologies and active site geometries. We performed sequence design and PLACER filtering for the complete reaction coordinate on round 3 backbones excluding **win** (fig. S1), and of 20 designs tested, two (**charliet2** and **kent1**) displayed significant esterase activity, with catalytic efficiencies of 180 M^{-1} s^{-1} and 1400 M^{-1} s^{-1} (Fig. 4H,I,J,K), suggesting that structural variability in intermediate states of the reaction coordinate may have limited otherwise functional designs. We also used sequence design combined with PLACER filtering to modify the substrate selectivity of **win1**, converting it from accepting only the small acyl group of 4MU-Ac to processing the larger substrates 4MU-phenylacetate (4MU-PhAc) (fig. S14).

To test the generality of RFdiffusion combined with PLACER filtering, we applied it to a different active site configuration in which the oxyanion hole consists of two backbone amides, rather than a backbone amide and a sidechain H-bond donor, and where the first backbone amide of the oxyanion hole is the residue following the catalytic serine (*N+1*) rather than the catalytic serine itself (*N*)

as in the previous designs (Fig. 4L). We used the RFdiffusion and LigandMPNN/FastRelax design pipeline to generate 66 designs for this new catalytic site and the larger 4MU-PhAc substrate (fig. S1). The most active of these, **momi**, displayed a k_{cat}/K_m of 1240 M^{-1} s^{-1} and a k_{cat} of 0.1 s^{-1}, a 5-fold faster rate than **win11**, the previous best design in terms of turnover number. The distribution of folds generated by RFdiffusion for this active site geometry differed from that of the original geometry, with more α/β fold solutions (as in the case of **momi**), showing how the RFdiffusion buildup approach crafts overall protein structure topology to the specific active site of interest. Natural esterases to our knowledge exclusively employ the **momi** *N+1* oxyanion hole motif, suggesting that it is particularly well suited for ester hydrolysis. The high activity achieved without any prior experimental characterization for this new catalytic site shows that filtering for preorganization across the reaction cycle can yield novel catalysts in one shot.

Several experimental results identify areas to address for improved function. First, **kent1** inactivates after roughly 10 turnovers, and mass spectra of the catalyst and the serine knockout incubated with substrate reveal stable acylated species (fig. S15), indicating that designs that hydrolyze the AEI are still susceptible to inactivation, potentially from off-mechanism acylation events in the active site or acylation-induced conformational changes. Second, mutation of the sidechain oxyanion hole residue had variable effects on activity. In three designs (**dadt1**, **charliet2**, **kent1**) from design rounds 4 and 5 that underwent stringent PLACER filtering, mutation of the sidechain oxyanion hole residue had a modest effect on activity, suggesting limited contribution to catalysis (fig. S12). Analysis of the oxyanion hole geometries in these designs and others in earlier design rounds reveal in-plane hydrogen bonds to the oxygen of the substrate carbonyl (fig. S16, Supplementary Text), in contrast to those found in nature, which are perpendicular to the plane of the carbonyl, where they likely stabilize the SP3 oxyanion transition state over the SP2 carbonyl ground state (*115–117*).

We next explored whether existing designs could be improved by rebuilding suboptimal regions using RFdiffusion. Using the **momi** backbone as input to RFdiffusion, we built out the N-terminus to further stabilize the active site but made no changes to the parent backbone or sequence (fig. S1 and fig.

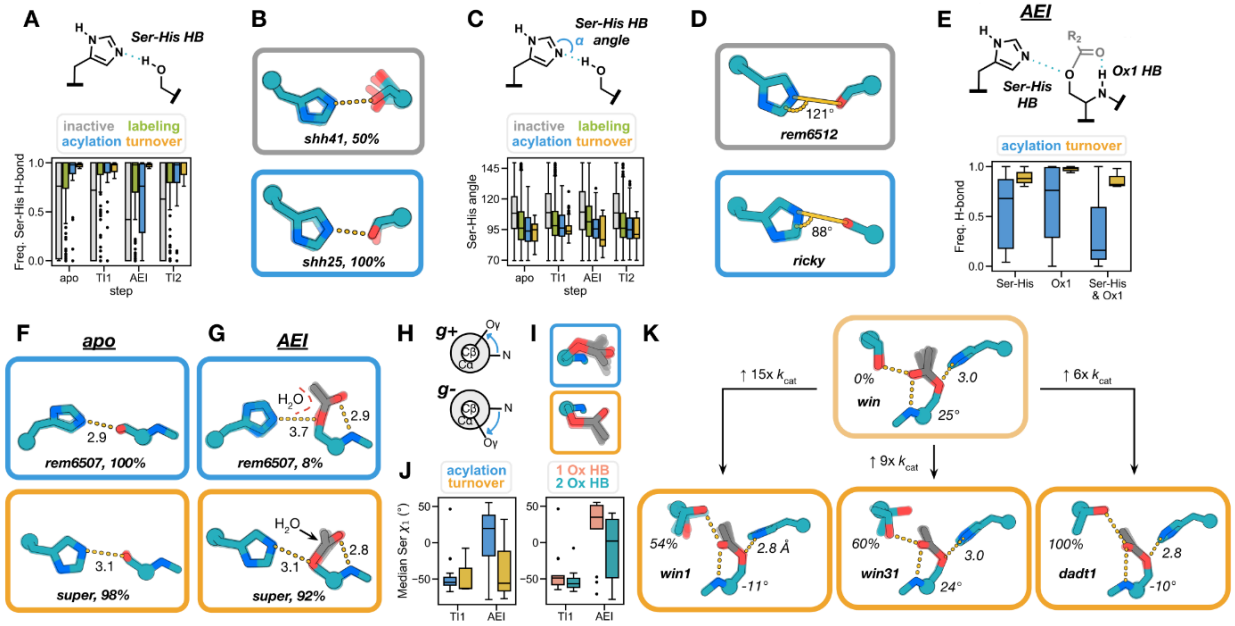

Figure 5. PLACER ensembles reveal geometric determinants of catalysis. (**A**) Frequencies of catalytic Ser-His H-bond formation in PLACER ensembles for each reaction step, grouped by experimental outcome. (**B**) Apo PLACER ensembles of representative inactive (top) and acylating (bottom) designs. (**C**) Median angle (α) between serine Oγ, histidine Nϵ and Cϵ across PLACER ensembles of inactive and acylating designs. (**D**) Apo PLACER ensembles of representative inactive (top) and acylating (bottom) designs, angle indicates median α. (**E**) AEI PLACER ensemble H-bond frequencies for designs that undergo acylation or full turnover. (**F**) PLACER ensembles of the apo state for an acylating (top) and multiple turnover design (bottom). (**G**) PLACER ensembles of the AEI state for a representative design that undergoes acylation (top) and a design that catalyzes turnover (bottom). Measurements shown represent median distances (Å) of key H-bonds indicated for each ensemble and percentages represent frequency of H-bond formation across all PLACER trajectories. (**H**) Newman projections of serine *g*+ and *g*- rotameric states (left). (**I**) PLACER ensembles of an acylating design (top) and a design that catalyzes turnover (bottom). (**J**) Median serine χ_1 angle across TI1 and AEI state PLACER ensembles for designs that catalyze acylation or turnover (left) and for the same designs grouped by number of oxyanion hole H-bonds. (**K**) AEI state PLACER ensembles for **win**, **win1**, **win31**, and **dadt1**, with percent of frames with correct oxyanion hole rotamer, Ser χ_1 angle, and catalytic Ser-His H-bond distance shown. Boxplots represent median, upper and lower quartiles; whiskers extend 1.5×IQR above and below the upper and lower quartiles (respectively). Observations falling outside these ranges plotted as outliers.

S17). Of 65 designs tested, all showed activity, and one design, **momi120**, displayed a catalytic efficiency of 4300 M^{-1} s^{-1}, 3.5-fold greater than **momi**, driven by a 2-fold increase in k_{cat} and 1.5-fold decrease in K_m (fig. S17). We also used RFdiffusion to improve the suboptimal in-plane (with respect to the substrate carbonyl) oxyanion hole H-bond formed by Gln71 in **super**. The serine protease subtilisin utilizes a chemically similar sidechain oxyanion hole, Asn155, with an amide positioned perpendicular to the plane of the substrate carbonyl (fig. S16A). Using the subtilisin oxyanion hole geometry as a guide, we mutated

Gln71 to Asn in **super**, and repositioned it to form an analogous out-of-plane H-bond to the substrate carbonyl, then rebuilt the surrounding backbone of the protein with RFdiffusion to accommodate this change (fig. S18). Of the 150 designs screened, the two most active designs, **superfast** and **supercool**, showed 8-fold and 7-fold improvements in k_{cat} over the parent design **super** (k_{cat} = 0.00137 s^{-1}), and 19-fold and 13-fold improvements in k_{cat}/K_m, respectively (fig. S18). These results highlight productive design interventions made possible by RFdiffusion that are not easily accessible with traditional engineering tools like rational mutagenesis and directed evolution, where the sequence can be readily changed but not easily augmented with new structural features.

We redesigned **momi120** for the hydrolysis of polyethylene terephthalate (PET) and screened 85 designs for activity on the sterically similar 4MU-PhAc substrate. All 85 designs displayed activity above background in a lysate screen and two of the most active designs were further kinetically characterized and found to have k_{cat}/K_m >10^4 M^{-1} s^{-1} (fig. S19). The most efficient design, **momi120-103**, has a k_{cat} for 4MU-PhAc of 0.057 s^{-1}, K_m of 0.26 μM, and a k_{cat}/K_m of 2.2x10^5 M^{-1} s^{-1} (Fig. 4N). PLACER and Chai-1 predictions suggest that 4MU-PhAc fits with high shape complementarity into the redesigned pocket; the substitutions lining the binding pocket, particularly F76G (fig. S19), appear to provide a deeper pocket that may be the structural basis of the sub-micromolar K_m.

3.6 – Structural determinants of catalysis

The high structural conservation of catalytic geometry in native serine hydrolases suggests that it is close to optimal for catalysis (*100*, *118*), but it is difficult to assess how activity depends on the detailed geometry of the interactions of the transition states with the catalytic serine, histidine, and oxyanion hole functional groups since while the identities of the catalytic residues can be readily changed by mutation, it is not straightforward to systematically vary backbone geometry. In contrast, our de novo buildup approach samples a wide range of catalytic geometries. To investigate how active site geometry and preorganization influence catalytic activity, we generated PLACER ensembles of all 812 experimentally

characterized designs, categorized as inactive, FP probe labeling, acylation, and catalytic turnover, for each reaction step in the hydrolysis of 4MU-Ac (including design rounds 1-3 and previous NTF2-based designs). We summarize the strongest trends in the following paragraphs.

Increased preorganization and bending of the Ser-His H-bond were associated with higher rates of probe-labeling, acylation, and turnover. All designs capable of catalyzing turnover displayed highly preorganized Ser-His H-bonds across all four states, while inactive designs often displayed rotamer shifts causing loss of the interaction (Fig. 5A,B). Designs that catalyzed turnover had Ser(Oγ):His(Nε-Cε) bond angles that were more acute (median, all states = 94°) than inactive designs (median, all states = 108°), which were more similar to serine-histidine hydrogen bonds across the PDB (~125°) (*115*) (Fig. 5C). This acute H-bond is consistent with the reaction mechanism, as this geometry allows histidine to participate, without changing conformation, in all of the necessary proton transfers involving serine, the leaving group oxygen in TI1, and the hydrolytic water (*102*, *119*). This compromise in positioning is observed not only in our active designs but also in many of those found in nature (*115*, *119*, *120*).

The geometry of the serine rotamer throughout the catalytic cycle was also strongly correlated with experimental outcome. For designs that display acylation or turnover, we found that serine largely occupies the active *g*- rotamer (*118*) in the apo state. Designs that display turnover retain the *g*- serine conformer upon formation of the AEI, but designs that irreversibly acylate switch to the *g*+ rotamer in the AEI (Fig. 5H,I,J). The *g*+ serine rotamer is catalytically incompetent in these designs because it leads to an acyl group conformation that occludes interaction of the hydrolytic water with histidine (Fig. 5G), increases the median Ser-His H-bond distance (Fig. 5G), and reduces the frequency that the Ser-His and oxyanion hole-acyl group H-bonds form (Fig. 5E). The same retention of the *g*- rotamer in the AEI is observed in native crystal structures (*102*). PLACER analysis also revealed that the presence of a second oxyanion hole residue favors the active *g*- serine rotamer: those designs with only one oxyanion hole H-bond (from the backbone amide of the serine nucleophile) shift from *g*- to *g*+ upon acylation, while designs with two oxyanion hole H-bonds predominantly occupy *g*- Ser rotamers (Fig. 5J, right). The

second oxyanion hole contact in serine hydrolases thus not only stabilizes the transition state but likely helps orient intermediates in catalytically productive conformations.

Differential preorganization may also explain activity trends in the **win**, **win1**, **win31**, and **dadt1** series. PLACER analysis of the crystal structures of these designs revealed that in the AEI state, the more active redesigns **win1**, **win31**, and **dadt1** sample the designed T99 oxyanion hole rotamer in 56, 60, and 100% of predictions, respectively, while the less active **win** never adopts this rotamer (Fig. 5K). Although both observed rotamers place T99 Oγ within hydrogen bonding distance of the oxyanion, the designed rotamer-oxyanion dihedral angle (91°) adopted by the redesigns much more closely matches the angles observed in native serine hydrolases, suggesting it is likely more optimal for selective transition state stabilization (*115–117*). We also observed differences in the serine rotameric state and the preorganization of the acyl group in the AEI state. Both **win** and **win31** occupy the catalytically unfavorable *g+* rotamer across the entire AEI ensemble, while **win1** and **dadt1** both display a less pronounced rotameric shift, which leads to shorter Ser-His H-bond distances (mean H-bond distance of 2.8 Å in **win1** and **dadt1** compared to 3.1 Å in **win** and **win31**). Overall, the acyl groups of **win1** and especially **win31** and **dadt1** display significantly less conformational heterogeneity than that of **win**, which may increase the likelihood of histidine-mediated water attack (Fig. 5K).

3.7 – Conclusion

The substantial catalytic efficiencies, the complexity of the active sites, and the atomic accuracy of the designs described here represent major advances in computational enzyme design. The serine catalytic triad plus oxyanion hole mechanism involves complex machinery that is challenging to scaffold (compared to, for example, the Kemp eliminase, which requires only a general base in a hydrophobic environment (*22*)), necessitates chemical activation of serine, and proceeds through a complex multistep mechanism that traverses a chemically stable AEI. The designed serine hydrolases described here have efficiencies up to 2.2×10^5 M^{-1} s^{-1}, a significant improvement in function for computationally designed

enzymes. For example, the previously designed esterase OE1 has a k_{cat}/K_m = 210 M^{-1} s^{-1} and reached an efficiency of 3190 M^{-1} s^{-1} after four rounds of directed evolution and screening over 12,000 clones, despite the use of a more activated N_δ-methylhistidine nucleophile (*85*). The closest comparable de novo design in terms of mechanism, in which a cysteine-based catalytic triad was mutated into a peptide-based helical barrel that proceeds via a more activated thioester intermediate (*83*), has a k_{cat}/K_m of 3.7 M^{-1} s^{-1} and k_{cat} of 0.0005 s^{-1}, 60000x less efficient and 400x slower than the most efficient (**momi120-103**) and highest turnover design (**momi120**) described here, respectively. The ability to accelerate the hydrolysis of a chemically stable acyl-enzyme intermediate has been a decades-old challenge in enzyme design. To approximate the deacylation rate enhancement, we compared the uncatalyzed rate of hydrolysis of ethyl acetate (2.5–5.0)x10^{-10} s^{-1}, (*121*)) to the lower limit of the deacylation rate constant of **momi** (k_{cat}, 0.076 s^{-1}, pH 7.0, 25°C), yielding an estimated rate enhancement of over 10^8. Taken together, the design of serine hydrolases spanning five folds not represented in natural esterases, the considerable improvement in activity over previously designed esterases, and the acceleration of deacylation represent key advances in enzyme design.

The designs described here are not as efficient as native serine hydrolases with their cognate substrates (e.g. the k_{cat}/K_m of acetylcholinesterase with acetylcholine is >10^8 M^{-1} s^{-1}) (*122*), but they have efficiencies comparable or better than natural proteases for activated esters (α-chymotrypsin with *p*-nitrophenyl acetate k_{cat}/K_m*:* 3530 M^{-1} s^{-1}, k_{cat}*:* 0.0053 s^{-1}; subtilisin with *p*-nitrophenyl acetate k_{cat}/K_m*:* 610 $M^{-1}s^{-1}$, k_{cat}*:* 0.23 s^{-1}) (*123*, *124*), and are within the distribution of efficiencies observed in nature (*122*). Higher k_{cat} could likely be achieved through optimization of the catalytic geometry, further preorganization of the active site (*60*, *77*), and increasing active site complexity. Acetylcholinesterase employs three backbone amide hydrogen bonds to the oxyanion and an additional network of hydrogen bonds to stabilize the catalytic aspartate (*125*, *126*). The current designs do not employ this machinery, and comparison of catalytic triad and oxyanion hole geometries to those found in highly efficient native serine hydrolases highlights differences that could be responsible for the remaining activity gap (see Supplementary Text). Our de novo buildup approach using RFdiffusion coupled with PLACER ensemble

analysis to ensure design accuracy and preorganization should allow us to test these hypotheses by direct construction, which should complement more traditional approaches based on structural examination, computational analysis, and optimization by experimental approaches like directed evolution.

Previous efforts to design catalytic triad-based designs have failed to achieve multiple turnover; in some cases, such as our preliminary NTF2-based designs, a backbone amide oxyanion hole was impossible to achieve due to scaffold limitations, while in others based on native scaffolds, the histidine geometry was difficult to control which likely limited activation of the leaving groups and water (fig. S20) (*27*). De novo backbone generation building outward from a specified active site with RFdiffusion, described here for serine hydrolases and also recently used to generate retroaldolases (77), overcomes these limitations by enabling generation of almost any desired catalytic geometry. We further show that the deep neural network PLACER can rapidly generate ensembles for a series of reaction intermediates to predict preorganization, and provide insights that would otherwise require labor-intensive structural studies. For example, PLACER revealed pervasive off-target conformational changes in the acyl-enzyme intermediate, providing feedback on design flaws that would go unnoticed when considering only a single state in the catalytic cycle. The value of this approach is evident in the dramatic improvement in experimental success rate upon filtering with PLACER, suggesting that such ensemble generation will be useful for enzyme design moving forward. While the designs described here do utilize a known mechanism, the geometries sampled and the folds that scaffold them are distinct from those found in native proteins, and the insights provided by PLACER for these geometries suggests that the approach should prove valuable for assessing catalytic geometries for which no native precedent exists. We anticipate that the ability to precisely position multiple catalytic groups using RFdiffusion, and to assess active site organization throughout a complex reaction cycle using PLACER should enable the design of a wide variety of new catalysts, such as PETases, amidases, and ligases, in the near future.

3.7 – Methods

NTF2 design campaign. Catalytic geometries from a previous analysis of native serine hydrolases (*100*) were used to generate constraint files for use in the RosettaMatch algorithm (*127*). The scaffold set used for matching was a set of idealized Nuclear Transport Factor 2 (NTF2) fold proteins generated with trRosetta (*28*). After matching, sequence design was performed using LigandMPNN and FastRelax and designs were filtered using AlphaFold2 as described below. An additional filter was used requiring that all catalytic hydrogen bonds in the active site be formed in the AlphaFold2 prediction.

Computational design of serine hydrolases.

Motif generation

Motifs were built in an iterative process. First, a substrate rotamer in a transition state geometry (either 4MU-Bu or 4MU-Ac) was placed in accordance with geometries in ref 33 in relation to a 3-residue stub of the serine and local oxyanion hole from one of two natural serine hydrolase crystal structures, in which all residues other than serine were mutated to alanine (*N* oxyanion hole: 1scn, residues 220-222; *N+1* oxyanion hole: 1lns, residues 347-349). The transition state geometry of the substrate ester group was determined by DFT geometry optimization (B3LYP-D3(BJ)/6-31G(d)). Next, positions and rotamers of histidine on 3-residue helical or strand stubs flanked by alanine were sampled around the catalytic serine and filtered for those structures in which the histidine simultaneously formed hydrogen bonds with the catalytic serine and the substrate leaving group oxygen. This process resulted in 108 unique motifs for design rounds 1 and 2. For the round 3 motifs, initially the aspartate or glutamate residue and second oxyanion hole hydrogen bond were added in a similar manner using geometric sampling of hydrogen-bonding conformations and rotamers. However, backbones produced from these motifs had exceedingly low AF2 success rates, presumably due to the generation of incompatible combinations of backbone conformations. To ensure that the remaining catalytic residue stubs were placed in physically plausible geometries, we generated 10,000 backbones with RFdiffusion using the simple

substrate-Ser-His motifs as input, and then searched these backbones using Rosetta for positions on secondary structure that could accommodate the aspartate or glutamate triad residue to hydrogen bond to histidine. These stubs were then extracted, and in a final step, the same process was repeated to generate stubs for the second oxyanion hole, considering all hydrogen bond donating sidechains, ultimately producing 2238 unique round 3 motifs with Ser-His-Asp/Glu catalytic triads, and Ser/Thr/Tyr/His/Trp oxyanion holes.

Backbone generation

See supplemental methods for a detailed description of CA diffusion, which was employed to generate backbones to scaffold the generated active sites.

Sequence design

We performed three cycles of LigandMPNN (*112*) and Rosetta FastRelax (*128*) to design sequences for backbones generated from RFdiffusion. To encourage formation of hydrogen bond contacts to the catalytic histidine (for round 1 motifs) and to the catalytic aspartate/glutamate (round 3 motifs), the log probabilities used by LigandMPNN to select residues were biased toward polar amino acids for all residues with Cα within 8 Å of the active site. Catalytic residues were kept fixed and Rosetta enzyme constraints (*127*, *129*) were applied during the relax steps to maintain the catalytic geometry during each LigandMPNN/FastRelax cycle. Constraints were defined for each hydrogen bonding interaction between the catalytic dyad, backbone oxyanion hole, and substrate using the starting motif geometry with tolerances of 0.1 Å for distances and 5° for angles and dihedrals. For designs with catalytic triads, the His-Asp interaction was constrained

Filtering

After sequence design, designs were filtered on the recapitulation of the motif catalytic geometry after FastRelax and the shape complementarity of the binding site to the substrate using Rosetta. Passing designs were used as input to AF2 (*15*) for single sequence structure prediction. AF2 was run using model

4 with three recycles. Designs were filtered for a global Cα RMSD < 1.5 Å, pLDDT > 75, and catalytic residue Cα RMSD < 1.0 Å. In the case of final round N+1 oxyanion hole designs, a modified version of Initial Guess AF2 was used to predict designs with sparse template information provided (see Supplementary Methods).

Designs that passed AF2 filters were subsequently analyzed using PLACER. PLACER is a denoising neural network trained on X-ray and EM structures from the PDB to recapitulate the correct atom positions from partially corrupted input structures provided the atom type and bond connectivity is known. PLACER predictions were done for a spatial crop of 600 atoms closest to the active site. The inputs to the network included the protein backbone coordinates within the crop and the amino acid sequence with side chain coordinates randomly initialized around the respective Cα atoms. For proteins without a crystal structure, the AF2 model was used. For every designed protein, we modeled 5 reaction states representing the chemical modifications the catalytic serine undergoes in the course of the reaction: 1) apo, 2) substrate bound, 3) tetrahedral intermediate 1 (TI1), 4) acylenzyme intermediate (AEI), and 5) tetrahedral intermediate 2 (TI2). We used 50 different seeds to generate an ensemble of 50 PLACER models for each reaction state (apo, substrate bound, TI1, AEI, and TI2). For each of the 50 models in a given ensemble, the presence and geometry of key hydrogen bonds in each individual model (see Supplemental Methods) were determined. To analyze native hydrolases with PLACER, a set of native crystal structures was collected (*115*) (PDB IDs: 1ACB_E, 1C5L_H, 1H2W_A, 1IC6_A, 1IVY_A, 1PFQ_A, 1QNJ_A, 1QTR_A, 1ST2_A, 2H5C_A, 2QAA_A, 3MI4_A, 5JXG_A), the active site locations identified, and the aforementioned process applied.

Backbone resampling for momi and super redesign campaigns. The design model of **momi** was provided as input to RFdiffusion and the entirety of the protein was fixed while a region of secondary structure was diffused at the N-terminus. The length of this region was randomly sampled from a range of 20 to 50 amino acids for 1000 independent diffusion trajectories. The contigs flag for RFdiffusion was as follows: contigs:{region_length},A1-160. For each backbone, the sequence of the original **momi** input

was kept fixed while the newly diffused region at the N-terminus was designed as described previously with LigandMPNN and FastRelax, with ten sequences generated per backbone.

To generate designs in complex with the PET substrate, **momi120** was redesigned around a 2-mer of the PET polymer. The PET 2-mer was aligned into the active site based on the geometry of the original **momi120** design in complex with 4MU-PhAc substrate. Two regions of secondary structure which clashed with the aligned PET substrate, region 1 (residues 66-87) which flanks the lower cleft of the active site and region 2 (residues 94-104) which sits above the catalytic histidine, were subsequently remodeled with RFdiffusion. The lengths of region 1 and 2 were randomly sampled from a range of 18 to 28 amino acids and 7 to 17 amino acids, respectively, for 1000 independent diffusion trajectories. The contigs flag for RFdiffusion was formatted as follows: contigs:A1-65,{region1_length},A88-93,{region2_length},A105-194. The sequence of the entire structure was designed as described above. Twenty sequences were generated per backbone and designs were filtered as previously described with AF2 and PLACER. For 74 backbones that passed AF2 and PLACER filters, sequences were designed again as described above with 1000 sequences generated per backbone and subsequently filtered for confidence and self-consistency by single sequence AF2 prediction.

To generate a version of **super** with an optimized oxyanion hole sidechain geometry, we started by superimposing the active sites of **super** and subtilisin (PDB: 1scn) by alignment of the catalytic serine backbone atoms. Residues 56-91 that flank the oxyanion hole residue Gln71 in **super** were removed and Asn155 that was aligned from subtilisin was copied into the structure. We used RFdiffusion2 (*130*), a backbone generation model capable of scaffolding individual atoms or functional groups, to reconstruct the removed region of **super** and scaffold the newly placed amide group of Asn. We sampled lengths between 48-58 residues to generate 10,000 unique backbones which were then designed and filtered as described above.

In-gel fluorescence screening with activity-based probes. DNA encoding the designed proteins was ordered from IDT as eblocks and the GoldenGate method was used to clone them into vector LM627 (addgene), which contains a C-terminal SNAC tag followed by a hexahistidine-tag. Resulting plasmid was transformed into BL21(DE3) cells and grown overnight in 1 mL of LB supplemented with 50 μg/ml kanamycin. For expression, 100 μL of overnight culture was used to inoculate 1 mL of LB media and grown for 1.5 hours at 37 °C on a Heidolph shaker at 1300 rpm and then 10 μL of 100 mM IPTG was added and cultures were incubated at 37 °C with shaking for an additional 3 hours. Cultures were centrifuged at 4000*g* for 10 minutes and supernatant removed. Cell pellets were resuspended in 200 μL of 20 mM HEPES (pH 7.4), containing 50 mM NaCl, 0.1 mg/mL lysozyme, and 0.01 mg/mL DNaseI. After 15 minutes, lysates were frozen in liquid nitrogen and subsequently thawed at room temperature. For labeling, 10 μL of lysate was incubated with 1 μM FP-TAMRA probe (10 μL of 2 μM stock in lysis buffer) for 1 hour at room temperature before quenching with 2x Laemmli sample buffer. Labeled samples were heated at 95°C for 5 minutes and 10 μL of each sample was separated on a BioRad AnykD Criterion precast gel and fluorescence imaging performed using a LI-COR Odyssey M imager. Gels were subsequently stained with coomassie blue and imaged again.

Lysate screening. DNA encoding the designed proteins was ordered from IDT as eblocks and cloned by the GoldenGate method into vector pCOOL1 which contains a C-terminal mScarlet-i3 fusion to enable normalization of activity in lysate by enzyme concentration. Resulting plasmid was transformed into BL21(DE3) cells and cultures were grown overnight at 1 mL scale in 2 mL deep-well 96-well round bottom plates on a Heidolph shaker at 1300 rpm and 37 °C. For expression, 50 μL of the overnight cultures were used to inoculate 1 mL of autoinduction media in 2 mL deep-well 96-well round bottom plates and incubated at 1300 rpm and 37 °C for approximately 24 hours. Cultures were centrifuged at 4000*g* for 10 minutes and supernatant decanted, washed with buffer (20 mM HEPES, 50 mM NaCl, pH 7.4), and incubated on a Heidolph shaker at 1300 rpm at room temp for 5 minutes to resuspend. Plates were centrifuged again at 4000*g* for 10 minutes and supernatant decanted. For lysis, cell pellets were

resuspended with 500 µL of lysis buffer (20 mM HEPES, 50 mM NaCl, 0.01 mg/mL DNAseI, 0.01 mg/mL lysozyme, 1 mM EDTA, 0.1% triton X-100) and incubated for 2 hours on a Heidolph shaker (1300 rpm, 37 °C). Plates were centrifuged at 4300*g* for 30 minutes and supernatant collected for screening. For activity screening, 4 or 6 µL of lysate was aliquoted into microtiter plates and reactions initiated by addition of 36 or 54 µL of buffer containing 111.1 µM 4MU-Ac or 4MU-Bu, 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO. Volume sizes were modified depending on plate type used, where half-area plates were used for 40 µL reaction volume and full-area plates were used with 60 µL reaction volume. Upon addition of substrate, microtiter plates were measured once for mScarlet-i3 signal and then subsequently monitored continuously for the generation of 4MU (ex: 365 nm, em: 445 nm) on a Neo2 plate reader.

Protein expression and purification. Genes encoding the designed proteins were ordered from IDT as eblocks and cloned via the Golden Gate method into vector LM627 as previously described (*38*). Resulting plasmid was transformed into BL21(DE3) cells and grown overnight in 1 mL of LB supplemented with 50 µg/ml kanamycin, after which 500 µL of overnight was used to inoculate 50 mL of autoinduction media (*131*), which was grown 4-6 hours at 37 °C and then overnight at 18 °C. Cultures were spun down at 4000*g* for 15 minutes, and supernatant decanted. Cell pellets were resuspended in 25 mL of cold wash buffer (40 mM imidazole, 500 mM NaCl, 50 mM sodium phosphate, pH 7.4) with 1 mg/mL lysozyme and 0.1 mg/mL DNAse I. Cell slurries were sonicated on ice for 2.5 minutes at 80% amplitude, 10s on 10s off. The resulting lysate was centrifuged at 14000*g* for 30 minutes and the supernatant was applied to 1 mL of Ni-NTA resin equilibrated with wash buffer. The resin was subsequently washed with 15 mL of wash buffer 3 times and once with 400 µL of elution buffer (400 mM imidazole, 500 mM NaCl, 50 mM sodium phosphate, pH 7.4) followed by elution with 1.3 mL elution buffer. The eluate was purified by size-exclusion chromatography on a Superdex 75 Increase 10/300 GL with running buffer of 20 mM HEPES, 50 mM NaCl, pH 7.4. Samples were either used immediately in downstream experiments or snap frozen in liquid nitrogen and stored at -80 C. Protein molecular weight

was confirmed by LC-MS.

Kinetic analysis. To characterize hits identified from in-gel fluorescence and lysate screens for catalytic turnover, we incubated purified protein samples with fluorogenic substrates 4MU-Ac, 4MU-Bu and 4MU-PhAc. Kinetic screens were either performed in 40 µL reaction volumes in 96-well half area plates or 60 µL reaction volume in 96-well full-area plates. Protein and substrate were prepared fresh in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO. Either 4 or 6 µL of enzyme was added to microtiter plates and the reactions were initiated by addition of substrate (36 or 54 µL). Generation of the fluorogenic product 4MU was monitored continuously (excitation 365 nm, emission 445 nm) on a Neo2 plate reader with incubation at 30 °C. Analysis of the resulting data was carried out using custom scripts (see computational methods). In cases where single-turnover activity was observed, initial velocities were used to determine k_2/K_m. For those designs that displayed a clear burst phase followed by a slower steady-state rate, straight-line fits of the steady-state velocities were used to determine Michaelis-Menten catalytic parameters. To determine the uncatalyzed reaction rate in assay buffer (20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO), substrate was diluted in buffer alone and rates determined at multiple substrate concentrations, after which the rate was determined from fitting [S] versus rate with an equation of the form rate = k_{buffer}[S].

Crystallography. Proteins for crystallography were prepared as described above, but SEC was done with SNAC tag cleavage buffer (*132*). After SEC, protein eluate was incubated with 500 mM guanidinium hydrochloride and 2 mM $NiCl_2$ overnight at room temperature to remove the C-terminal His tag. The SNAC cleavage reaction was applied to a nickel column equilibrated with wash buffer to remove any uncleaved product and resulting eluate applied to a Superdex 75 Increase 10/300 GL column with 20 mM HEPES, 50 mM NaCl, pH 7.4 as the running buffer. Samples were concentrated and stored at -80 °C or immediately used for crystallization. Crystallization screening was performed using a Mosquito LCP by STP Labtech and resulting crystals were harvested directly from the screening plate. Crystallization conditions for each design were as follows: **n8** (15 mg/mL) in 0.1 M Bis-Tris pH 5.5, 25% (w/v) PEG

3350, **super** (50 mg/mL) in 0.2 M Potassium fluoride, 20% (w/v) PEG 3350, **win** (42 mg/mL) in 0.1 M Sodium acetate pH 4.6, 8% (w/v) PEG 4000, **win1** (54 mg/mL) in 60% v/v Tacsimate pH 7.0, **win31** (60 mg/mL) in 0.2 M diammonium tartrate and 20% (w/v) PEG 3350, and **dadt1** (27 mg/mL) in 0.1 M Potassium chloride, 0.02 M Tris pH 7.0, and 20% PEG4000. Data were processed with XDS (*133*), phased and refined with Phenix (*134*), and model building performed with COOT (*135*). Percent Ramachandran favored, allowed, and outliers for each structure are as follows: **n8** (98.21, 1.79, 0.00), **super** (99.37, 0.63, 0.00), **win** (97.99, 2.01, 0.00), **win1** (99.68, 0.32, 0.00), **win31** (99.36, 0.64, 0.00), and **dadt1** (100, 0, 0). Coordinates are deposited in the PDB with PDB IDs of 9DED (**n8**), 9DEE (**super**), 9DEF (**win**), 9DEG (**win1**), 9DEH (**win31**), and 9MRB (**dadt1**).

Mass spectrometry. Intact mass spectra of protein samples were obtained by reverse-phase LC/MS on an Agilent G6230B TOF after desalting using an AdvanceBio RP-Desalting column. Deconvolution using a total entropy algorithm was performed using Bioconfirm. In some cases, protein samples (1 mg/mL) were incubated overnight with substrate (300 μM) in SEC running buffer at room temperature prior to mass spectrometry analysis.

Structural similarity search of the PDB and AFDB. To assess the structural novelty of our designed enzymes, we used FoldSeek (*136*) to compare our crystal structures and select design models against all available databases. Searches were performed in TM-align mode and the highest TM-score hit was used for structural comparison.

Chapter 4: *De novo* design of PETases

4.1 – Introduction

In 2024, an estimated 220 million tons of plastic waste was generated; while already staggering, that figure is expected to double by 2040 (*137*). Plastic pollution has become ubiquitous in the environment, while microplastics are increasingly being found in human tissues. Plastic waste presents an immediate and well-documented threat to marine environments, where it entangles or is consumed by wildlife, but its impact on human health is not yet fully understood (*138*, *139*). While many countries have plastic recycling facilities in place, current technologies cannot keep pace with the growing quantity of waste being generated. A majority of the plastic waste that is processed undergoes mechanical recycling, in which high heat and pressure is applied to a sorted plastic waste stream and composite material is extruded (*140*). This material exhibits decreased performance because of the extreme conditions to which it is subjected, resulting in inferior quality products. The process is also energy- and labor-intensive due to the requirements for high heat, high pressure, and sorting of waste streams by polymer composition. As a result, plastic material can be thermomechanically recycled a finite number of times, and the production of virgin material is ultimately more financially viable.

Recycling via enzymatic depolymerization addresses many of these issues. Firstly, as enzymes frequently are active under ambient conditions, enzymatic processing would require significantly less energy. Additionally, enzymes would also negate the need for plastic waste stream sorting, as enzymes would selectively depolymerize their substrate from a mixed sample. These soluble monomers can then be separated from the insoluble intact plastic and synthesized into new plastic which retains the properties of virgin material, or upcycled into value-added products.

However, plastic recycling also presents unique challenges to biocatalysis. Firstly, because the polymer chains of plastics are much less physically accessible than small molecule substrates, plastic hydrolysis reactions are often performed at or above the glass transition temperature (T_g), at which the

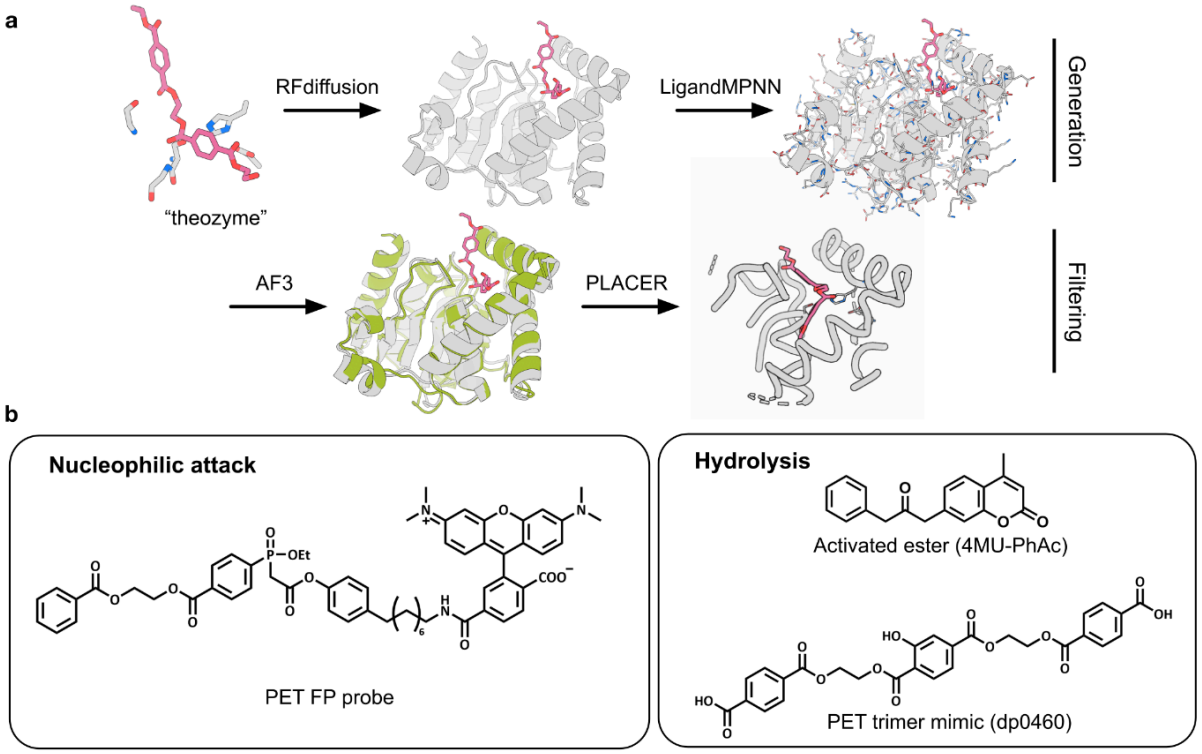

Figure 1. Design and screening methods. (**A**) Given the geometry of a possible active site configuration, RFdiffusion denoising trajectories generate backbone coordinates which scaffold the site. Sequence design is performed with LigandMPNN. Outputs are filtered for self-consistency between design models and predicted structures from AlphaFold2. Active site ensembles generated with PLACER are then evaluated for pre. The coordinates of the sidechains around the active site and any bound small molecule for the step in the reaction being considered are randomized, and n samples are carried out to generate an ensemble of predictions. (**B**) Activity-based substrates for evaluation of hydrolysis. To screen the initial nucleophilic attack step, a fluorophosphonate-based probe with a PET mimic moiety was used. Designs with an activated serine nucleophile become covalently bound to the probe and are thus fluorescently labeled by the TAMRA group. 4MU-PhAc is a coumarin derivative which was used to screen for complete catalytic turnover on an activated ester substrate. dp0460 is a fluorescent reporter compound mimicking a PET trimer, thus providing a high throughput screen for quasi-PETase activity.

---

chains become more flexible (T_g of polyethylene terephthalate is ~67-81 °C) (*141*). Thus, plastic-degrading enzymes must be stable and active at or above this temperature range. Additionally, because the accessibility of the scissile bond is limited in polymeric substrates, active sites must be solvent-accessible, forming a groove, rather than the more common binding pocket feature.

In 2016, a bacterium isolated from a recycling facility in Japan was found to be capable of metabolizing the plastic polyethylene terephthalate (PET) as its primary carbon source. Genomic analysis identified two enzymes responsible for the reaction, dubbed PET hydrolase (PETase) and MHET

hydrolase (MHETase) which together hydrolyze bulk PET to its constituent monomers, terephthalic acid and ethylene glycol (*90*). Subsequent work established PETase and its homologs as viable catalysts for industrial PET recycling, and the dual-enzyme system is currently in use at multiple companies across Europe (*89*, *142*). However, despite being the most deeply studied and optimized plastic-degrading enzyme, PETases have thus far been constrained to the serine hydrolase fold, with little sequence and structural diversity (*143*, *144*).

4.2 – Design and characterization of PETases

Given the progress in the design of serine hydrolases, PETases are a natural model reaction for the development of methods for the design of plastic degrading enzymes. Additionally, the limited soluble yield and stability of natural PETases could be greatly improved through design, as evidenced by the remarkable stability of many previously designed enzymes. Concurrent to efforts to design simple serine hydrolases, we applied the same methods to the design of PETases. Polyesters are inherently more challenging to hydrolyze both due to the chemical energy of their substituent ester bonds as compared to the activated ester substrates described in chapter 3 and the requirements for high thermostability and exposed active sites.

To design PETases, we began with the same active site theozymes as used in chapter 3 with a dimer of the PET molecule situated in the catalytic center. RFdiffusion was used to generate protein backbones which perfectly scaffolded the active site motif around the PET substrate. LigandMPNN was used to assign amino acid sequences to each backbone, introducing favorable interactions between the protein and substrate. To enrich for backbones with an exposed surface groove active site, designs which completely enveloped any part of the substrate mimic were eliminated. Designs were then modeled and filtered with PLACER as described above (Fig. 1A).

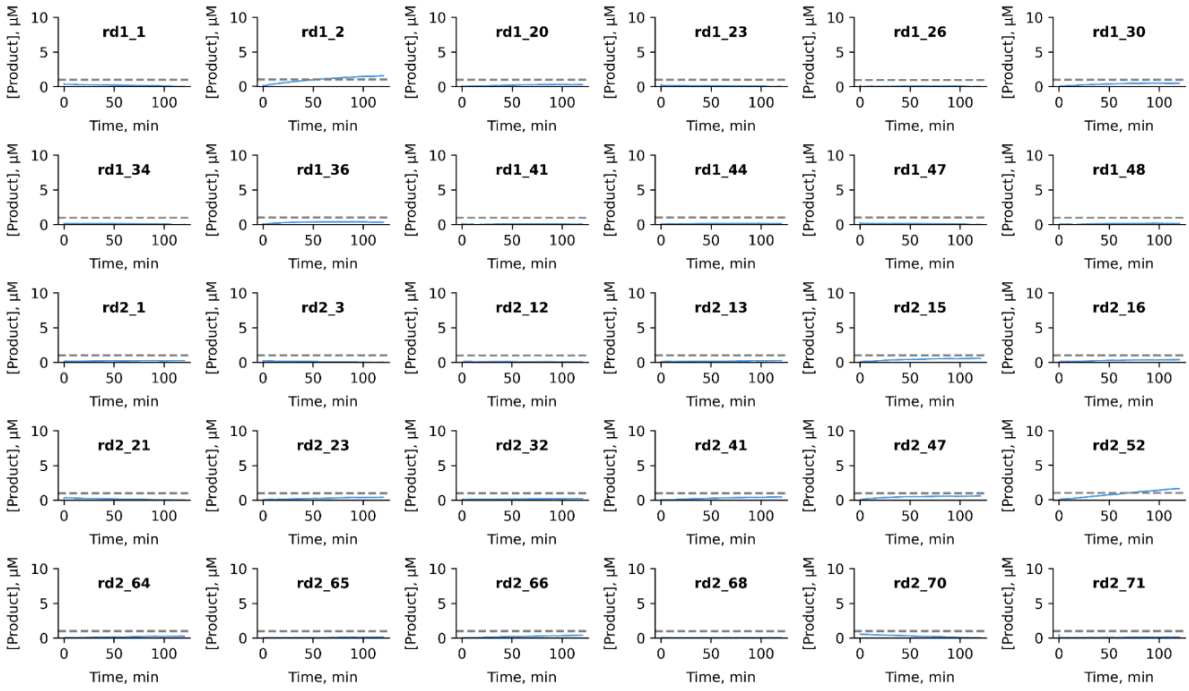

Figure 2. Catalytic turnover of activated esters. Reaction progress curves of 4MU-PhAc substrate incubated with purified designs from rounds 1 and 2. Dashed lines represent enzyme concentration. Two designs showed multiple turnovers (accumulation of product at a greater concentration than that of enzyme).

PETase designs were experimentally screened in four activity-based assays of increasing difficulty: covalent labeling of a FP probe with a PET mimic moiety, hydrolysis of 4MU-PhAc (the fluorescent activated ester substrate with the most structural similarity to PET), hydrolysis of a fluorescent PET trimer mimic reporter (dp0460), and hydrolysis of bulk PET substrate (Fig. 1B).

In the first two rounds of design, theozymes exclusively containing the *N*-motif were used. Resulting designs were exclusively composed of alpha helical content. Designs were incubated with the PET mimic FP probe and 14 from round 1 (28%) and 18 from round 2 (25%) labeled (fig. S2A,B). Reactive designs were purified and incubated with 4MU-PhAc and dp0460 in separate assays; two hydrolyzed 4MU-PhAc and none hydrolyzed dp0460 (Fig. 2).

We hypothesized that utilizing the aforementioned *N+1* motif, in which there are two backbone amide oxyanion contacts, one of which is the residue following the serine nucleophile, may enhance

activity and increase the complexity of folds. Indeed, designs generated from these theozymes were overwhelmingly of a mixed alpha-helical/beta-sheet fold in stark contrast to previous rounds (fig. S1). When tested experimentally, 56 of 96 ordered designs (58%) labeled with the PET FP probe (fig. S2C). The 7 designs with strongest labeling were purified and evaluated for esterase activity; 3 hydrolyzed 4MU-PhAc and none hydrolyzed dp0460 (Fig. 3A). One design, **rd3_N1_95**, was faster and more efficient than all of our previous serine hydrolase designs (k_{cat} = 0.0135 ± 0.0005 s^{-1}, K_m = 31 ± 3 μM), demonstrating the potential of the *N+1* motif (Fig. 3B).

We also designed general serine hydrolases for 4MU-PhAc hydrolysis using the *N+1* motif, described above. One such design, **momi120**, demonstrated particularly high efficiency. Hoping to exploit this activity for PET, we redesigned the active site by docking in a PET dimer and performing sequence design with LigandMPNN. Of 159 designs ordered, all hydrolyzed 4MU-PhAc and two hydrolyzed dp0460 (fig. S3A,B). We characterized the kinetic parameters of the two most active designs, which both have catalytic efficiencies on the order of 10^5 M^{-1} s^{-1} on 4MU-PhAc (Fig. 3D,F). To our knowledge, these catalytic efficiencies surpass that of any previously reported *de novo* designed enzyme, and begin to approach the range seen in natural enzymes (10^5 - 10^8 M^{-1} s^{-1}).

We then performed sequence redesign of **momi120-103** for 4MU-PhAc hydrolysis to determine if explicit design for the activated ester substrate could enhance the k_{cat}. Interestingly, the most active design from this set, **rd3_11**, was more active on both 4MU-PhAc and dp0460 than previous designs (Fig. 3H, fig. S4A,B). Due to its activity on dp0460, we proceeded to evaluate its capability to hydrolyze bulk PET. The enzyme was incubated with PET film and powder, and the supernatant was run on HPLC to monitor the accumulation of MHET and TPA, suggesting degradation. In both, breakdown products above background could be observed (Fig. 3I). As high temperatures are generally needed for industrial plastic recycling conditions, we decided to test the thermostability of **rd3_11**. Excitingly, it showed a T_m > 95 °C by circular dichroism spectroscopy, which is higher than that of any reported PETase (Fig. 3J).

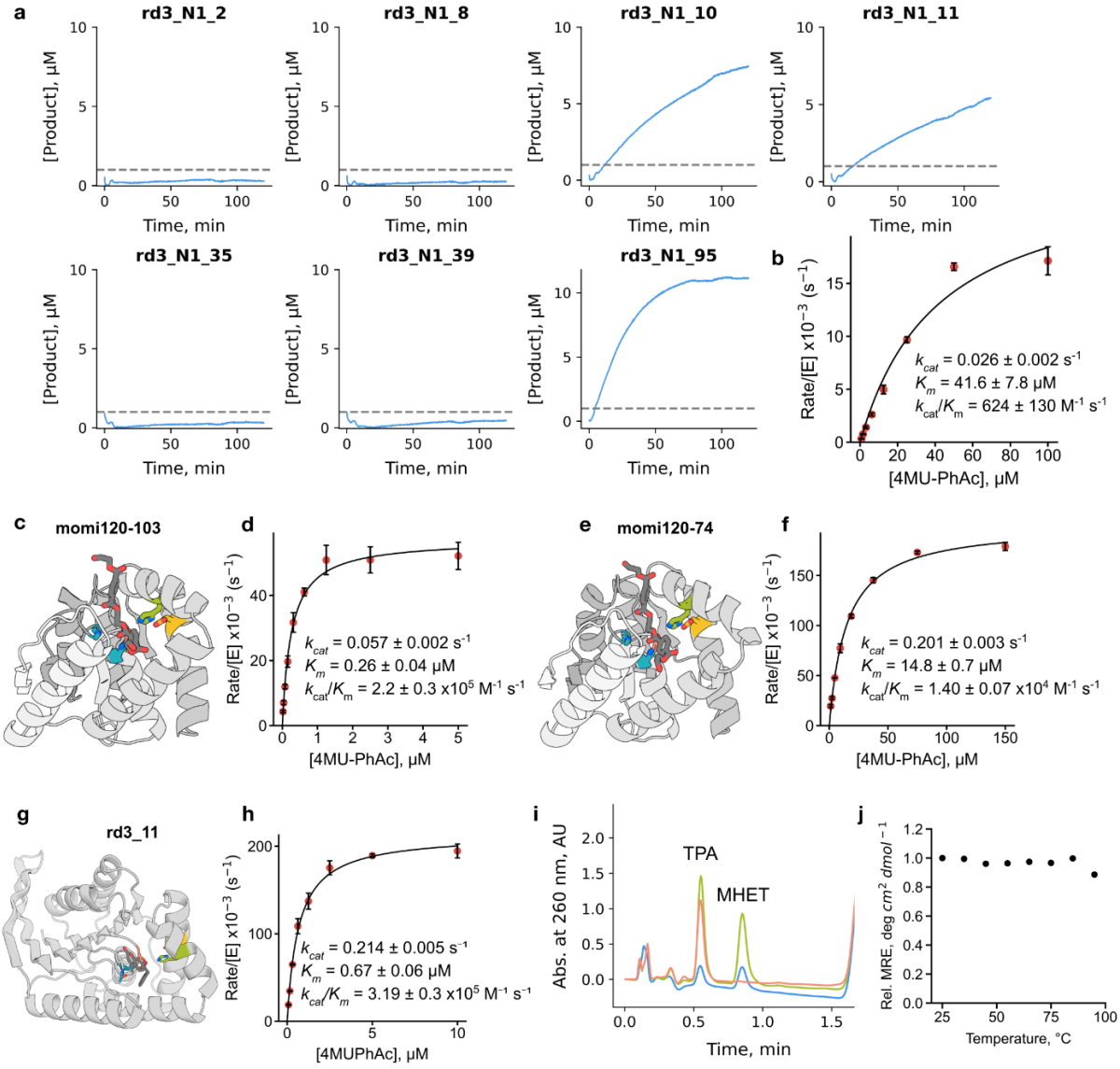

Figure 3. Complex folds enable unprecedented efficiencies and activities. (**A**) Reaction progress curves of 4MU-PhAc substrate incubated with purified designs from round 3 utilizing the *N+1* motif. Three of seven tested performed multiple turnovers of the substrate. (**B**) Kinetic parameters of **rd3_N1_95**. One-shot design with *N+1* motifs resulted in parameters rivaling those of optimized designs from *N* motif campaigns. Design models (**C,E,G**) and Michaelis-Menten plots (**D,F,H**) for active designs with distinct folds. (**I**) Hydrolysis products are generated from amorphous PET film after incubation with **rd3_11**. (**J**) CD melting temperature plot of **rd3_11** (signal reported in molar residue ellipticity (MRE)).

4.3 – Conclusion

To our knowledge, **rd3_11** is the first completely *de novo* designed enzyme which can hydrolyze bulk PET. Additionally, with a catalytic efficiency of $3.2x10^5$ M^{-1} s^{-1} on 4MU-PhAc, it is the most efficient designed enzyme reported to date. This work represents a significant advance in multiple aspects of enzyme design: unprecedented catalytic efficiencies, scaffolding of a complex active site constellation, and hydrolysis of a chemically- and physically-challenging substrate. While **rd3_11** has significantly lower activity than optimized PETases, it has extremely high soluble yield and thermostability, and thus offers improved carbon efficiency – a key principle of green chemistry. Additionally, we envision that the platform developed here can be broadly applied to many polymer types. For other polyesters, a similar serine hydrolase motif can be utilized with active site design to change substrate binding specificity. Other classes of polymers may necessitate different mechanisms, but with the correct theozymes can be designed accordingly.

4.4 – Methods

Note: Many methods utilized are described in Section 3.7 of this document. Below are methods unique to this project.

Screening with PET fluorescent reporter. Kinetic screens were performed in 40 μL reaction volumes in 96-well half area plates. Protein and substrate were prepared fresh in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO. An enzyme solution at 2X concentration (2 μM) was added to the microtiter plate and reactions were initiated by addition of a 2X concentrated substrate solution. Generation of the fluorogenic product was monitored continuously (excitation 310 nm, emission 430 nm) on a Neo2 plate reader with incubation at 40 °C.

Screening with bulk PET substrate. (Adapted from the method described in Bell *et al.* 2022 (*12*).) Degradation assays were performed in 500 μL reaction volumes in 1.6-mL Eppendorf tubes. Protein and

substrate were prepared fresh in 50 mM Gly-OH, pH 9.2 4% BugBuster. Either a 6 mm x 6 mm square of amorphous PET film (Goodfellow ES30-FM-000145) or 13 mg of crystalline PET powder (Goodfellow ES30-PD-000132) was added to the tube and reactions were initiated with addition of enzyme. Hydrolysis products were analyzed by HPLC as described below.

Chromatographic analysis of PET degradation products. (Adapted from the method described in Bell *et al.* 2022 (*12*).) Samples were analyzed by HPLC using a Kinetex XB-C18 100 Å, 5 µm, 50 × 2.1 mm, LC Column with a stepped, isocratic solvent ratio method. Mobile phase A was water containing 0.1% formic acid and mobile phase B was acetonitrile at a flow rate of 1.1 ml min^{-1}. 2 µl of sample was injected, after which the mobile phase was set to 13% buffer B for 52 s to separate TPA and MHET, stepped up to 95% buffer B for 33 s to separate larger reaction products and contaminants, and then stepped back down to 13% buffer B for column re-equilibration until a total run time of 1.8 min. Using this method, TPA is eluted at roughly 0.4 min, MHET at around 0.6 min and small amounts of bis(2-hydroxyethyl) terephthalate (BHET) and longer oligomers at around 1–1.2 min.

Conclusion

Natural enzymes have evolved to catalyze a diversity of reactions with exquisite selectivity and efficiency, making them invaluable biochemical tools. While protein design holds the promise of the generation of completely bespoke enzymes for any reaction, enzyme design remains an outstanding challenge in the field due to the complexity and precision of the catalytic mechanisms employed. Here, I describe recent technologies which have significantly enhanced our ability to effectively design enzymes, as well as novel methods to apply them to the design of highly efficient serine hydrolases and plastic-degrading enzymes.

This work represents a significant advance in the field, with enzymes presented here reaching catalytic efficiencies of 10^5 M^{-1} s^{-1} – the highest of any designed enzyme. Additionally, evidence of bulk PET hydrolysis suggests that we were successful in generating the first *de novo* designed PETase. The

designs possess entirely unique topologies, are significantly smaller than their natural counterparts, and produce high soluble yield in heterologous expression systems. These physical features lend well to large-scale production, enabling application in industry.

We envision that these findings will support further enzyme design campaigns for new chemistries of interest. The catalytic motif utilized here performs a broad diversity of reactions in natural enzymes and thus can be directly co-opted for compatible substrates. For reactions that cannot be catalyzed by the serine hydrolase motif, our methods can be adapted while still probing the same critical features: active site preorganization, ligand binding, and protein folding. Many natural enzymes catalyze valuable reactions, but are not used at industrial scale due to their unfavorable properties. The work described above, wherein the physical properties of valuable natural enzymes were enhanced through sequence redesign with the same deep learning tools, additionally enables the utilization of such problematic enzymes when *de novo* design may be infeasible.

While significant progress has been made in the commercial application of enzymatic PET depolymerization, most unnatural polymers are not yet known to be degraded by enzymes. Amongst these polymers there is essentially infinite structural diversity as, although they are composed of few backbones, their physical properties are modulated by the incorporation of unique side chain substituents. Fewer mechanisms are known for the depolymerization of plastics utilizing amide or olefin backbones as compared to polyesters (such as PET), but designing enzymes to cleave these functional groups could lead to enzymatic degraders for entire plastic classes. As methods advance, degradation of recalcitrant substrates will become increasingly feasible, as the work here demonstrates.

Supplementary Information

Chapter 2 Supplement

Computational details

Myoglobin backbone idealization with inpainting. The backbone idealization of Rosetta-relaxed crystal structure of human myoglobin (PDB: 3RGK) was performed using RoseTTAFold joint inpainting (*52*). Two separate design trajectories were performed. In first, the following regions were considered for idealization: 9 N-terminal residues, 10 C-terminal residues, positions 73-88 connecting the E and F helices. In the second strategy, in addition to the above, positions 47-59 in the CD-loop region were considered for remodeling (Figure 2A). Furthermore, positions in the fixed parts of the protein that are in contact with the remodeled regions and are not part of the heme binding site were allowed to be redesigned using the "inpaint_seq" option.

The following settings were included in the input JSON files to perform the design:

Strategy 1:

```
[{"pdb": "../3RGK_fr.pdb",
"task": "hal",
"dump_all": true,
"inf_method": "multi_shot",
"n_cycle": 15,
"num_designs": 20,
"tmpl_conf": "0.9",
"contigs": ["6-10,A10-72,14-19,A89-139,8-12"],
"inpaint_seq": ["A130","A134","A137"],
"out": "3RGK_inpaint1"}]
```

Strategy 2:

```
[{"pdb": "../3RGK_fr.pdb",
"task": "hal",
"dump_all": true,
"inf_method": "multi_shot",
```

```
"n_cycle": 15,
"num_designs": 10,
"tmpl_conf": "0.9",
"contigs": ["6-10,A10-46,10-16,A60-72,14-19,A89-139,8-12"],
"inpaint_seq": ["A26","A30","A34","A62","A130","A134","A137"],
"out": "3RGK_inpaint2"}]
```

ProteinMPNN design of myoglobin. The following command was used to perform ProteinMPNN(*17*) sequence redesign of the native myoglobin as well as the structures obtained from inpainting backbone idealization.

```
python $MPNN_PATH/protein_mpnn_run.py --jsonl_path ../parsed_pdbs_bb.jsonl
--fixed_positions_jsonl ../masked_pos.jsonl --batch_size 1 --out_folder ./
--num_seq_per_target 20 --sampling_temp "0.1 0.2 0.3" --omit_AAs='MC'
--checkpoint_path $MPNN_PATH/vanilla_model_weights/v_48_020.pt
```

Where `parsed_pdbs_bb.jsonl` contains the parsed PDB file information, created with the script

`$MPNN_PATH/helper_scripts/parse_multiple_chains.py`

`masked_pos.jsonl` file contains the positions that were kept fixed during sequence design:

```
{"3RGK": {"A": [39, 42, 43, 45, 64, 67, 68, 71, 72, 89, 92, 93, 97, 99, 104,
107, 138]}}
```

For each of the outputs from the inpainting backbone idealization, the fixed position numbers were readjusted to correspond to the positions in the parent structure.

ProteinMPNN design of TEV protease. The following command was used to perform sequence design with ProteinMPNN on TEV protease.

```
python $MPNN_PATH/protein_mpnn_run.py \
        --jsonl_path ../parsed_pdbs_bb.jsonl \
        --chain_id_jsonl ../assigned_chains.jsonl \
        --fixed_positions_jsonl ../masked_pos.jsonl \
        --out_folder $MPNN_OUTDIR \
        --num_seq_per_target 16 \
        --sampling_temp "0.1 0.2 0.3" \
        --batch_size 8 \
```

```
        --omit_AAs='XC'
```

Where `../assigned_chains.jsonl` contains the parsed PDB chain information: `{"TEVd": [["A"]]}`

Sets of designs were distinguished by selection of fixed residues.

Designs with only the amino acid identities of the active site fixed during sequence design had the following residues fixed:

```
[31, 32, 44, 46, 81, 134, 135, 139, 146, 147, 148, 149, 150,151, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178,204, 208, 209, 211, 213, 214, 215, 216, 217, 218, 219, 220]
```

Designs with the amino acid identities of active site residues and the 30% most conserved residues fixed during sequence design had the following residues fixed:

```
[3, 7, 9, 10, 11, 12, 14, 19, 25, 34, 36, 38, 42, 44, 46, 47, 48, 51, 52, 53, 55, 61, 62, 64, 68, 81, 88, 89, 90, 92, 94, 100, 101, 103, 110, 113, 116, 117, 126, 127, 129, 139, 140, 142, 143, 144, 146, 149, 151, 152, 154, 156, 160, 161, 163, 165, 167, 169, 177, 186, 190, 198, 202, 211, 212, 221]
```

Designs with the amino acid identities of active site residues and the 50% most conserved residues fixed during sequence design had the following residues fixed:

```
[2, 3, 7, 8, 9, 10, 11, 12, 13, 14, 21, 23, 25, 26, 27, 31, 32, 34, 35, 36, 37, 38, 41, 42, 43, 44, 46, 47, 48, 51, 52, 53, 55, 59, 61, 62, 64, 68, 70, 72, 76, 81, 85, 88, 89, 90, 91, 92, 93, 94, 95, 98, 100, 101, 103, 107, 109, 112, 113, 115, 116, 117, 119, 123, 125, 126, 127, 129, 133, 134, 135, 139, 140, 141, 142, 143, 144, 146, 147, 148, 149, 150, 151, 152, 153, 154, 156, 157, 160, 161, 163, 165, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 182, 183, 186, 190, 198, 200, 202, 204, 205, 208, 209, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221]
```

Designs with the amino acid identities of active site residues and the 70% most conserved residues fixed during sequence design had the following residues fixed:

```
[1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 21, 22, 23, 25, 26, 27, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 53,
```

55, 57, 59, 61, 62, 63, 64, 66, 68, 69, 70, 71, 72, 73, 76, 79, 80, 81, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 100, 101, 103, 107, 108, 109, 111, 112, 113, 115, 116, 117, 118, 119, 120, 122, 123, 124, 125, 126, 127, 129, 131, 133, 134, 135, 137, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 160, 161, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 182, 183, 186, 187, 189, 190, 194, 196, 198, 200, 202, 203, 204, 205, 206, 207, 208, 209, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221]

Supplementary Figures

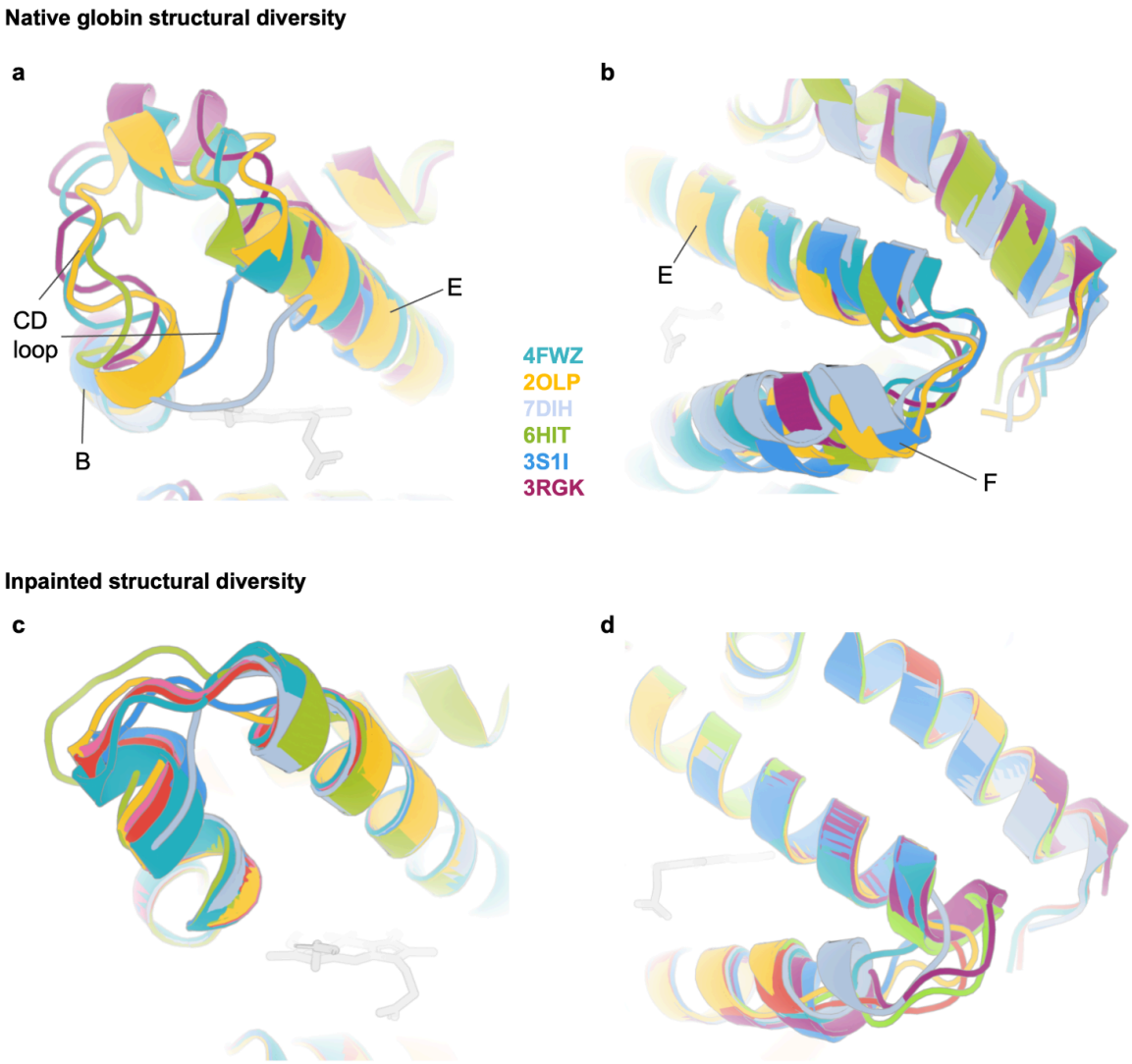

Figure S1. Inpainting samples different backbone structures compared to native globins. (A) Diversity of the CD-loop region in selected native globins. (B) Diversity in the loop connecting helices E and F in selected native globins. (C) Diversity of inpainted motifs replacing the CD-loop region. (D) Diversity of inpainted loops connecting helices E and F.

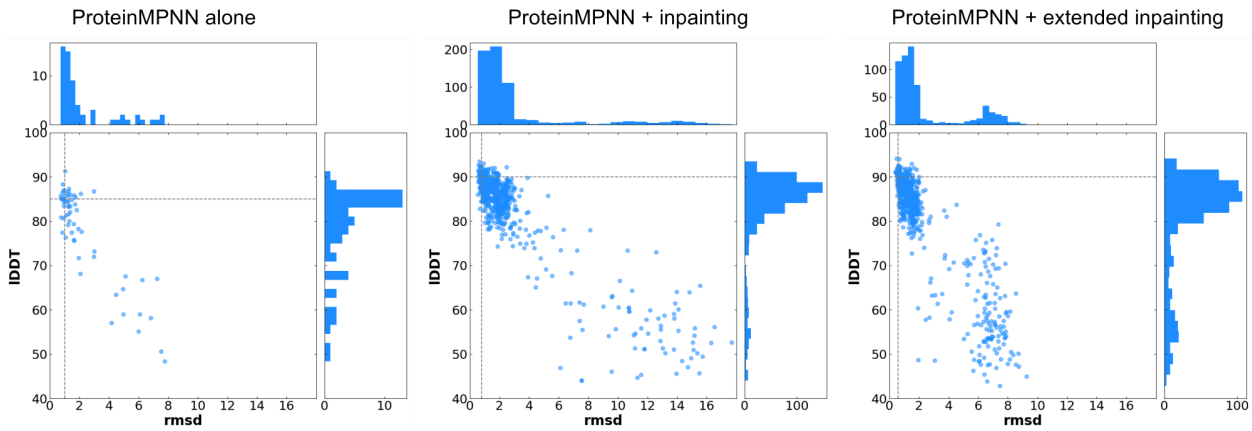

Figure S2. Extensive backbone remodeling with RoseTTAFold joint inpainting improves structure prediction metrics. Designs made with only sequence redesign had the lowest-scoring structure prediction metrics (lDDT and RMSD to design model) amongst all designs, while designs subjected to the most aggressive backbone remodeling strategy scored the highest in these metrics. Dashed lines indicate lDDT and RMSD cutoffs used for design selection, with the top left sector containing successful designs.

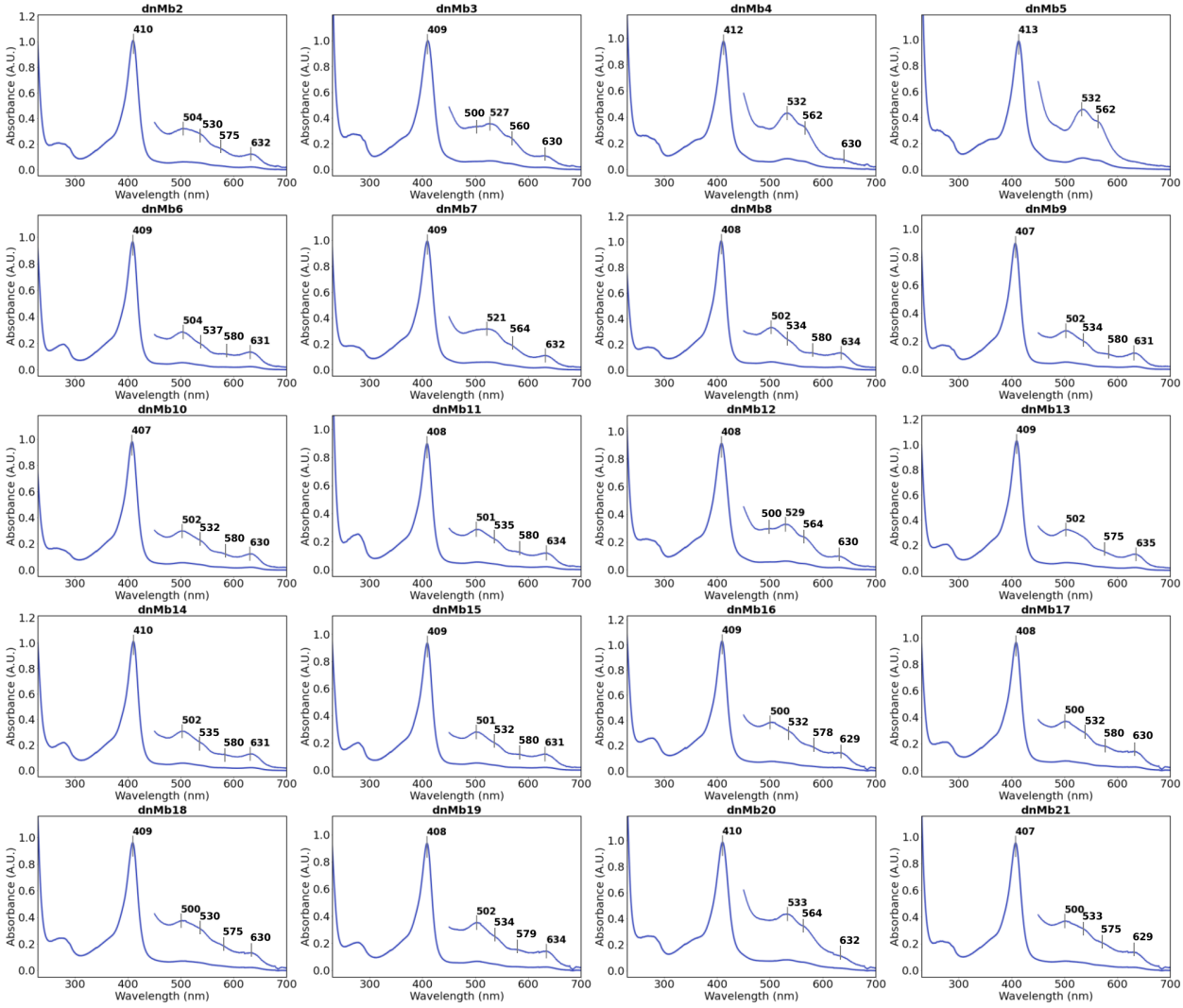

Figure S3. UV/Vis spectra of myoglobin variants. Spectroscopic data of most designs is in close agreement with that of native myoglobin (Soret maximum at 409 nm; Q band features at 500, 537, 582 and 630 nm), suggesting pentacoordinate heme-binding. A few designs (dnMb3, dnMb4, dnMb5, dnMb12, and dnMb20) show some degree of hexacoordinate heme-binding (potentially through incorporation of imidazole from the purification buffer), indicated by the major Q band features at ~530 and ~560 nm. Spectra were recorded in a buffer containing 25 mM Tris-HCl and 300 mM NaCl at pH 8.2.

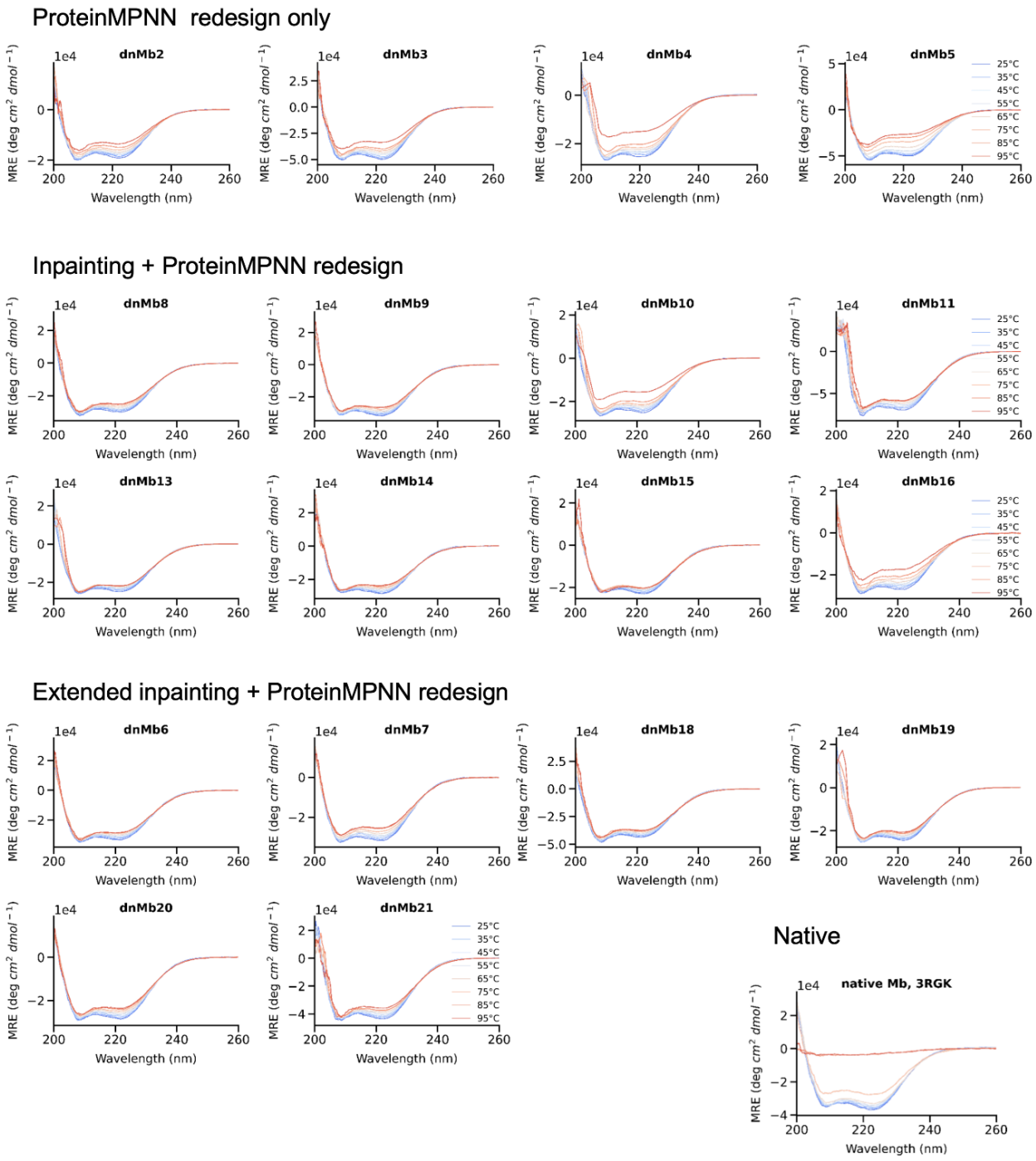

Figure S4. Many myoglobin designs show increased thermostability over parent. CD spectroscopy signal of myoglobin designs and parent sequence nMb over a temperature gradient from 25 °C to 95 °C indicates elevated resistance to unfolding in designs. CD signal reported in molar residue ellipticity (MRE).

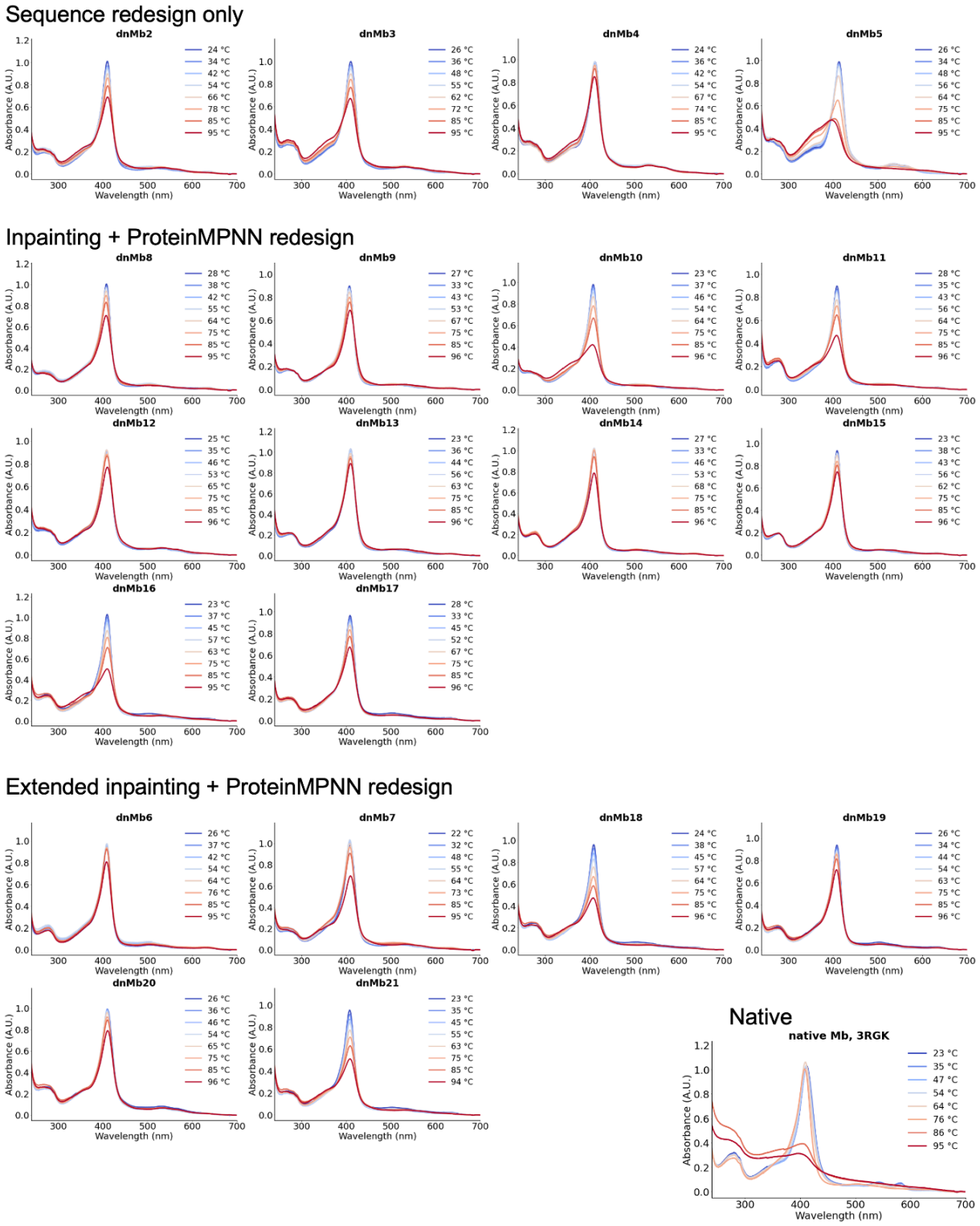

Figure S5. Myoglobin designs retain heme binding at higher temperatures than parent. Heme binding as measured by UV/Vis absorbance over a temperature gradient from 25 °C to 95 °C indicates retention of function at higher temperatures in designs. Higher melting temperatures of designs indicate more temperature-stable binding sites.

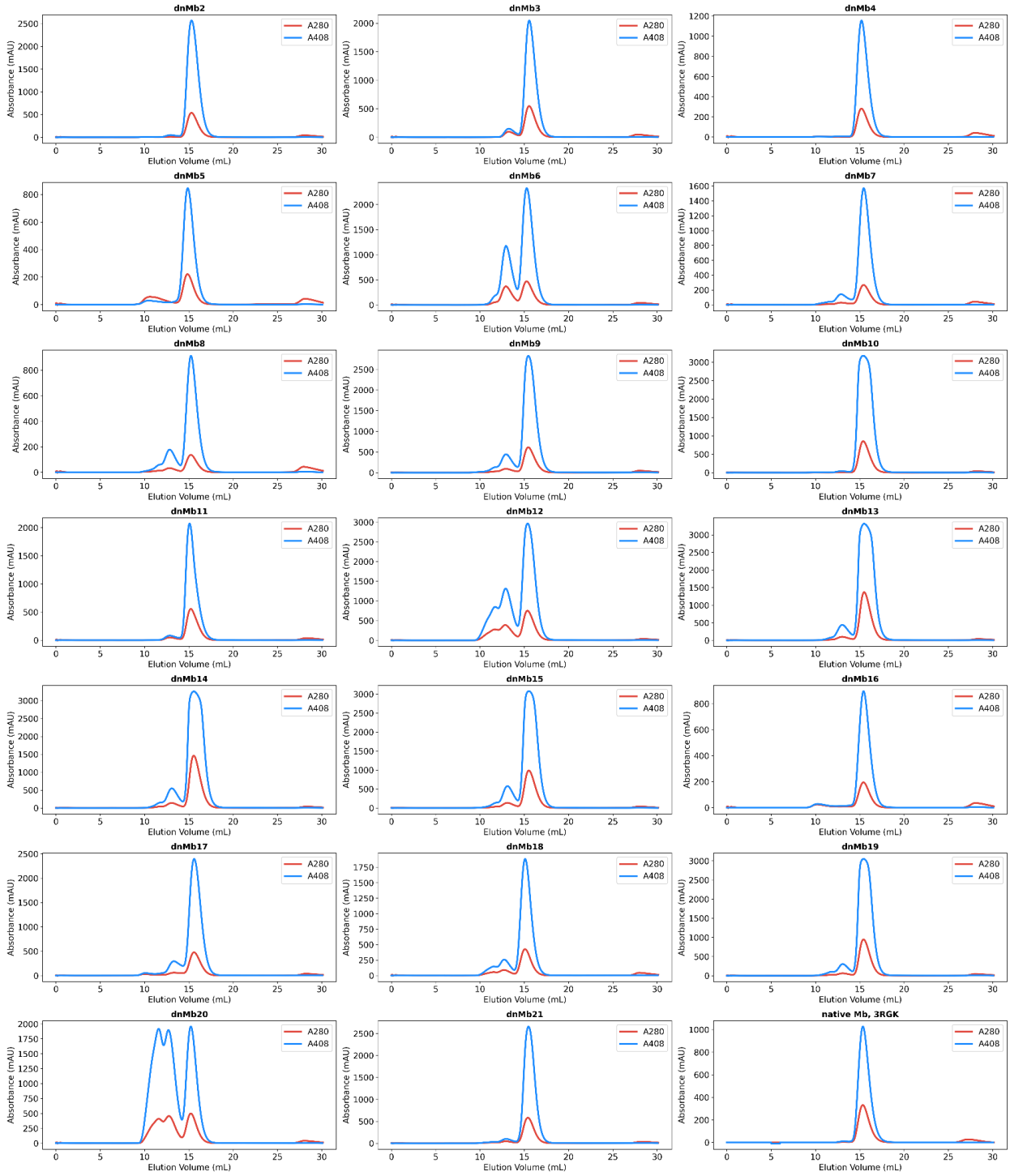

Figure S6. Size-exclusion chromatograms of heme-loaded myoglobin variants. Data were collected using a Superdex Increase 75 10/300 GL column (GE Healthcare) in a buffer containing 25 mM Tris-HCl and 300 mM NaCl at pH 8.2. Void volume of the column is 8.5 mL. Blue chromatograms were obtained by following the absorbance at 408 nm, indicating elution of heme-containing species. Red chromatograms were obtained from absorbance at 280 nm.

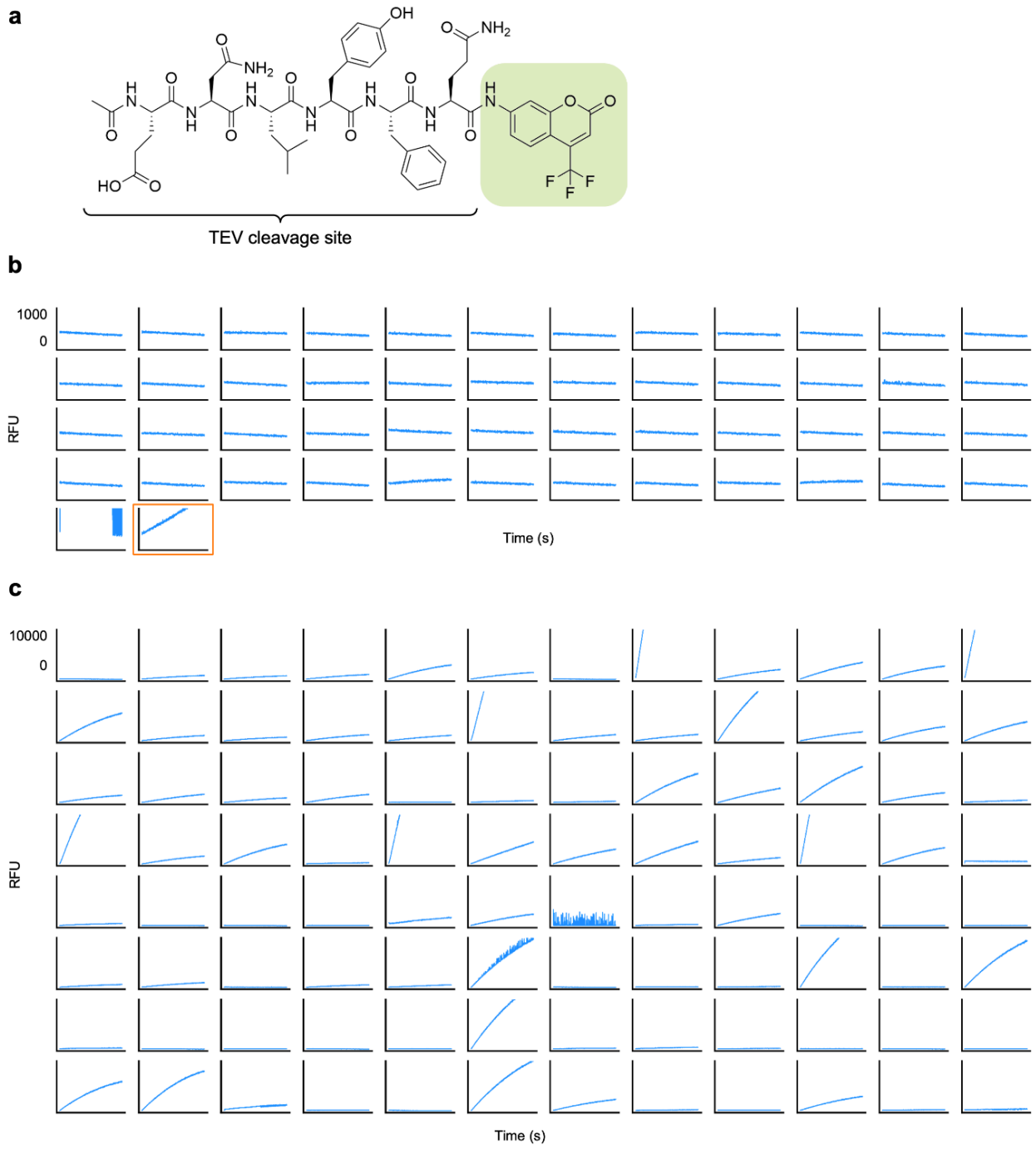

Figure S7. Initial screen of proteolytic activity on fluorescent reporter substrate. Pure protein was normalized to 1 μM and assayed against 10 μM substrate, AFC, in an initial screen for catalytic turnover. (A) Structure of the peptide-coumarin substrate, AFC, used to assay proteolytic activity. (B) Raw fluorescence data (in raw fluorescence units, RFU) for designs generated with only active site residues fixed or with active site residues and 30% most conserved residues fixed during design. TEVd plot outlined in orange. (C) Raw fluorescence data for designs generated with active site residues fixed and 50% most conserved residues fixed or with active site residues fixed and 70% most conserved residues fixed.

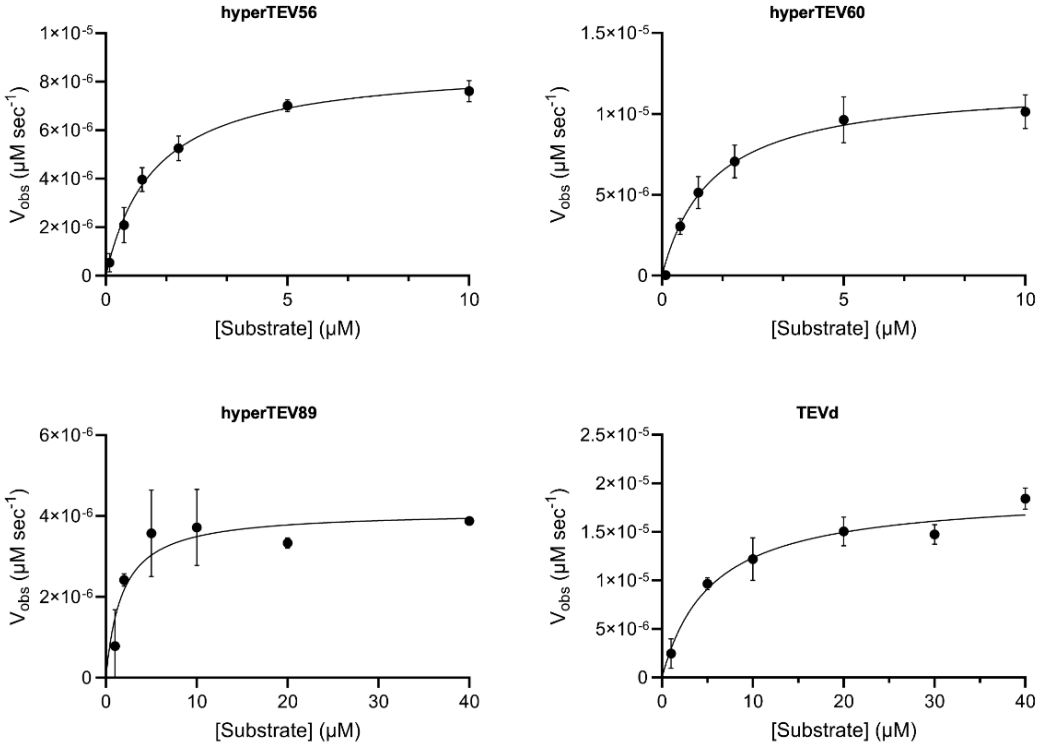

Figure S8. Michaelis Menten kinetics of TEV redesigns and parent. Michaelis Menten plots for three TEV designs and TEVd. Error bars represent standard deviation from three technical replicates. hyperTEV designs were assayed at 50 nM while TEVd was assayed at 500 nM.

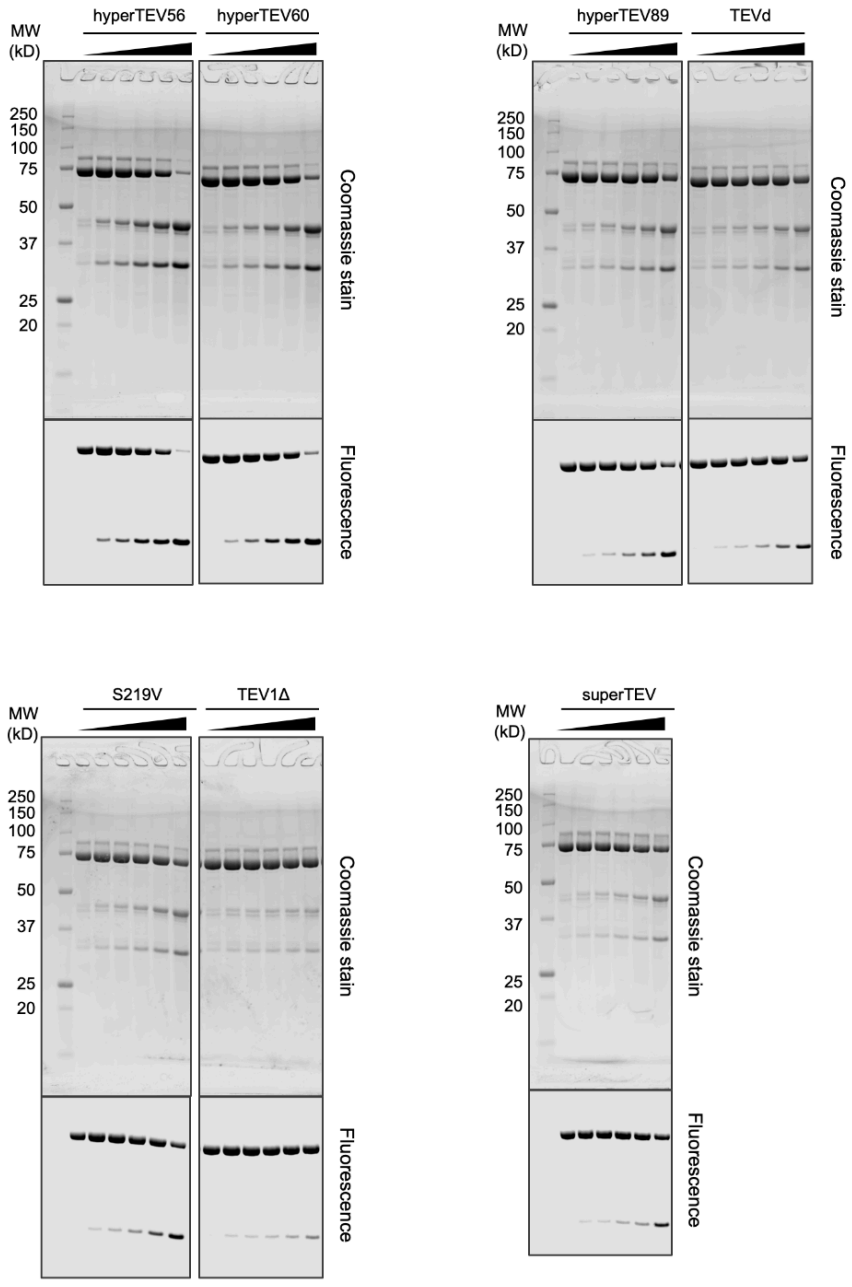

Figure S9. SDS-PAGE gels of protein substrate cleavage by TEV designs. Protein standard molecular weight ladder is shown on the left, with molecular weight markers indicated in kD. For each gel, the coomassie-stain is shown above and the EGFP fluorescence image is shown below.

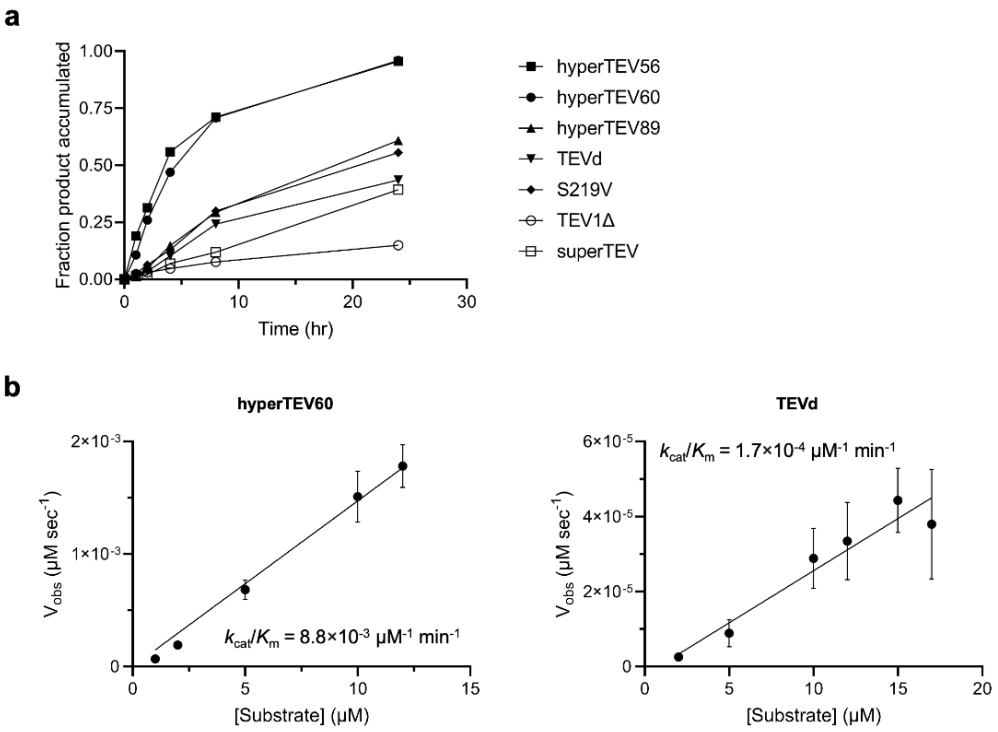

Figure S10. Activity of TEV redesigns in a gel-based activity assay. (A) Plot of accumulated product normalized to fluorescence intensity of uncleaved substrate over time. Fluorescence intensity was quantified with ImageJ software. Designs hyperTEV56 and hyperTEV60 show increased turnover rate compared to reported TEV variants. (B) Straight-line fit for initial turnover rates in gel assay for hyperTEV60 and TEVd. Curves were fitted from monitoring of substrate depletion for hyperTEV60 and production accumulation for TEVd. Error bars represent standard deviation from three technical replicates.

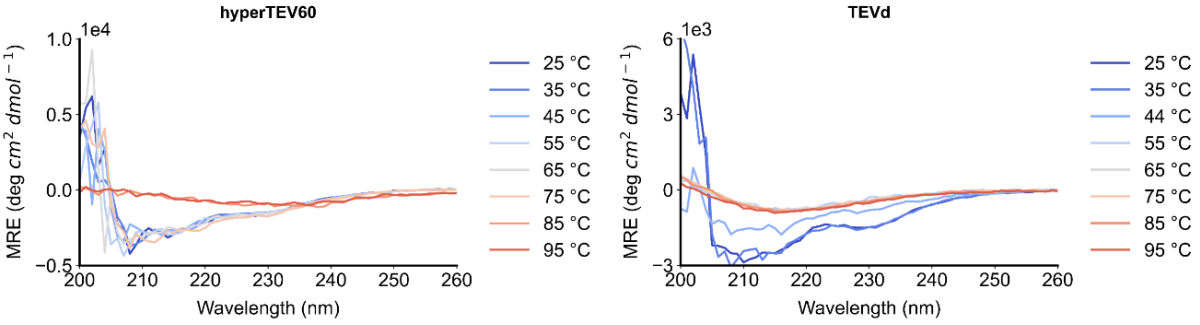

Figure S11. TEV design shows increased thermostability over parent. CD spectroscopy signal of hyperTEV60 and TEVd over a temperature gradient from 25 °C to 95 °C indicates elevated resistance to unfolding in ProteinMPNN design hyperTEV60. CD signal reported in molar residue ellipticity (MRE).

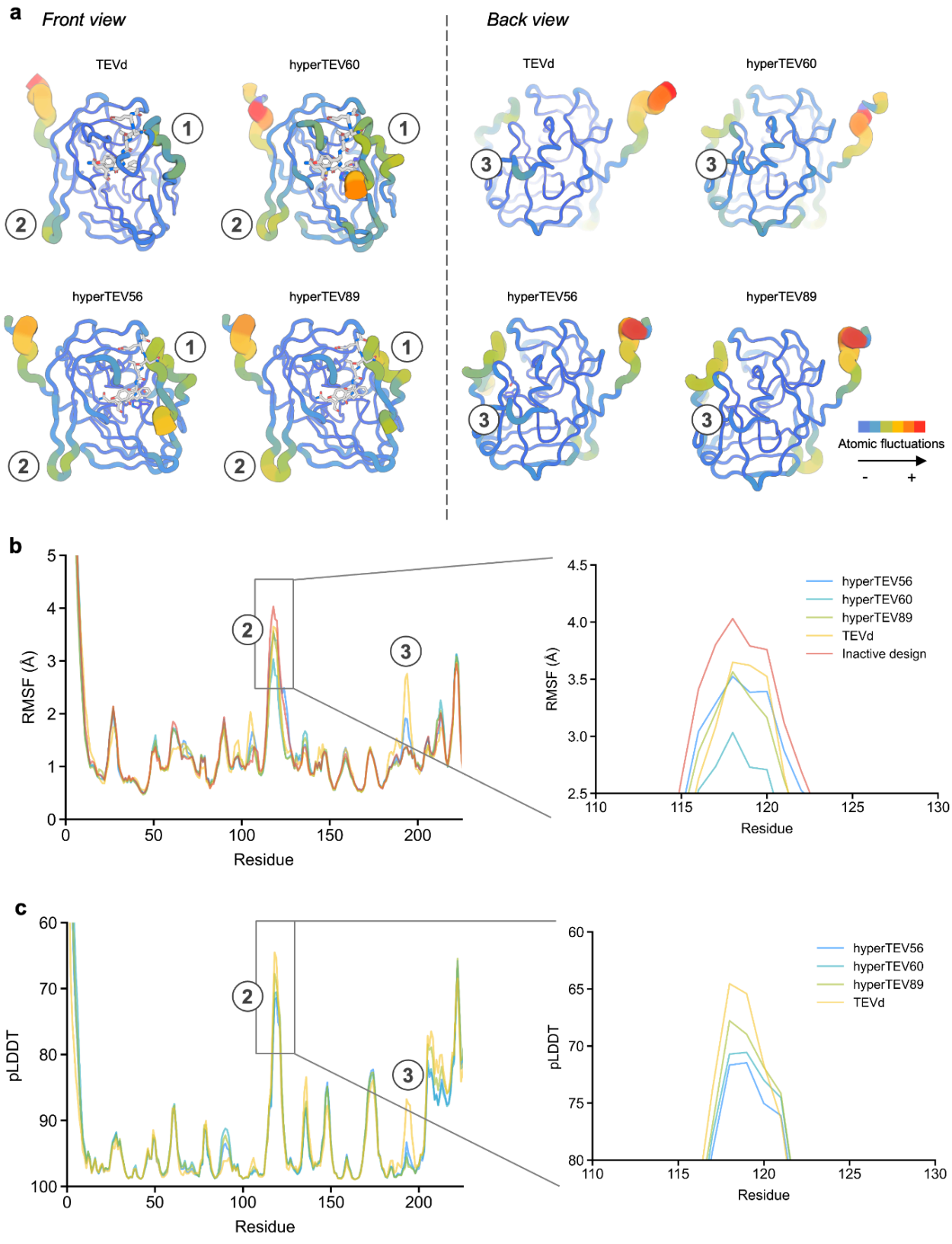

Figure S12. Designs show trends in rigidification and activity in molecular dynamics simulations. (A) Molecular dynamics (MD) simulations revealed trends of rigidification of several loops (marked with numbered circles) in the

redesigned structures as compared to the parent. Directly adjacent to the peptide binding site, region 1 (residues 206-215) shows diminishing mobility in hyperTEV60 and other redesigns as compared to TEVd. An internal loop designated as region 3 (residues 192-194) shows significant loss of atomic fluctuation relative to TEVd. (B) Cα root mean square fluctuation (RMSF) of designs in region 2 (residues 115-124) denoted in (A) shows a positive correlation between activity and rigidification, with TEVd and a design inactive on the peptide substrate showing most flexibility in this region. (C) Per-residue pLDDT values from AlphaFold2 ensemble prediction exhibit similar trends of increased rigidification in more active designs.

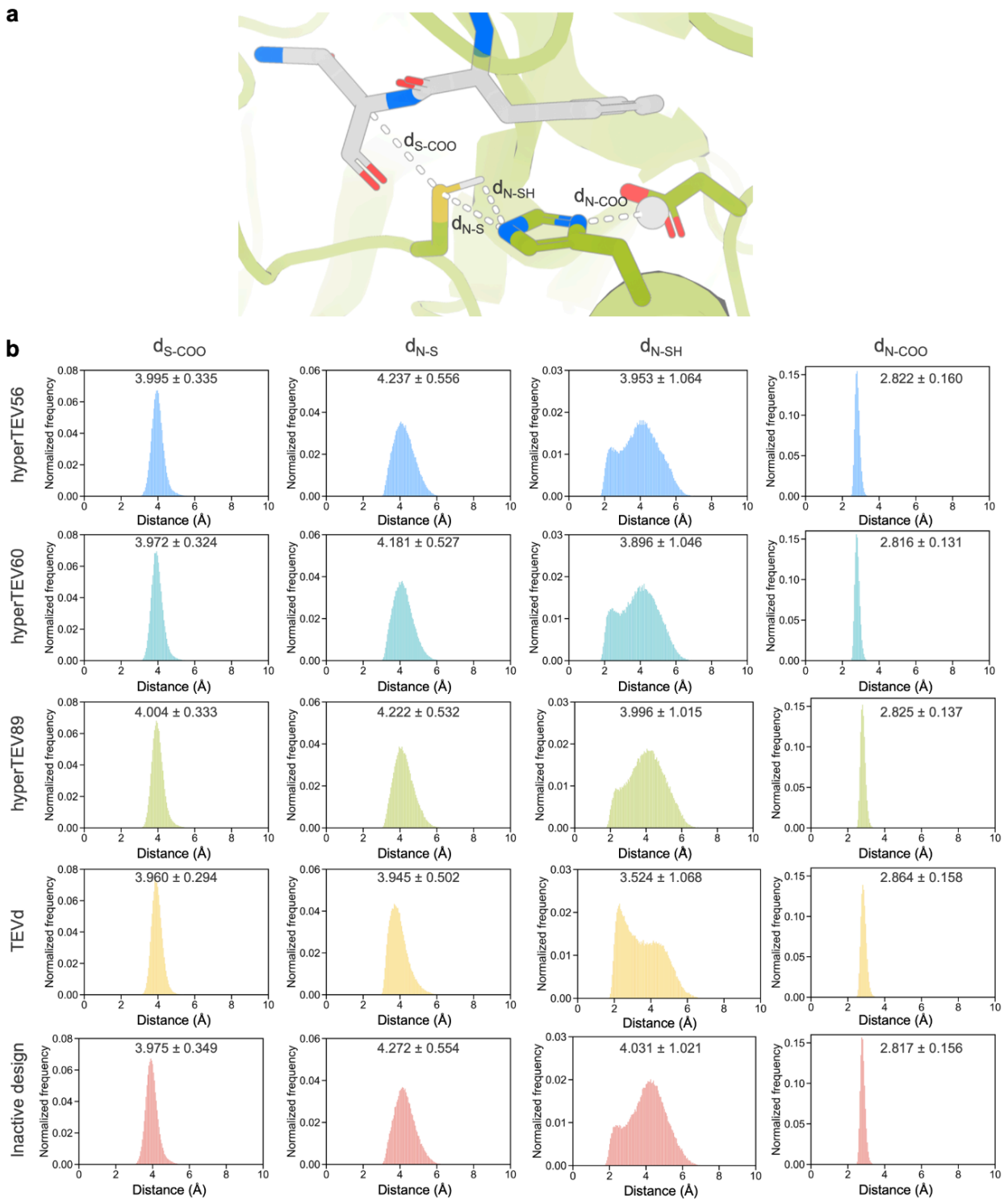

Figure S13. Population of catalytically competent dyad conformers correlates with activity in MD simulations. (A) Key distances in the TEV catalytic triad as shown on TEVd. Peptide substrate is shown in gray. (B) Distances of key interactions in the catalytic triad were measured across MD simulations. Average distances for each interaction in each TEV variant are inset. Catalytically competent conformers of the Cys-His dyad ($d_{N\text{-}SH}$) are less populated in designs as compared to TEVd. Among designs, the highest activity variant hyperTEV60 has the highest percentage of competent dyad conformers.

Table S1. Myoglobin sequence similarity analysis against UniRef100

Variant	Design method	Sequence similarity with parent (3RGK)	Highest sequence similarity	Most similar UniRef100 ID
dnMb2	ProteinMPNN only	47%	51%	P02182
dnMb3	ProteinMPNN only	49%	52%	UPI00148EC9BB
dnMb4	ProteinMPNN only	51%	55%	P02182
dnMb5	ProteinMPNN only	46%	51%	Q0KIY3
dnMb6	Inpaint CD+EH; ProteinMPNN	40%	44%	A0A8C5V5K1
dnMb7	Inpaint CD+EH; ProteinMPNN	42%	48%	UPI001C20C4EB
dnMb8	Inpaint EH; ProteinMPNN	45%	47%	A0A8C9A9W3
dnMb9	Inpaint EH; ProteinMPNN	46%	53%	UPI001C20C4EB
dnMb10	Inpaint EH; ProteinMPNN	46%	51%	UPI00148EC9BB
dnMb11	Inpaint EH; ProteinMPNN	49%	55%	P02185
dnMb12	Inpaint EH; ProteinMPNN	51%	54%	A0A4W2F1N8
dnMb13	Inpaint EH; ProteinMPNN	44%	49%	UPI001CA46E1B
dnMb14	Inpaint EH; ProteinMPNN	46%	51%	A0A8C9A9W3
dnMb15	Inpaint EH; ProteinMPNN	46%	50%	UPI000011026E
dnMb16	Inpaint EH; ProteinMPNN	39%	45%	P02169
dnMb17	Inpaint EH; ProteinMPNN	42%	45%	F6PMG4
dnMb18	Inpaint CD+EH; ProteinMPNN	42%	43%	R9RZ90
dnMb19	Inpaint CD+EH; ProteinMPNN	39%	41%	UPI0003C8C8C2
dnMb20	Inpaint CD+EH; ProteinMPNN	45%	48%	P02182
dnMb21	Inpaint CD+EH; ProteinMPNN	39%	44%	P02182

Table S2. Mass spectrometry data for myoglobin variants.

Variant	Expected mass	Observed mass
dnMb2 (Met missing)	19186	19054
dnMb3 (Met missing)	18981	18850
dnMb4 (Met missing)	19054	18923
dnMb5 (Met missing)	18441	18310
dnMb6 (Met missing)	19604	19473
dnMb7 (Met missing)	18728	18597
dnMb8 (Met missing)	19604	19472
dnMb9 (Met missing)	19579	19448
dnMb10 (Met missing)	18690	18559
dnMb11 (Met missing)	19464	19333
dnMb12 (Met missing)	19536	19405
dnMb13 (Met missing)	19178	19047
dnMb14 (Met missing)	19675	19544
dnMb15 (Met partially missing)	19598	19467,19598
dnMb16 (Met partially missing)	19441	19310,19441
dnMb17 (Met missing)	20247	20115
dnMb18 (Met partially missing)	19427	19296,19427
dnMb19 (Met missing)	19452	19321
dnMb20 (Met partially missing)	18814	18683,18814
dnMb21 (Met missing)	19133	19002
nMb 3RGK (Met missing)	18751	18620

Table S3. The extinction coefficients of the Soret band and R_z values (A_{Soret} / A_{280}) of myoglobin variants.

Variant	**Extinction coefficient (mM^{-1} cm^{-1})**	**R_z**
dnMb2	128 ± 3	5.7
dnMb3	166 ± 15	4.0
dnMb4	127 ± 1	4.6
dnMb5	154 ± 8	4.9
dnMb6	181 ± 14	5.4
dnMb7	159 ± 6	5.8
dnMb8	186 ± 2	6.8
dnMb9	182 ± 5	5.5
dnMb10	157 ± 2	7.4
dnMb11	177 ± 8	4.1
dnMb12	123 ± 3	4.9
dnMb13	154 ± 2	5.2
dnMb14	174 ± 1	4.8
dnMb15	156 ± 1	4.9
dnMb16	171 ± 4	4.5
dnMb17	170 ± 1	5.4
dnMb18	175 ± 15	3.1
dnMb19	171 ± 3	5.1
dnMb20	150 ± 4	3.7
dnMb21	153 ± 1	4.6

Crystallographic data

Protein sample for crystallography was prepared following the general procedure for myoglobin production. The holoprotein was purified using Ni-affinity and size exclusion chromatography. The C-terminal hexahistidine tag was left intact. The holo dnMb19 was crystallized at 17 mg mL−1 in a buffer containing 25 mM Tris-HCl, 300 mM NaCl, pH 8.2.

The crystallization experiment for the designed protein was conducted using the sitting drop vapor diffusion method. Crystallization trials were set up in 200 nL drops using the 96-well plate format at 20°C. Crystallization plates were set up using a Mosquito LCP from SPT Labtech, then imaged using UVEX microscopes from JAN Scientific. Diffraction quality crystals formed in 0.1 M Bis-Tris pH 6.5, 28% w/v Polyethylene glycol monomethyl ether 2,000 (Index crystallization screen, Hampton Research, well D11).

Diffraction data were collected at ALS-ENABLE beamline 8.2.2. X-ray intensities and data reduction were evaluated and integrated using XDS (*133*) and merged/scaled using Pointless/Aimless in the CCP4 program suite (*145*). Structure determination and refinement starting phases were obtained by molecular replacement using Phaser (*146*) using the designed model structure. Following molecular replacement, the models were improved using phenix.autobuild (*147*). Structures were refined in Phenix (*147*). Model building was performed using COOT (*135*). The final model was evaluated using MolProbity (*148*). Data collection and refinement statistics are recorded in Table S4. Data deposition, atomic coordinates, and structure factors reported for the protein in this paper have been deposited in the Protein Data Bank (PDB), http://www.rcsb.org/ with accession code 8U5A.

Table S4. Crystallographic statistics for dnMb19.

	dnMb19
PDB accession number	8U5A
Wavelength (Å)	1.0
Resolution range	42.64 - 2.0 (2.05 - 2.0)
Space group	P 1 2_1 1
Unit cell dimensions a, b, c, (Å) α, β, γ (°)	 31.589 41.669 128.439 90 95.13 90
Unique reflections	22200 (1513)
Multiplicity	4.3 (4.1)
Completeness (%)	97.30 (95.52)
Mean I/sigma(I)	10.95 (1.58)
Wilson B-factor	36
R-merge	0.07531 (0.9919)
R-pim	0.04041 (0.5512)
CC1/2	0.997 (0.773)
Reflections used in refinement	22200 (1513)
R-work	0.2301 (0.3312)
R-free	0.2581 (0.3741)
Number of non-hydrogen atoms	2612
macromolecules	2440
ligands	87
solvent	85
Protein residues	298
RMS(bonds)	0.002
RMS(angles)	0.39
Ramachandran favored (%)	99.32
Ramachandran allowed (%)	0.68
Ramachandran outliers (%)	0
Average B-factor	44
macromolecules	44.03
ligands	40.92
solvent	46.5

Sequence information

Alignment of TEV hit sequences

```
TEVd            GESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGPFIITNKHLFRRNNGTLLVQSL    60
hyperTEV89      AESAAPGPRDYNPISSTIVRLTNTSDGHSISLFGIGFGPLIITNAHLFRRNNGTLTITSL    60
hyperTEV56      MESAAPGPRDYNPISDTIVKLTNTSDGYSISLYGIGFGPLIITNAHLFRRNNGTLTVTSK    60
hyperTEV60      AESAAPGPRDYNPISDTIVLLTNTSDGYSISLYGIGFGPLIITNAHLFRRNNGTLTITSK    60
                 **   *********.**  *** ***:: **:******:**** ********** : *

TEVd            HGVFKVKNTTTLQQHLIDGRDMIIIRMPKDFPPFPQKLKFREPQREERICLVTTNFQTKS    120
hyperTEV89      HGTFTISNTTTLKLHLIEGRDLVIIKMPKDFPPFPTTLEFREPVVGEDIVLVTRNFQDKD    120
hyperTEV56      HGTFTIENTTTLQLHLIEGRDLVIIKMPKDFPPFPTDLVFREPVEGEKITLVTRNFQTKE    120
hyperTEV60      HGTFTISNTTTLKLHLIEGRDLVLIEMPKDFPPFPTNLVFREPVVGEEIVLVTRNFQTKT    120
                **.*.:.*****: ***:***:::*.*********  * ****   * * *** *** *

TEVd            MSSMVSDTSCTFPSSDGIFWKHWIQTKDGQCGSPLVSTRDGFIVGIHSASNFTNTNNYFT    180
hyperTEV89      PTSEVSDTSTTEPSSDGVFWKHWIPTKDGQCGSPMVSVSDGSIVGIHSASNFTNTNNYFT    180
hyperTEV56      PTSEVSDVSSTYPSSDGVFWKHWIPTKDGQCGSPMVSVEDGSIVGIHSASNFTNTNNYFT    180
hyperTEV60      PTSEVSDVSTTYPSSDGVFWKHWIPTKDGQCGSPMVSVTDGSIVGIHSASNFTNTNNYFT    180
                 :* ***.* * *****:****** **********:**. ** ******************

TEVd            SVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHKVFMDKP   221
hyperTEV89      AVPPNFMDLLTDPSLQKWISGWSLNADSVDWGGHKVFMDKP   221
hyperTEV56      AVPPDFMDLLTNDSLQKWISGWSLNSDSVEWGGHKVFMDKP   221
hyperTEV60      AVPPDFMRLLTDPSLQKWVSGWSLNSDSVEWGGHKVFMDKP   221
                :** :** ***: . *:*:*** **:*** ***********
```

Alignment of myoglobin sequences

```
3RGK        -GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKA-S 58
dnMb02      -GLTEEEQKLVWEIFERFEEDLEGFGLDVLIRAFTEHPETLKKFPRFADLKSEAELRA-S 58
dnMb03      -GLTAEEQKLVRDIWAEVEKDREGFGLEVLLLTFTEHPETLKKFPRFAHLKSAEELRA-S 58
dnMb04      -GLTAEEQALVRAIWAKVREDLEGFGLAVLLKTFTEHPETLKKFPRFKDLKSEEEILA-S 58
dnMb05      -GLSDEEQALVLSIFEKVKEDLAGFGLDVLLLAFTKNPATLEKFPRFADLKSEAELLA-S 58
dnMb20      --LSEEEWKIVLEIFALVREDLAGVGAAVLERTFATHPETLKKFPRFLAAAEAGVLD--R 56
dnMb16      ---AEEEKEKVLSIFKLVEKDKKTIGSEVLIITFTKNPETKKKFPRFKDLKTVEELKA-S 56
dnMb19      ---SEEKAALVLALFDRVEADREEIGAAVLRRTFEEHPETLKKFPRFLELYKKGSPEL-D 56
dnMb18      ---DEEKKKLVLEAFELVEKDIEGIGAEVLKLTFEKHPETLEKFPRLKELHAAGSPEL-E 56
dnMb15      ---EEEEKQIVLELFAKVEEDLEGIGLEVLILTFTKHPETRKKFPRFAHLTTEAQLQA-S 56
dnMb17      IKLSEEEKKLVLEIFKLFEENLEEFGKEVLITTFTKHPETKKKFPRFAHLKTEEEFLA-S 59
dnMb13      ----DERNKLVLSAFALVREDLEEIGAEVLILTFTENPETLKKFPRFAHLKTEEELKK-S 55
dnMb14      ---EKEKNELVLKAFELIEKDLEGFGSEVLILTFTKHPETLKKFPRFKHLKTEEEFKA-S 56
dnMb21      -SLTPEELAIVKALFARVREDLEGVGAEVLRLTFEKHPETLKKFPRFLELKKAGSPEL-E 58
dnMb07      -KLTPEEKAIVLRIFALVREDRAGIGAAILRRTFEAHPETLEKFPRLRALRAAGREAELE 59
dnMb06      -KLSEEEKEIVLKIFELVEKDVEEIGLRVLELTFEKHPETLEKFPRLRELLAAGRLEELE 59
dnMb10      ---DAEKQALVASIFAKFEADLEGFGKAVLIKTFTKHPETRKKFPRFKHLKSVEELEK-S 56
dnMb11      -KLSEEEKEIVLKIFALVEKDLEGFGKEVLIKTFLKYPETLKKFPRFKHLKTEEELKA-S 58
dnMb12      LNLSPEDKAKVLEIFALVEEDLEGFGREVLILTFTKHPETLKKFPRFAHLKTEEELRA-S 59
dnMb08      IKISEEEFEIVLEIFELVKKDLAGIGKEVLILTFTKHPETLKKFPRFAHLKTVEELEA-S 59
dnMb09      SKLTEEEWKTVFKIFALVEKDLEGFGLAVLIRTFTRYPETLKKFPRFAHLKTVEELRA-S 59
                      *   :  .. :    *  :*   *   * * :** *:

3RGK        EDLKKHGATVLTALGGI---LKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVL 115
dnMb02      PRLREHGVTVLKALIKI---FKKGEDFAEEVKPLAESHSKVHKIPVSDLEVIAAAILATA 115
dnMb03      PEAKAHGVTVLDALSKI---LKKGSNFEEEIKPLAESHYKEHKIPIEDLKVIADAIIAVL 115
dnMb04      EKAKKHGVTVLTALFAI---FDKGENFEEEIKPLAESHYKEHKIPISDLKVIADAIVAVL 115
dnMb05      EKAKEHGITVLTALFAI---FEKGDDFDAEVEPLATSHTREHKIPTSDLEVIAAAILETA 115
dnMb20      ALLAAHGETVLTALIEIAES----KLDPELIKKLAESHVKEHKIPIEYLRAIADSLIAVL 112
dnMb16      EKVKDHGVTVLDALIEWARLHVEGKDYDSLVKKLAESHKKEHKIPIEDLKSIADALIEVL 116
dnMb19      ALLKEHGKTVLDALIEIARLRYSGEDYRSLIKELAKSHKEEHKIPIEDLRHIAEALLAVL 116
dnMb18      ELLKEHGATVLKALIEIARLKISGGDYLSLVKELAKSHKEEHKIPIEDLKKIAEALLEVL 116
dnMb15      PELKQHGVTVLKALITIAKLYYEGKDYESLIKELAKSHKEEHKIPIEYLEYISESILEVL 116
dnMb17      PELAKHGVTVLNALIEIAKLYLEGKDYRSLIKKLAKSHKLEHKIPIEDLKYIADAIIEVA 119
dnMb13      PLLKEHGVTVLNALIEIAELKYSGGDYESLVKELAKSHKEKHKIPIEDLKAIAEAILKVL 115
dnMb14      EELKEHGVTVLKALIEIAKLKVSGEDYDSLIKELAKSHKTKHKIPIEYLKYIADAILEVA 116
dnMb21      AELRAHGVTVLTALIELADNY---EGNNETLEKLAESHTKVHKIPVSDLKNIAAAIIEVL 115
dnMb07      ALLREHGVTVLDALIEI---V--ENDDEELLKKLAESHKTTHKIPIEHLEHIAAALLEVL 114
dnMb06      AYLREHGVTVLKALIEA---I--KNEDEELLEKLAKSHKEEHKIPIEYLKYIADSIIEVL 114
dnMb10      EELKEHGVTVLTALREIS--L--GENQDKKIKDLATSHKEKHKIPIEDLEVIAAAILEVA 112
dnMb11      EELKEHGVTVLKALIEI---F--KNEDEEKLKELAKSHKEEHKIPIEDLEKIAEAIIEVL 113
dnMb12      EELKEHGVTVLKALRAI---L--EKGDEELLKKLAESHTKEHKIPVSDLEVIAESIIEVA 114
dnMb08      PLLAEHGVTVLKALIKIVEEL--KKGDTSLIKELAKSHKTEHKIDIKDLKYIAESIIEVL 117
dnMb09      PLLREHGVTVLKALTKIAEEL--KKGKTGTLKKLAESHSKVHKIPISDLERIAEAIIEVL 117
                 ** *** **               :. ** **   ***  . *. *: ::: .

3RGK        QSKHPGDFGADAQGAMNKALELFRKDMASNYKEL-    149
dnMb02      KERFPEFFNEKAQAALTKALQQFIDAIAAEYKKL-    149
dnMb03      KKRFPTAFNSAAQAAVTKALQQFIDALEKEFKKL-    149
dnMb04      KEAFPEAFDAKAQAAFTKALEQFIKAFEEEYKKL-    149
dnMb05      KERFPTEFDEEAQAALEKALAQFIAAYAAQAAKL-    149
dnMb20      KERYPERFGEKAQEAVKKFLDLFIEKFEEEAEKEK    147
dnMb16      KKFYPEEFGEEAQAAVQKLLNYFIEKLKQYYE---    148
dnMb19      AERFPDEFGPEARAALTDFLDWFIAEIEEEYKK--    149
dnMb18      KEKYPEEFGEETQEALKEFLDWFIEELEKEFKE--    149
dnMb15      KKRFPEFFGEKAQAAVRKFLDFFISKLKEYYE---    148
dnMb17      KKFFPEKFGEKAQEALKKFLNYFIEELEKEYEKL-    153
dnMb13      KKRYPEEFGEKTQAALKEFLDEFIELTEKYYK---    147
dnMb14      KKRFPKEFDEKTYAALKEFLDYFIEKIEKYYK---    148
dnMb21      KERFPEEFGEEAQAAFTKFLDKFIKDIAELQKKFE    150
dnMb07      AEKYPEEFGPEARAAVTKALELFIKKLAEFYE---    146
dnMb06      EEKFPKEFNEKAREALKKALEYFIEELEKYYK---    146
dnMb10      KERFPEEFDEAAQAALQEFLDDFISKLKEYFE---    144
dnMb11      KEKYPEEFDEEAEEAVKKFLKLFIEKLKEYRE---    145
dnMb12      KKRFPEEFGEEAQAALKKFLEEFIEKWKEYQEEFK    149
dnMb08      KKRFPEEFDEKAKEAVEKVLNLFIEKIEEFYKK--    150
dnMb09      EERFPEEFDEKAKEAVKKFLDLFIEKHAEFVKK--    150
             . .*  *.  :  *. . *  *
```

Myoglobin sequences

dnMb2

```
ATGTCAGGAGGCCTGACCGAAGAAGAACAGAAACTGGTGTGGGAAATTTTTGAACGCTTTGAAGAGGATCTGGAAGGCTTTGGCCT
GGATGTGCTGATTCGCGCGTTTACCGAACATCCAGAAACCCTGAAAAAATTTCCGCGCTTTGCGGATCTGAAAAGCGAAGCGGAAT
TACGTGCGAGCCCGCGCCTGCGCGAACATGGCGTGACCGTGCTGAAAGCGCTGATTAAAATCTTTAAAAAAGGCGAAGATTTTGCC
GAAGAAGTGAAACCGCTGGCGGAAAGCCATAGCAAAGTGCATAAAATTCCGGTGAGCGATCTCGAAGTGATTGCGGCGGCGATTCT
GGCGACCGCGAAAGAACGCTTTCCGGAATTTTTTAATGAAAAAGCGCAGGCGGCGCTGACCAAAGCACTGCAGCAGTTTATTGATG
CGATCGCCGCGGAATATAAAAAACTGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGGLTEEEQKLVWEIFERFEEDLEGFGLDVLIRAFTEHPETLKKFPRFADLKSEAELRASPRLREHGVTVLKALIKIFKKGEDFA
EEVKPLAESHSKVHKIPVSDLEVIAAAILATAKERFPEFFNEKAQAALTKALQQFIDAIAAEYKKLGGGSGSHHWGSTHHHHHH
```

dnMb3

```
ATGTCAGGAGGCCTGACCGCGGAAGAACAGAAACTGGTGCGCGATATTTGGGCGGAAGTGGAAAAAGATCGCGAAGGCTTTGGCCT
GGAAGTGCTGTTGCTGACCTTTACCGAACATCCAGAAACCTTAAAAAAATTTCCACGTTTTGCGCATCTGAAAAGCGCCGAGGAAC
TGCGCGCGAGCCCGGAAGCGAAAGCGCATGGCGTGACCGTGCTGGATGCGCTGAGCAAAATTTTGAAGAAAGGCAGCAATTTCGAA
GAAGAAATTAAACCGCTGGCGGAAAGCCATTATAAAGAACATAAAATTCCGATTGAAGATTTGAAAGTGATTGCGGATGCGATTAT
TGCGGTGCTGAAAAAACGCTTTCCGACCGCGTTTAATAGCGCGGCGCAGGCGGCGGTGACCAAAGCGCTGCAGCAGTTTATTGATG
CACTGGAAAAGGAATTTAAGAAACTGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGGLTAEEQKLVRDIWAEVEKDREGFGLEVLLLTFTEHPETLKKFPRFAHLKSAEELRASPEAKAHGVTVLDALSKILKKGSNFE
EEIKPLAESHYKEHKIPIEDLKVIADAIIAVLKKRFPTAFNSAAQAAVTKALQQFIDALEKEFKKLGGGSGSHHWGSTHHHHHH
```

dnMb4

```
ATGTCAGGAGGTTTAACCGCGGAAGAACAAGCGCTGGTGCGCGCGATTTGGGCGAAAGTGCGCGAAGATCTGGAAGGCTTTGGCCT
GGCGGTGCTGCTGAAAACCTTTACCGAACATCCGGAAACCCTGAAAAAATTTCCGCGCTTTAAAGACTTGAAAAGCGAAGAAGAAA
TTCTGGCGAGCGAAAAGGCGAAAAAACATGGCGTGACCGTGCTGACTGCGCTGTTTGCGATTTTTGATAAAGGTGAGAATTTTGAA
GAGGAAATTAAACCGCTGGCGGAAAGCCATTATAAAGAACATAAAATTCCGATTAGCGATCTGAAAGTGATTGCGGATGCGATTGT
GGCCGTGTTGAAAGAAGCGTTTCCGGAAGCATTTGATGCGAAAGCGCAGGCGGCGTTTACCAAAGCACTGGAACAGTTTATTAAAG
CGTTCGAGGAAGAATATAAAAAACTGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGGLTAEEQALVRAIWAKVREDLEGFGLAVLLKTFTEHPETLKKFPRFKDLKSEEEILASEKAKKHGVTVLTALFAIFDKGENFE
EEIKPLAESHYKEHKIPISDLKVIADAIVAVLKEAFPEAFDAKAQAAFTKALEQFIKAFEEEYKKLGGGSGSHHWGSTHHHHHH
```

dnMb5

```
ATGTCAGGAGGCCTGAGCGATGAAGAACAGGCGCTGGTGCTGAGCATTTTTGAAAAAGTGAAAGAAGATCTGGCGGGCTTTGGCCT
GGATGTGCTGTTGCTGGCGTTTACCAAAAATCCGGCGACCCTGGAAAAATTTCCGCGCTTTGCGGATCTGAAAAGCGAAGCGGAAC
TGTTGGCGAGCGAAAAGGCCAAAGAACATGGCATTACCGTGCTGACCGCGCTGTTTGCGATTTTCGAGAAAGGCGATGATTTTGAT
GCGGAAGTTGAACCGCTGGCGACCAGCCATACCCGCGAACATAAAATTCCGACGAGCGATCTGGAAGTGATTGCGGCGGCGATTCT
GGAAACCGCGAAGGAACGCTTTCCAACCGAATTTGATGAAGAAGCGCAGGCGGCCTTAGAAAAAGCGTTGGCGCAGTTTATTGCAG
CGTATGCGGCGCAAGCCGCGAAACTGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGGLSDEEQALVLSIFEKVKEDLAGFGLDVLLLAFTKNPATLEKFPRFADLKSEAELLASEKAKEHGITVLTALFAIFEKGDDFD
AEVEPLATSHTREHKIPTSDLEVIAAAILETAKERFPTEFDEEAQAALEKALAQFIAAYAAQAAKLGGGSGSHHWGSTHHHHHH
```

dnMb6

```
ATGTCAGGAAAACTGAGCGAAGAAGAAAAAGAAATTGTGCTGAAAATTTTTGAACTGGTGGAAAAGGATGTGGAAGAAATTGGCCT
GCGCGTGCTGGAACTGACCTTTGAAAAACATCCAGAAACCCTGGAGAAATTTCCACGCTTACGCGAATTATTAGCGGCGGGCCGCC
```

TGGAGGAACTGGAAGCGTATCTGCGCGAACATGGCGTGACCGTGTTAAAAGCGCTGATTGAAGCGATTAAAAATGAAGATGAAGAA
CTGTTGGAAAAACTGGCGAAAAGCCATAAAGAGGAACATAAAATTCCGATTGAATATCTGAAATATATTGCGGATAGCATTATTGA
AGTGTTAGAAGAGAAGTTTCCGAAAGAATTTAATGAAAAGGCGCGCGAAGCGTTGAAGAAAGCACTGGAATATTTTATTGAGGAGC
TGGAGAAATATTATAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGKLSEEEKEIVLKIFELVEKDVEEIGLRVLELTFEKHPETLEKFPRLRELLAAGRLEELEAYLREHGVTVLKALIEAIKNEDEE
LLEKLAKSHKEEHKIPIEYLKYIADSIIEVLEEKFPKEFNEKAREALKKALEYFIEELEKYYKGGGSGSHHWGSTHHHHHH

dnMb7

ATGTCAGGAAAATTAACCCCAGAAGAAAAAGCGATTGTTTTACGTATTTTTGCGTTAGTTCGTGAAGATCGTGCGGGTATTGGTGC
GGCGATTTTGCGTCGTACCTTTGAAGCGCATCCAGAAACCTTAGAAAAATTTCCACGTTTACGTGCGTTACGCGCCGCGGGCCGCG
AAGCGGAACTGGAAGCGCTGTTGCGTGAACATGGCGTGACCGTGCTGGATGCGCTGATTGAAATTGTGGAAAATGATGATGAAGAA
CTGCTGAAAAAACTGGCGGAAAGCCATAAAACCACCCACAAAATTCCAATTGAACATTTAGAACATATTGCGGCGGCGCTGCTGGA
AGTGCTGGCCGAGAAATATCCGGAAGAATTTGGTCCGGAGGCGCGCGCAGCGGTGACCAAAGCCTTGGAACTGTTTATTAAAAAGC
TCGCGGAATTTTATGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGKLTPEEKAIVLRIFALVREDRAGIGAAILRRTFEAHPETLEKFPRLRALRAAGREAELEALLREHGVTVLDALIEIVENDDEE
LLKKLAESHKTTHKIPIEHLEHIAAALLEVLAEKYPEEFGPEARAAVTKALELFIKKLAEFYEGGGSGSHHWGSTHHHHHH

dnMb8

ATGTCAGGAATTAAAATTAGCGAAGAAGAATTTGAAATTGTGCTGGAAATTTTTGAACTGGTGAAAAAAGATCTGGCGGGCATTGG
CAAAGAAGTGCTGATTCTGACCTTTACCAAACATCCAGAAACCCTGAAGAAATTTCCACGTTTTGCGCATCTGAAAACCGTGGAAG
AACTGGAAGCGAGCCCGCTGCTGGCGGAACATGGCGTGACCGTGCTGAAAGCGCTGATTAAGATCGTGGAGGAACTGAAGAAAGGC
GATACCAGCCTGATCAAAGAACTGGCGAAAAGCCATAAAACCGAACATAAGATTGATATTAAGGATTTGAAATATATTGCGGAAAG
CATTATTGAAGTTTTAAAAAAACGCTTTCCGGAAGAGTTCGATGAAAAAGCGAAAGAAGCGGTGGAAAAAGTGTTGAATCTGTTTA
TCGAGAAAATCGAAGAATTTTATAAAAAGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGIKISEEEFEIVLEIFELVKKDLAGIGKEVLILTFTKHPETLKKFPRFAHLKTVEELEASPLLAEHGVTVLKALIKIVEELKKG
DTSLIKELAKSHKTEHKIDIKDLKYIAESIIEVLKKRFPEEFDEKAKEAVEKVLNLFIEKIEEFYKKGGGSGSHHWGSTHHHHHH

dnMb9

ATGTCAGGAAGCAAACTGACCGAAGAAGAATGGAAAACCGTGTTTAAAATTTTTGCGCTGGTGGAAAAAGATCTGGAAGGCTTTGG
CCTGGCGGTGCTGATTCGCACCTTTACCCGTTATCCAGAAACCTTGAAAAAATTTCCACGTTTCGCGCATCTGAAGACCGTGGAAG
AATTGCGTGCGAGCCCGCTGCTGCGCGAACATGGCGTGACCGTGCTGAAAGCGCTGACCAAAATTGCGGAAGAACTGAAGAAAGGC
AAAACCGGCACCCTCAAAAAACTGGCGGAAAGCCATAGCAAAGTGCATAAAATTCCGATTAGCGATTTAGAACGCATTGCCGAAGC
GATTATTGAAGTGCTGGAAGAACGCTTTCCGGAAGAGTTTGATGAAAAAGCGAAAGAAGCGGTGAAGAAGTTTCTGGATCTGTTTA
TCGAAAAACATGCGGAATTTGTGAAAAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGSKLTEEEWKTVFKIFALVEKDLEGFGLAVLIRTFTRYPETLKKFPRFAHLKTVEELRASPLLREHGVTVLKALTKIAEELKKG
KTGTLKKLAESHSKVHKIPISDLERIAEAIIEVLEERFPEEFDEKAKEAVKKFLDLFIEKHAEFVKKGGGSGSHHWGSTHHHHHH

dnMb10

ATGTCAGGAGATGCGGAAAAACAGGCGCTGGTGGCGAGCATTTTTGCGAAATTTGAAGCGGATCTGGAAGGCTTTGGCAAAGCGGT
GCTGATTAAAACCTTTACCAAACATCCGGAAACCCGCAAAAAATTTCCGCGCTTTAAACATCTGAAAAGCGTGGAAGAACTGGAAA
AAAGCGAAGAACTGAAAGAACATGGCGTGACCGTGCTGACCGCGCTGCGCGAGATTAGCCTGGGCGAAAATCAGGATAAAAAGATT
AAAGATCTGGCGACCAGCCATAAAGAAAAGCATAAAATTCCGATTGAAGATTTGGAAGTGATTGCGGCGGCGATTTTAGAAGTGGC
GAAGGAACGCTTTCCGGAAGAATTTGATGAGGCGGCGCAGGCAGCGCTGCAGGAATTTCTGGATGATTTTATTAGCAAATTAAAAG
AATATTTTGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGDAEKQALVASIFAKFEADLEGFGKAVLIKTFTKHPETRKKFPRFKHLKSVEELEKSEELKEHGVTVLTALREISLGENQDKKI
KDLATSHKEKHKIPIEDLEVIAAAILEVAKERFPEEFDEAAQAALQEFLDDFISKLKEYFEGGGSGSHHWGSTHHHHHH

dnMb11

ATGTCAGGAAAACTGAGCGAAGAAGAAAAAGAAATTGTGCTGAAAATTTTTGCGCTGGTGGAAAAGGATCTGGAAGGCTTTGGCAA
AGAAGTGCTGATTAAAACCTTTCTGAAATATCCGGAAACCCTGAAAAAATTTCCGCGCTTTAAACATCTGAAAACCGAGGAAGAAC
TGAAAGCGTCGGAAGAGTTGAAAGAACATGGCGTGACCGTGTTAAAAGCGCTGATTGAAATCTTTAAAAATGAAGATGAAGAAAAA
CTGAAGGAGCTGGCGAAAAGCCATAAAGAAGAGCATAAAATTCCGATTGAAGATTTAGAGAAAATTGCGGAAGCGATTATTGAAGT
ACTGAAAGAGAAATACCCGGAAGAATTTGATGAAGAAGCGGAGGAGGCGGTGAAAAAGTTTTTAAAACTGTTTATCGAGAAGCTCA
AAGAATATCGCGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGKLSEEEKEIVLKIFALVEKDLEGFGKEVLIKTFLKYPETLKKFPRFKHLKTEEELKASEELKEHGVTVLKALIEIFKNEDEEK
LKELAKSHKEEHKIPIEDLEKIAEAIIEVLKEKYPEEFDEEAEEAVKKFLKLFIEKLKEYREGGGSGSHHWGSTHHHHHH

dnMb12

ATGTCAGGACTGAATCTGAGCCCGGAAGATAAAGCGAAAGTGCTGGAAATTTTTGCGCTGGTGGAAGAAGATCTGGAAGGCTTTGG
CCGCGAAGTTCTGATTCTGACCTTTACCAAACATCCGGAAACCCTGAAAAAATTTCCACGCTTTGCGCATCTGAAAACCGAAGAGG
AACTGCGCGCGAGCGAAGAACTGAAAGAACATGGCGTGACCGTGCTGAAAGCGCTGCGTGCGATTCTGGAAAAAGGCGATGAAGAG
CTGTTGAAGAAACTGGCGGAAAGCCATACCAAAGAACATAAAATTCCGGTGAGCGATTTGGAAGTGATTGCGGAAAGCATTATTGA
AGTGGCGAAAAAACGCTTTCCGGAGGAATTTGGTGAAGAAGCGCAGGCGGCGCTGAAGAAGTTTTTAGAAGAATTTATCGAAAAAT
GGAAAGAATATCAGGAGGAGTTTAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGLNLSPEDKAKVLEIFALVEEDLEGFGREVLILTFTKHPETLKKFPRFAHLKTEEELRASEELKEHGVTVLKALRAILEKGDEE
LLKKLAESHTKEHKIPVSDLEVIAESIIEVAKKRFPEEFGEEAQAALKKFLEEFIEKWKEYQEEFKGGGSGSHHWGSTHHHHHH

dnMb13

ATGTCAGGAGATGAACGCAATAAACTGGTGCTGAGCGCGTTTGCGCTGGTGCGCGAAGATCTGGAAGAAATTGGCGCGGAAGTGCT
GATTCTGACCTTTACCGAAAATCCGGAAACCCTGAAAAAATTTCCGCGCTTTGCGCATCTGAAAACCGAAGAAGAACTGAAAAAAA
GCCCGCTGCTGAAAGAACATGGCGTGACCGTGCTGAATGCGCTGATTGAAATTGCGGAATTAAAATATAGCGGCGGCGATTATGAA
AGCCTGGTGAAAGAGCTGGCGAAAAGCCATAAAGAAAAACATAAAATTCCGATTGAAGATTTGAAAGCGATTGCAGAAGCCATCTT
GAAAGTCTTAAAAAAACGCTATCCAGAAGAATTTGGTGAGAAGACCCAGGCGGCGTTGAAGGAGTTTCTGGATGAATTTATTGAAC
TGACCGAGAAATATTATAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGDERNKLVLSAFALVREDLEEIGAEVLILTFTENPETLKKFPRFAHLKTEEELKKSPLLKEHGVTVLNALIEIAELKYSGGDYE
SLVKELAKSHKEKHKIPIEDLKAIAEAILKVLKKRYPEEFGEKTQAALKEFLDEFIELTEKYYKGGGSGSHHWGSTHHHHHH

dnMb14

ATGTCAGGAGAAAAAGAAAAAAATGAACTGGTGCTGAAAGCGTTTGAACTGATTGAAAAGGATCTGGAAGGCTTTGGCAGCGAAGT
GCTGATTCTGACCTTTACCAAACATCCGGAAACCCTGAAAAAATTTCCGCGCTTTAAACATCTGAAAACCGAAGAAGAATTTAAAG
CGAGCGAAGAACTGAAAGAACATGGCGTGACCGTGTTGAAGGCGCTGATCGAAATTGCGAAATTAAAAGTGAGCGGCGAAGATTAT
GATAGCCTGATTAAAGAGCTGGCGAAAAGCCATAAAACCAAACATAAAATTCCGATTGAATATTTGAAATATATTGCGGATGCGAT
TTTGGAAGTGGCAAAAAAACGCTTTCCGAAAGAGTTTGATGAGAAGACCTATGCGGCGTTAAAGGAATTTCTGGATTATTTTATTG
AGAAAATTGAGAAATATTATAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGEKEKNELVLKAFELIEKDLEGFGSEVLILTFTKHPETLKKFPRFKHLKTEEEFKASEELKEHGVTVLKALIEIAKLKVSGEDY
DSLIKELAKSHKTKHKIPIEYLKYIADAILEVAKKRFPKEFDEKTYAALKEFLDYFIEKIEKYYKGGGSGSHHWGSTHHHHHH

dnMb15

ATGTCAGGAGAAGAAGAAGAAAAACAGATTGTGCTGGAACTGTTTGCGAAAGTGGAAGAGGATCTGGAAGGCATTGGCCTGGAAGT
GCTGATTCTGACCTTTACCAAACATCCAGAAACCCGTAAAAAATTTCCACGTTTTGCGCATCTGACCACCGAAGCGCAGCTGCAGG
CGAGCCCGGAACTGAAACAGCATGGCGTGACCGTGCTGAAAGCGCTGATTACCATTGCCAAACTGTATTATGAAGGCAAAGATTAC

GAAAGCCTGATTAAAGAACTGGCGAAAAGCCATAAAGAGGAACATAAAATTCCGATTGAATATTTGGAATATATTAGCGAAAGCAT
TTTGGAGGTTCTGAAAAAACGCTTTCCGGAATTTTTTGGTGAGAAAGCCCAGGCGGCGGTGCGCAAATTTCTGGATTTTTTTATTA
GCAAATTGAAAGAATATTACGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGEEEEKQIVLELFAKVEEDLEGIGLEVLILTFTKHPETRKKFPRFAHLTTEAQLQASPELKQHGVTVLKALITIAKLYYEGKDY
ESLIKELAKSHKEEHKIPIEYLEYISESILEVLKKRFPEFFGEKAQAAVRKFLDFFISKLKEYYEGGGSGSHHWGSTHHHHHH

dnMb16

ATGTCAGGAGCGGAAGAAGAAAAAGAAAAAGTGCTGAGCATTTTTAAACTGGTGGAAAAGGATAAAAAAACCATTGGCAGCGAAGT
CCTGATTATTACCTTTACCAAAAATCCGGAAACCAAGAAGAAATTTCCGCGCTTTAAAGATCTGAAAACCGTGGAAGAACTGAAAG
CGAGCGAAAAAGTGAAAGATCATGGCGTGACCGTGCTGGATGCGCTGATTGAATGGGCGCGCCTGCATGTGGAAGGCAAAGATTAT
GATAGCCTGGTGAAAAAACTGGCGGAAAGCCATAAGAAGGAACATAAAATTCCGATTGAAGATTTGAAAAGCATTGCGGACGCCTT
GATCGAAGTTTTAAAGAAATTTTATCCAGAGGAATTTGGCGAAGAAGCGCAGGCGGCGGTGCAGAAACTGCTGAATTATTTTATTG
AGAAGTTGAAACAGTATTATGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGAEEEKEKVLSIFKLVEKDKKTIGSEVLIITFTKNPETKKKFPRFKDLKTVEELKASEKVKDHGVTVLDALIEWARLHVEGKDY
DSLVKKLAESHKKEHKIPIEDLKSIADALIEVLKKFYPEEFGEEAQAAVQKLLNYFIEKLKQYYEGGGSGSHHWGSTHHHHHH

dnMb17

ATGTCAGGAATTAAACTGAGCGAAGAAGAAAAAAAAACTGGTGCTGGAAATTTTTAAGCTGTTTGAAGAGAATCTGGAAGAATTTGG
CAAAGAAGTGCTGATTACCACCTTTACCAAACATCCGGAAACCAAAAAAAAAATTTCCGCGCTTTGCGCATCTGAAAACCGAGGAGG
AATTTCTGGCGAGCCCGGAACTGGCGAAACATGGCGTGACCGTGCTGAATGCGCTGATTGAAATTGCGAAACTGTATTTAGAAGGC
AAGGATTATCGCAGCCTGATTAAAAAGCTGGCAAAAAGCCATAAACTGGAACATAAAATTCCGATTGAAGATTTGAAATATATTGC
GGATGCGATTATTGAAGTGGCCAAAAAGTTCTTTCCAGAAAAATTCGGTGAAAAAGCGCAGGAAGCGCTGAAGAAGTTCCTGAATT
ATTTTATTGAGGAGTTGGAAAAAGAATATGAAAAACTGGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCAC
CACCAC

MSGIKLSEEEKKLVLEIFKLFEENLEEFGKEVLITTFTKHPETKKKFPRFAHLKTEEEFLASPELAKHGVTVLNALIEIAKLYLEG
KDYRSLIKKLAKSHKLEHKIPIEDLKYIADAIIEVAKKFFPEKFGEKAQEALKKFLNYFIEELEKEYEKLGGGSGSHHWGSTHHHH
HH

dnMb18

ATGTCAGGAGATGAAGAAAAGAAAAAACTGGTGCTGGAAGCGTTTGAATTGGTGGAAAAAGATATTGAAGGCATTGGCGCGGAAGT
GCTGAAACTGACCTTTGAAAAACATCCGGAAACCCTGGAGAAATTTCCGCGCCTGAAAGAATTACATGCGGCGGGCAGCCCGGAAC
TGGAAGAACTGTTGAAGGAACATGGCGCGACCGTGTTGAAAGCGCTGATTGAAATTGCGCGCTTAAAAATTAGCGGCGGCGATTAT
CTGAGCCTGGTGAAAGAGCTGGCGAAAAGCCATAAAGAAGAGCATAAAATTCCGATTGAAGATCTGAAGAAAATCGCCGAAGCCCT
GCTGGAGGTTTTGAAGGAAAAATATCCAGAAGAATTTGGCGAAGAAACCCAGGAGGCTCTGAAAGAATTTCTGGATTGGTTTATCG
AGGAGCTGGAAAAGGAATTTAAGGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGDEEKKKLVLEAFELVEKDIEGIGAEVLKLTFEKHPETLEKFPRLKELHAAGSPELEELLKEHGATVLKALIEIARLKISGGDY
LSLVKELAKSHKEEHKIPIEDLKKIAEALLEVLKEKYPEEFGEETQEALKEFLDWFIEELEKEFKEGGGSGSHHWGSTHHHHHH

dnMb19

ATGTCAGGAAGCGAAGAAAAAGCGGCGTTGGTGCTGGCGCTGTTTGATCGCGTGGAAGCGGATCGCGAAGAAATTGGTGCGGCGGT
GCTGCGCCGCACCTTTGAAGAACATCCGGAAACCCTGAAAAAATTTCCGCGCTTTCTGGAACTGTATAAAAAAGGTAGCCCAGAAC
TGGATGCGCTGCTGAAAGAGCATGGCAAAACCGTGCTGGACGCGCTGATTGAAATTGCGCGCCTGCGCTATAGCGGCGAAGATTAT
CGCAGCCTGATTAAAGAGCTGGCGAAAAGCCATAAAGAAGAACATAAAATTCCAATTGAAGATCTGCGCCATATTGCGGAAGCGTT
GTTGGCCGTTTTGGCGGAACGCTTTCCGGATGAATTTGGTCCAGAAGCGCGCGCGGCACTGACCGATTTTTTGGATTGGTTTATCG
CGGAAATTGAGGAGGAATATAAGAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGSEEKAALVLALFDRVEADREEIGAAVLRRTFEEHPETLKKFPRFLELYKKGSPELDALLKEHGKTVLDALIEIARLRYSGEDY
RSLIKELAKSHKEEHKIPIEDLRHIAEALLAVLAERFPDEFGPEARAALTDFLDWFIAEIEEEYKKGGGSGSHHWGSTHHHHHH

dnMb20

ATGTCAGGACTGAGCGAAGAAGAATGGAAAATTGTGCTGGAAATTTTTGCGTTAGTTCGTGAAGATTTAGCGGGTGTTGGTGCGGC
GGTTTTAGAACGTACCTTTGCGACCCATCCAGAAACCTTAAAAAAATTTCCACGTTTTCTCGCGGCAGCGGAAGCGGGCGTGTTGG
ATCGTGCGCTGCTGGCCGCGCATGGCGAAACCGTGCTGACCGCGCTGATTGAAATTGCGGAAAGCAAACTGGATCCGGAACTGATT
AAGAAACTGGCGGAAAGCCATGTGAAAGAACATAAAATTCCGATTGAATATCTGCGCGCGATTGCCGATAGCCTGATCGCGGTCCT
GAAAGAACGCTATCCAGAGCGCTTTGGTGAAAAAGCGCAGGAGGCGGTTAAAAAGTTTCTGGATCTGTTTATTGAAAAGTTTGAAG
AAGAAGCGGAAAAAGAAAAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGLSEEEWKIVLEIFALVREDLAGVGAAVLERTFATHPETLKKFPRFLAAAEAGVLDRALLAAHGETVLTALIEIAESKLDPELI
KKLAESHVKEHKIPIEYLRAIADSLIAVLKERYPERFGEKAQEAVKKFLDLFIEKFEEEAEKEKGGGSGSHHWGSTHHHHHH

dnMb21

ATGTCAGGAAGCTTAACCCCAGAAGAATTAGCGATTGTGAAAGCGCTGTTTGCGCGCGTGCGCGAAGATCTGGAAGGCGTGGGCGC
GGAAGTGCTGCGCTTGACCTTTGAAAAACATCCAGAAACCTTAAAAAAATTTCCACGTTTTTTGGAATTGAAGAAAGCGGGCAGCC
CGGAACTGGAAGCGGAGCTGCGTGCGCATGGCGTGACCGTGCTGACCGCGCTGATTGAACTTGCGGATAATTATGAAGGCAATAAT
GAAACTCTGGAAAAACTGGCCGAAAGCCATACCAAAGTGCATAAAATTCCGGTGAGCGATCTGAAGAATATTGCGGCGGCGATTAT
TGAAGTCCTGAAAGAACGCTTTCCGGAAGAGTTTGGCGAAGAAGCGCAGGCGGCGTTTACCAAATTTTTAGATAAATTTATTAAAG
ATATTGCGGAGCTGCAGAAAAAATTTGAAGGCGGCGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGSLTPEELAIVKALFARVREDLEGVGAEVLRLTFEKHPETLKKFPRFLELKKAGSPELEAELRAHGVTVLTALIELADNYEGNN
ETLEKLAESHTKVHKIPVSDLKNIAAAIIEVLKERFPEEFGEEAQAAFTKFLDKFIKDIAELQKKFEGGGSGSHHWGSTHHHHHH

Native Mb, 3RGK

ATGTCAGGAGGCCTGAGCGATGGCGAATGGCAGCTGGTGCTGAACGTGTGGGGCAAAGTGGAAGCGGATATTCCGGGCCATGGCCA
GGAAGTGCTGATTCGCCTGTTTAAAGGTCATCCGGAAACCCTGGAAAAGTTCGACCGCTTTAAACATCTGAAATCTGAAGATGAGA
TGAAAGCGAGCGAAGATCTGAAGAAACATGGCGCGACCGTGCTGACCGCGCTGGGCGGTATTCTGAAGAAGAAAGGCCATCACGAA
GCCGAAATTAAACCGCTGGCGCAGAGCCATGCGACCAAACATAAAATTCCGGTGAAGTACCTGGAATTTATCAGCGAAGCGATTAT
TCAGGTGCTGCAGAGCAAACATCCGGGCGATTTTGGCGCGGATGCGCAGGGTGCGATGAACAAAGCGCTGGAACTGTTTCGCAAAG
ATATGGCGAGCAACTATAAAGAACTGGGTTCCGGCAGCCATCATTGGGGCAGCACCCACCACCACCACCACCAC

MSGGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHE
AEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGSGSHHWGSTHHHHHH

TEV sequences

N-terminal tag + linker and C-terminal linker are highlighted in green.

hyperTEV56

ATGAGCCACCACCACCACCACCACCACTCAGGAATGGAAAGCGCGGCGCCGGGCCCGCGCGATTATAACCCGATCAGCGATACCATTGT
GAAACTGACCAACACCTCTGATGGCTATAGCATTAGCCTGTATGGCATTGGCTTTGGCCCGCTGATCATTACCAACGCGCATTTAT
TTCGCCGCAACAACGGCACCCTGACCGTGACCAGCAAACATGGCACCTTCACCATTGAAAACACCACCACCCTGCAGCTGCATCTG
ATTGAAGGCCGCGATCTGGTGATTATTAAAATGCCGAAAGATTTTCCGCCGTTTCCGACCGATCTGGTGTTTCGCGAACCGGTGGA
AGGCGAGAAAATTACCCTGGTGACCCGCAACTTTCAGACCAAAGAACCGACCAGCGAAGTGAGCGATGTGAGCAGCACCTATCCGA
GCAGCGATGGCGTGTTTTGGAAACATTGGATTCCGACCAAAGATGGTCAGTGCGGCAGCCCGATGGTGAGCGTGGAAGATGGCAGC
ATTGTGGGCATTCATAGCGCGAGCAACTTTACCAATACCAACAACTATTTTACCGCGGTGCCGCCGGATTTCATGGATCTGCTGAC
CAACGATAGCCTGCAGAAATGGATTAGCGGCTGGAGCCTGAACAGCGATAGCGTTGAATGGGGCGGCCATAAAGTGTTTATGGATA
AACCGGGTTCC

MSHHHHHHSGMESAAPGPRDYNPISDTIVKLTNTSDGYSISLYGIGFGPLIITNAHLFRRNNGTLTVTSKHGTFTIENTTTLQLHL
IEGRDLVIIKMPKDFPPFPTDLVFREPVEGEKITLVTRNFQTKEPTSEVSDVSSTYPSSDGVFWKHWIPTKDGQCGSPMVSVEDGS
IVGIHSASNFTNTNNYFTAVPPDFMDLLTNDSLQKWISGWSLNSDSVEWGGHKVFMDKPGS

hyperTEV60

ATGAGCCACCACCACCACCACCACTCAGGAGCGGAAAGCGCGGCGCCGGGCCCGCGCGATTATAACCCGATTAGCGATACCATTGT
TCTGCTGACCAATACCAGCGATGGCTATAGCATTAGCCTGTATGGCATTGGCTTTGGCCCGCTGATTATTACCAACGCGCACCTGT
TTCGCCGCAACAACGGCACCCTGACCATTACCAGCAAACATGGCACCTTTACCATTAGCAACACCACCACCCTGAAACTGCATCTG
ATCGAAGGCCGCGATCTGGTGCTGATTGAAATGCCGAAAGATTTTCCGCCGTTTCCGACCAACCTGGTGTTTCGTGAACCGGTGGT
GGGCGAAGAAATTGTGCTGGTGACCCGCAACTTTCAGACCAAAACCCCGACCAGCGAAGTGAGCGATGTGAGCACCACCTATCCGA
GCTCCGATGGCGTGTTTTGGAAACATTGGATTCCGACGAAAGATGGCCAGTGCGGCAGCCCGATGGTGAGCGTGACCGATGGCAGC
ATTGTGGGCATTCATAGCGCGAGCAACTTTACCAACACCAACAACTATTTTACCGCGGTGCCGCCGGATTTTATGCGCCTGCTGAC
CGATCCGAGCCTGCAGAAATGGGTGAGCGGCTGGAGCCTGAACAGCGATAGCGTGGAATGGGGCGGCCATAAAGTGTTTATGGATA
AACCGGGTTCC

MSHHHHHHSGAESAAPGPRDYNPISDTIVLLTNTSDGYSISLYGIGFGPLIITNAHLFRRNNGTLTITSKHGTFTISNTTTLKLHL
IEGRDLVLIEMPKDFPPFPTNLVFREPVVGEEIVLVTRNFQTKTPTSEVSDVSTTYPSSDGVFWKHWIPTKDGQCGSPMVSVTDGS
IVGIHSASNFTNTNNYFTAVPPDFMRLLTDPSLQKWVSGWSLNSDSVEWGGHKVFMDKPGS

hyperTEV89

ATGAGCCACCACCACCACCACCACTCAGGAGCGGAAAGCGCGGCGCCGGGCCCGCGCGATTATAACCCGATTAGCAGCACCATTGT
GCGCCTGACCAACACCAGCGACGGCCATAGCATTAGCCTGTTTGGCATTGGCTTTGGCCCGCTGATTATTACCAACGCGCATTTAT
TTCGCCGCAACAACGGCACCCTGACCATTACCAGCCTGCATGGCACCTTTACCATTAGCAACACCACCACCCTGAAACTGCATCTG
ATTGAAGGCCGCGATCTGGTGATTATCAAAATGCCGAAAGATTTTCCGCCGTTTCCGACCACCCTGGAATTTCGCGAACCGGTGGT
GGGCGAAGATATTGTGCTGGTGACCCGCAACTTTCAGGATAAAGATCCGACCAGCGAAGTGAGCGATACCAGCACCACCGAACCGA
GCAGTGATGGCGTGTTTTGGAAACATTGGATCCCGACCAAAGATGGCCAGTGCGGCAGCCCGATGGTGAGCGTTAGCGATGGCAGC
ATTGTGGGCATTCATAGCGCGAGCAACTTCACCAATACCAACAACTATTTTACCGCGGTGCCGCCGAACTTTATGGATCTGCTGAC
CGATCCGAGCCTGCAGAAATGGATTAGCGGCTGGAGCCTGAACGCGGATAGCGTGGATTGGGGCGGCCATAAAGTGTTCATGGATA
AACCGGGTTCC

MSHHHHHHSGAESAAPGPRDYNPISSTIVRLTNTSDGHSISLFGIGFGPLIITNAHLFRRNNGTLTITSLHGTFTISNTTTLKLHL
IEGRDLVIIKMPKDFPPFPTTLEFREPVVGEDIVLVTRNFQDKDPTSEVSDTSTTEPSSDGVFWKHWIPTKDGQCGSPMVSVSDGS
IVGIHSASNFTNTNNYFTAVPPNFMDLLTDPSLQKWISGWSLNADSVDWGGHKVFMDKPGS

TEVd (PDB: 1LVM)(*55*)

ATGAGCCACCACCACCACCACCACTCAGGAGGCGAAAGCCTGTTCAAAGGCCCGCGCGATTATAACCCGATTAGCAGCACCATTTG
CCATCTGACCAACGAAAGCGATGGCCATACCACCAGCCTGTATGGCATTGGCTTTGGCCCGTTTATCATTACCAACAAACATTTAT
TTCGCCGCAACAACGGCACCCTGCTGGTGCAGAGCCTGCATGGCGTGTTTAAAGTGAAGAATACCACGACCCTGCAGCAGCATCTG
ATTGATGGCCGCGATATGATTATTATTCGCATGCCGAAAGATTTTCCGCCGTTTCCGCAGAAACTGAAATTTCGCGAACCGCAGCG
TGAAGAACGCATTTGTCTGGTGACCACCAACTTTCAGACGAAAAGCATGAGCAGCATGGTGAGCGATACCAGCTGCACCTTTCCGA
GCAGCGATGGTATCTTTTGGAAACATTGGATTCAGACCAAAGATGGTCAGTGCGGCAGCCCGCTGGTGAGCACCCGTGATGGCTTT
ATTGTGGGCATTCATAGCGCGAGCAACTTTACCAATACCAATAACTATTTTACCAGCGTGCCGAAGAACTTTATGGAACTGCTGAC
CAACCAGGAAGCGCAGCAGTGGGTGAGCGGCTGGCGCCTGAACGCGGATAGCGTGCTGTGGGGCGGCCATAAAGTGTTTATGGATA
AACCGGGTTCC

MSHHHHHHSGGESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGPFIITNKHLFRRNNGTLLVQSLHGVFKVKNTTTLQQHL
IDGRDMIIIRMPKDFPPFPQKLKFREPQREERICLVTTNFQTKSMSSMVSDTSCTFPSSDGIFWKHWIQTKDGQCGSPLVSTRDGF
IVGIHSASNFTNTNNYFTSVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHKVFMDKPGS

S219V(*55*)

ATGAGCCACCACCACCACCACCACTCAGGAGGCGAAAGCCTGTTCAAAGGCCCGCGCGATTATAACCCGATTAGCAGCACCATTTG
CCATCTGACCAATGAAAGCGATGGCCATACCACCAGCCTGTATGGCATTGGCTTTGGCCCGTTTATTATTACCAACAAACATTTAT
TTCGCCGCAACAACGGCACCCTGCTGGTGCAGAGCCTGCATGGCGTGTTTAAAGTTAAGAACACCACCACCCTGCAGCAGCATCTG
ATTGATGGCCGCGATATGATCATTATTCGCATGCCGAAAGATTTTCCGCCGTTTCCGCAGAAACTGAAATTTCGCGAACCGCAGCG

CGAAGAACGCATTTGCCTGGTGACCACCAACTTTCAGACCAAAAGCATGAGCAGCATGGTGAGCGATACCAGCTGCACCTTTCCGA
GCAGCGATGGTATCTTTTGGAAACATTGGATTCAGACGAAAGATGGCCAGTGCGGCAGCCCGCTGGTGAGCACCCGTGATGGCTTT
ATTGTGGGCATTCATAGCGCGAGCAACTTTACCAACACCAACAACTATTTTACCAGCGTGCCGAAGAATTTTATGGAACTGCTGAC
CAACCAGGAAGCGCAGCAGTGGGTGAGCGGCTGGCGCCTGAACGCGGATAGCGTGCTGTGGGGCGGCCATAAAGTGTTTATGGTGA
AACCGGAAGAACCGTTTCAGCCGGTGAAAGAAGCGACCCAGCTGATGAACGAAGGTTCC

MSHHHHHHSGGESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGPFIITNKHLFRRNNGTLLVQSLHGVFKVKNTTTLQQHL
IDGRDMIIIRMPKDFPPFPQKLKFREPQREERICLVTTNFQTKSMSSMVSDTSCTFPSSDGIFWKHWIQTKDGQCGSPLVSTRDGF
IVGIHSASNFTNTNNYFTSVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHKVFMVKPEEPFQPVKEATQLMNEGS

TEV1Δ(*59*)

ATGAGCCACCACCACCACCACCACTCAGGAGGCGAAAGCCTGTTTAAAGGCCCGCGCGATTATAACCCGATTAGCAGCACCATTTG
CCATCTGACCAACGAAAGCGATGGCCATACCACCAGCCTGTATGGCATTGGCTTTGGCCCGTTTATTATTACCAATAAACATTTAT
TTCGCCGCAACAACGGCACCCTGCTGGTGCAGAGCCTGCATGGCGTGTTTAAAGTGAAGAATACCACGACCCTGCAGCAGCATCTG
ATTGATGGCCGCGATATGATTATTATTCGCATGCCGAAAGATTTTCCGCCGTTTCCGCAGAAACTGAAATTTCGCGAACCGCAGCG
CGAAGAACGCATCTGCCTGGTGACCACCAACTTTCAGACCAAAAGCATGAGCAGCATGGTGAGCGATACCAGCTGCACCTTTCCGA
GCAGCGACGGCATTTTCTGGAAACATTGGATTCAGACGAAAGATGGTCAGTGCGGCAACCCGCTGGTGAGCACCCGCGATGGCTTT
ATTGTGGGCATTCATAGCGCGAGCAACTTTACCAACACCAACAACTATTTTACCAGCGTGCCGAAGAACTTTATGGAACTGCTGAC
CAACCAGGAAGCGCAGCAGTGGGTGAGCGGCTGGCGCCTGAACGCGGATAGCGTGCTGTGGGGCGGCCATAAAGTGTTTATGGTGG
GTTCC

MSHHHHHHSGGESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGPFIITNKHLFRRNNGTLLVQSLHGVFKVKNTTTLQQHL
IDGRDMIIIRMPKDFPPFPQKLKFREPQREERICLVTTNFQTKSMSSMVSDTSCTFPSSDGIFWKHWIQTKDGQCGNPLVSTRDGF
IVGIHSASNFTNTNNYFTSVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHKVFMVGS

superTEV(*8*)

ATGAGCCACCACCACCACCACCACTCAGGACCGCGCGATTATAACCCGATTAGCAGCACCATTGTGCATCTGACCAACGAAAGCGA
TGGCCATACCACCAGCCTGTATGGCATTGGCTTTGGCCCGTTTATTATTACCAACAAACATTTATTTCGCCGCAACAACGGCACCC
TGCTGGTGCAGAGCCTGCATGGCGTGTTTAAAGTTAAGAACACCACCACCCTGCAGCAGCATCTGATTGATGGCCGCGATATGATT
ATTATCCGCATGCCGAAAGATTTTCCGCCGTTTCCGCAGAAACTGAAATTTCGCGAACCGCAGCGTGAAGAACGTATTGTGCTGGT
GACCACCAACTTTCAGACCAAAAGCATGAGCAGCATGGTGAGCGATACCAGCAGCACCTTTCCGAGCAGCGATGGTATTTTCTGGA
AACATTGGATCCAGACCAAAGATGGCCAGTGCGGCAGCCCGCTGGTGAGCACCCGTGATGGCTTTATTGTGGGCATTCATAGCGCG
AGCAACTTTACCAACACCAATAACTATTTTACCAGCGTGCCGAAGAACTTTATGGAACTGCTGACCAATCAGGAAGCGCAGCAGTG
GGTGAGCGGCTGGCGCCTGAACGCGGATAGCGTGCTGTGGGGCGGCCATAAAGTGTTTATGGATAAACCGGGTTCC

MSHHHHHHSGPRDYNPISSTIVHLTNESDGHTTSLYGIGFGPFIITNKHLFRRNNGTLLVQSLHGVFKVKNTTTLQQHLIDGRDMI
IIRMPKDFPPFPQKLKFREPQREERIVLVTTNFQTKSMSSMVSDTSSTFPSSDGIFWKHWIQTKDGQCGSPLVSTRDGFIVGIHSA
SNFTNTNNYFTSVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHKVFMDKPGS

MBP-TEVcs-FKBP-EGFP substrate
TEVcs is highlighted in orange, FKBP is highlighted in green, and EGFP is highlighted in yellow.

MKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKAF
QDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKY
ENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKP
FVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAV
RTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNNNNLGIEGRISTSGSGGGGGSMSENLYFQGSMGVQVETISPGDGRTFPKRGQ
TCVVHYTGMLEDGKKFDSSRDRNKPFKFMLGKQEVIRGWEEGVAQMSVGQRAKLTISPDYAYGATGHPGIIPPHATLVFDVELLKL
NEGGSGGSGGSGGSMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCF
SRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQ
KNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKSGRHH
HHHH

Chapter 3 Supplement

Supplementary Text

Motif generation via conformational sampling. To generate constellations of histidine rotamers around the serine backbone motif, distances, angles, and torsions were first specified in enzyme constraint file format along with sampling ranges and frequencies for the serine-substrate and serine-histidine interactions. These were input to a script which (1) sampled probable histidine rotamers from the Dunbrack library at the specified geometries with respect to the serine nucleophile and substrate, (2) built out protein stubs in helical or strand conformation flanking the histidine using Rosetta, (3) and filtered for clashes between the substrate, serine stub, and histidine stub before outputting these conformations as PDB files. The code and a detailed description of this script can be found here: https://github.com/ikalvet/invrotzyme .

Evaluation of hydrogen bonds in PLACER predictions. Hydrogen-bonding interactions in PLACER predictions were evaluated by the measuring of distances, angles, and torsions between the acceptor and donor heavy atoms. Upper and lower bound cutoffs for each parameter (distance, angle, dihedral) were used to determine if the interaction constituted a hydrogen bond, depending on the hybridization of the participating heavy atoms.

Active site composition and geometric features. To identify potential strategies for improving k_{cat} in our designed hydrolases, we performed a comparative analysis of the active sites of our designs and natural hydrolases, identifying a number of deviating features in the catalytic triad, oxyanion hole, and surrounding molecular environment. Differences in active site makeup could contribute to decreased k_{cat} values in our designs. In nature, the catalytic aspartate or glutamate is supported by a network of additional hydrogen bonds from surrounding sidechains and backbone amides (61), a feature rarely seen in our designs, which have at most one sidechain hydrogen bond to the catalytic aspartate. This highly

polar environment helps aspartate remain in a charged state to stabilize the catalytic histidine, and clearly plays an important role in our designs as well, as evidenced by the significant effect of knockout mutants of these second shell contacts on activity (fig. S12). Although some notable exceptions exist, such as the serine protease subtilisin, in most natural serine hydrolases, at least 7 two backbone amide groups comprise the oxyanion hole, while all but one of the active designs herein contain one backbone amide and one sidechain hydrogen bond. Sidechains are expected to be significantly more challenging to preorganize than backbone amides due to additional rotatable bonds, and crystal structures of our designs often show deviations from the designed oxyanion hole sidechain conformations. Indeed, the design with the highest k_{cat}, momi120, has two backbone amide hydrogen bonds. Geometric deviations within the catalytic triad and oxyanion hole could also account for the differences in k_{cat} between our designs and native serine hydrolases. In nature, the catalytic aspartate and histidine form a particularly ideal hydrogen bond, characterized by short distances and near linear bond angles (33), while crystal structures of our designs display longer distances and deviations from ideal angles. Alignment of the active sites of natural enzymes and our active designs also reveals a difference in the angle of approach of the histidine to the serine in our designs (fig. S16). This shift may affect the role of the histidine in shuffling protons between the serine, substrate and hydrolytic water. Finally, the oxyanion hole hydrogen bonds of natural hydrolases are positioned to make out-of-plane interactions with the ester carbonyl of the substrate, which has been hypothesized to lead to selective stabilization of the oxyanion containing transition states and intermediates over the ground state (34, 51, 52). Our designs exhibit hydrogen bonds that are closer to in-plane hydrogen bonds that may over-stabilize the sp2 ground state of the substrate carbonyl.

Design of catalytic triad containing proteins in NTF2 scaffolds. Theozymes containing catalytic triads and one or two sidechain oxyanion hole residues were docked into a set of hallucinated NTF2 scaffolds (40) using RosettaMatch, designed using three cycles of RosettaDesign with enzyme constraints applied. Designs were input to AF2 for single-sequence structure prediction as described in the Methods

("Filtering"), and filtered for a global Cα RMSD < 0.7 Å, pLDDT > 93, and catalytic residue Cα RMSD < 1.0 Å.

Structure generation with a new variant of all-atom RFdiffusion. In this study, we follow the familiar protocol of generating protein backbone structures with a diffusion model, followed by a fixed backbone sequence design step. The diffusion model we use is a newly trained variant of RFdiffusionAA, which we call 'CA RFdiffusion' (as in, alpha-carbon). In this section, we detail the training and inference protocols for this model. All training and inference code will be released upon manuscript publication.

Introduction

CA RFdiffusion comprises a set of two models, a diffusion model and a "refinement" model, that are both fine-tuned versions of the same pretrained RosettaFold All-Atom structure prediction network. There are two main differences between CA RFdiffusion and any previous variant of RFdiffusion(AA). First, the diffusion model itself produces only alpha-carbon coordinate positions, instead of full N-CA-C frames with both a translation and orientation (as in RFdiffusion (36) and RFdiffusionAA (41)). This means that a second step after generating the "CA trace" must be taken to obtain a fully constrained polypeptide backbone which can be assigned a sequence (one could also design sequences from CA coordinates alone, but we did not attempt this). We perform a second "refinement" step by training a second separate model to produce a full protein backbone given just the trace of alpha carbons.

The second main difference is that instead of taking in motif structure information via static 3D coordinates supplied to the network (as is the case in RFdiffusion and RFdiffusionAA), CA RFdiffusion takes in motif structure information via the available inter-residue pairwise distance and orientation input. In RosettaFold All-Atom (and other previous structure prediction networks) this input is normally used for supplying structural information of proteins homologous to the query sequence during structure prediction (see (41, 75, 76)). For CA RFdiffusion, we use this input to encode motif structures as their pairwise distances and orientations, and have the network reconstruct the motif in its output 3D

predictions. It is noteworthy that at least one other group (77) has developed an analogous motif-scaffolding technique concurrently and independently of the method we describe here. One main benefit of this motif input style is that, as Lin et al. (77) also point out, a user is able to present multiple discontiguous motifs to the network for scaffolding without specifying their relative rigid body transform (as is mandatory, by definition, for methods that take in 3D coordinates of the motif).

Training CA RFdiffusion. Here we detail the two-part training of CA RFdiffusion, which comprises training a diffusion model for generating CA traces, followed by training a "refinement" model which produces full protein backbones from the CA traces.

Architecture. Training the diffusion model for CA RFdiffusion is similar to (36), but there are several differences in the inputs to the network and loss function. We begin with the architecture and pretrained weights of RoseTTAFold All-Atom (41). To this architecture, we expand the dimensions of the template inputs (see Krishna et al. 2024 for architecture details) to include one additional per-token and per-token-pair indicator feature. These denote whether or not a token, or pair of tokens, is part of a motif being scaffolded by the network. During training, we supply three distinct templates (analogous to three separate homologous structure inputs during structure prediction), as shown in Table S3.

We initialize the weights associated with the expanded feature dimensions as all zeros, such that predictions with the expanded architecture completely ignore the new features at first. However, as training progresses, the network can learn to correlate the motif indicator features with the perfect motif information supplied in template #3, helping distinguish that particular structural input from the rest of the structural inputs.

Preparing training examples. Training examples are prepared in two stages: First, a dataset is selected to draw a structure from; either (A) PDB protein monomers without small molecules, or (B) PDB protein monomers in complex with small molecules. Then, a masking strategy is chosen for the drawn structure. The masking strategies available to either dataset are not the same, and they are outlined in Table S4

below. Once the dataset and masking strategy are chosen, any structural components which are considered motif (any ligand, and unmasked protein fragments) are encoded in the motif (third) template with appropriate indicator features.

To noise a selected example, we uniformly sample from the discrete set of times t ∈ [1,200] and then adopt precisely the CA-coordinate noising strategy employed by (36) to noise the CA positions of each N-CA-C frame in the backbone and all atoms in any ligand. We do not use any noising strategy for the frame orientations, and instead set all residue frame orientations to the identity rotation, which removes any information about the native structure contained in the frame orientations. It is noteworthy that this step breaks the equivariance of a network forward pass, because the process of setting all frame orientations to an arbitrary orientation is not equivariant to arbitrary rotations of the input molecules. A summary of training example generation hyperparameters can be found in Table S5. It is worth emphasizing that because the motif information is encoded in the template inputs to the network (described above), all residues and small molecule atoms are noised in 3D space, and the network is tasked with reconstructing the motif in its output – a well-defined problem since the motif is templated.

Loss function. The loss function to train the diffusion model for CA RFdiffusion is:

$$L_{diffusion} = W_{disp} L_{disp} + W_{disto,anglo} L_{dist,ang} + W_{FAPE,prot} L_{FAPE,prot} + W_{FAPE,prot\text{-}sm} L_{FAPE,prot\text{-}sm} + W_{FAPE,sm} L_{FAPE,sm}$$

Where L_{disp} is the CA coordinate displacement loss (as described in (36)), $L_{dist,ang}$ is the pairwise inter-residue distogram and anglogram cross entropy loss (as described in (36)), $L_{FAPE,prot}$ is intra-protein frame-aligned point error (FAPE) (50) loss on the motif – applied only to pairs of residues which were supposed to be constrained with respect to each other in the motif template definition using Algorithms 1, 2, and 3. $L_{FAPE,prot\text{-}sm}$ is the inter-protein-ligand FAPE loss, only computed for frame-point pairs between motif amino acids and ligand atoms (recall, ligand atoms are always considered motif). $L_{FAPE,sm}$ is the FAPE associated with ligands internally only. The values of their respective weights can be found in Table S5.

While L_{disp} and $L_{disto,anglo}$ are identical to how they were defined and applied in (36), the application of FAPE losses on protein and ligand motif fragments is new to the CA RFdiffusion diffusion model. The FAPE losses were applied to encourage precise and robust reconstruction of ideal motif geometries in the final output of the network. Note that because FAPE by definition scores the quality of an N-CA-C rigid frame orientation, the diffusion model will produce full backbone heavy atom predictions for motif regions, and CA-only predictions in non-motif regions. All small molecules are considered to be a motif and thus encoded in the template structure input.

Training the refinement model. Because the diffusion model only produces CA positions in non-motif regions, a second model ("the refinement model") was trained to take in CA-only descriptions of backbones and produce a refined backbone with all N-CA-C positions and orientations fully described. In short, a model was trained to take in natural protein structures with slightly noised CA positions and random N-CA-C frame orientations, and reconstruct the native backbone heavy atoms. Note this is not a diffusion process, but instead a single shot structure noising and refinement.

Architecture. The architecture of this model is precisely identical to that of the diffusion model described above.

Preparing training examples. Training dataset, masking strategy and motif templating are very similar to that described above for the diffusion model. The only differences are (1) Instead of noising the resulting structures in a forward diffusion process, a single instance of 3D Gaussian noise is added to all CA atoms and ligand atoms in the structure, and the orientation of N-CA-C frames is randomized. (2) The model is trained in two stages in which the datasets sampled are different: We trained for the first 4 epochs with variance 1.0 Å2 on the PDB monomer only dataset (i.e., no ligands), then for 3 additional epochs with variance 1.5 Å2 on both the PDB monomer only dataset and the monomer + ligands dataset. For N-CA-C frame orientation randomization, we simply use the scipy.spatial.transform.Rotation object

to produce random rotation matrices for our frames via Rotation.random(L), and orient them accordingly. These examples are then fed to the model, and it is tasked with reproducing the native protein structure.

Loss function. The loss function for refinement is very similar to $L_{diffusion}$ defined above:

$$L_{refinement} = L_{diffusion} + W_{FAPE,prot\ (non\text{-}motif)}\, L_{FAPE,prot\ (non\text{-}motif)} + W_{pLDDT}\, L_{pLDDT}$$

Where $L_{FAPE,prot\ (non\text{-}motif)}$ is the FAPE loss associated with any non-motif regions (which are always protein only regions), $W_{FAPE,prot\ (non\text{-}motif)}$ is its corresponding weight, L_{pLDDT} is the loss associated with predicting the LDDT of the refined structure (see (41) for details of function) with weight factor W_{pLDDT}. Additionally, for refinement training we set $W_{disp} = 0$ (i.e., no loss from the coordinate displacement loss function) and let FAPE serve as the dominant loss signal. See Table S5 for the weight values for specific components of the loss function.

Running inference. Inference is run by running a denoising diffusion trajectory with the trained diffusion model, templating a desired motif and constraining inter-motif-chunk DOFs as the user desires (in this study, we constrain all components of the motif with respect to all others). The output from this denoising trajectory is then fed into the refinement model, which receives a template of the original perfect motif, as opposed to a template of the motif as it was reconstructed by the diffusion model. The refinement model then produces a full heavy atom prediction of the backbone, usually reconstructing the motif to within less than 0.2 Å of the native/input. For sidechains within the motif, we rebuild the rotamers from the input motif onto the corresponding residues in the output, using the sidechain dihedral angles from the original motif, and ideal bond lengths and angles. At this stage, the output is ready for sequence design.

Supplementary Figures

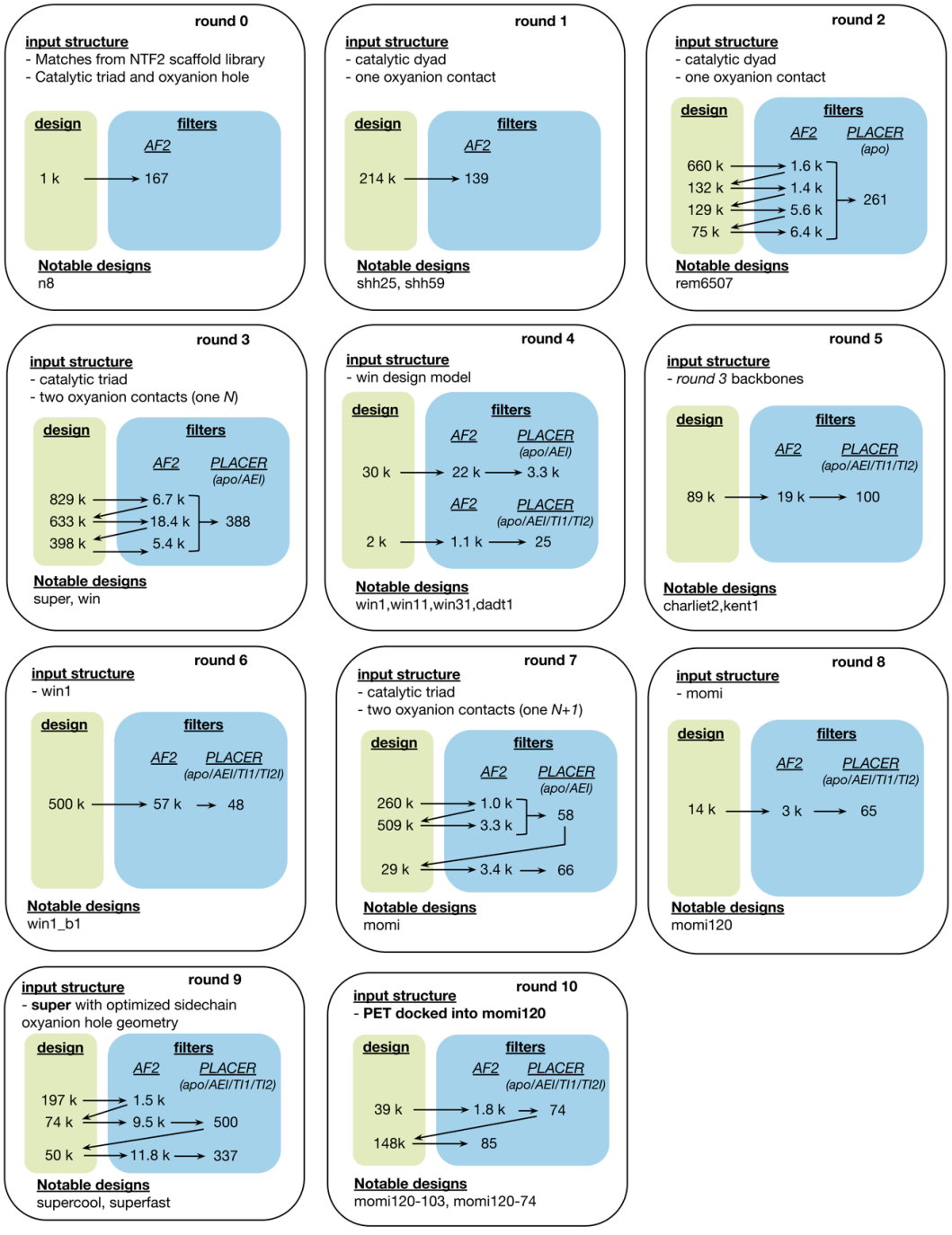

Figure S1. Summary of design rounds. Numbers represent generated designs (green boxes) or number of designs passing filters (blue boxes). Diagonal arrows indicate designs that passed filters in a previous round and were redesigned again with LigandMPNN and FastRelax to generate more designs. Key designs described in the main text are listed in the notable designs section.

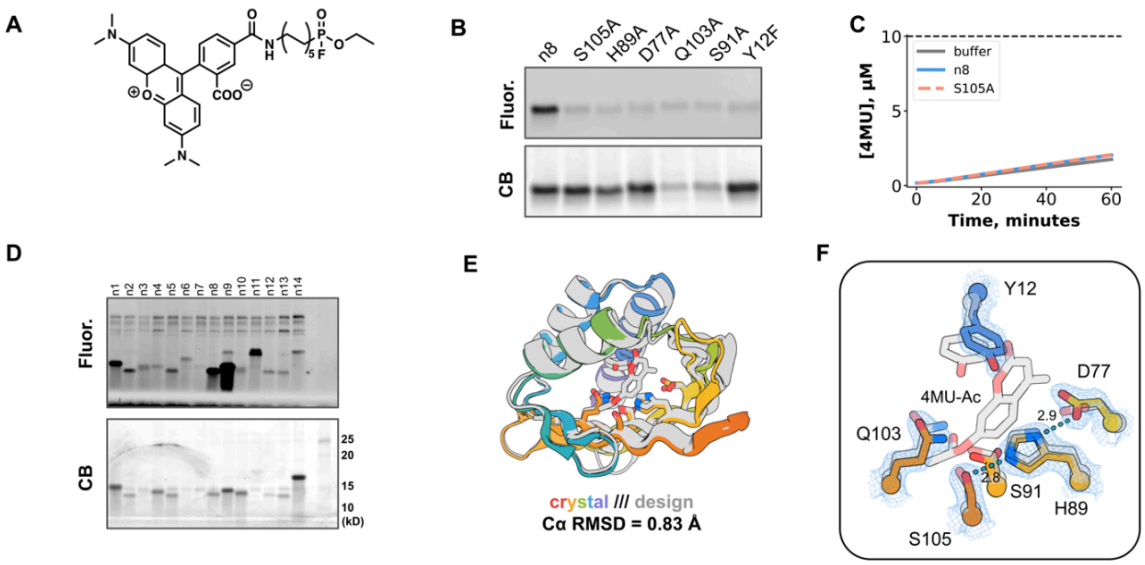

Figure S2. Serine hydrolase design and characterization with hallucinated NTF2s. (**A**) Chemical structure of TAMRA-conjugated fluorophosphonate probe (FP-probe). (**B**) In-gel fluorescence of catalytic residue alanine knockouts in probe-reactive design **n8**. Fluor. is fluorescence imaging and CB is Coomassie blue stained. (**C**) Reaction progress curves of 10 µM **n8** incubated with 100 µM 4MU-Ac in 20 mM HEPES, 50 mM NaCl, pH 7.4. Shaded area represents standard deviation of three technical replicates and dashed line represents enzyme concentration. (**D**) In-gel fluorescence of NTF2-based serine hydrolase design cell lysates after 1 hour incubation with 1 µM FP-TAMRA. Fluor. is fluorescence imaging and CB is Coomassie blue stained. Molecular weights of designs range from 15 to 18 kD. (**E,F**) Structural superposition of **n8** crystal structure and design model with 4MU-Ac docked in place of fluorophosphonate probe.

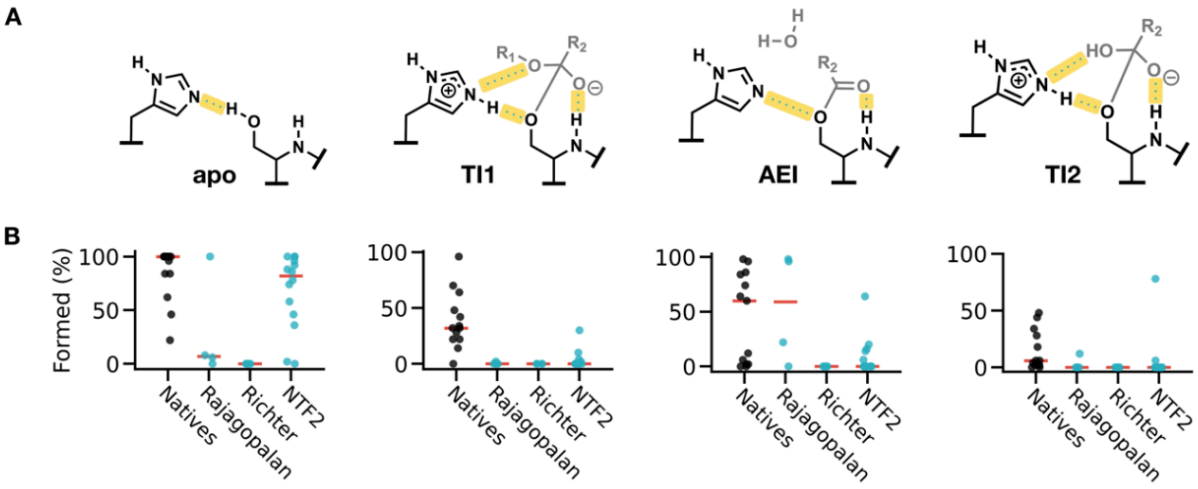

Figure S3. Comparing preorganization of native and designed serine hydrolases with PLACER. (**A**) Reaction states modeled with PLACER. Key hydrogen bonding interactions are highlighted in yellow. (**B**) Percent PLACER ensemble frames in which all of the key hydrogen bonding interactions are simultaneously formed for each step (apo, TI1, AEI, TI2 from left to right), with designs in teal, natural hydrolases in black, and medians for each distribution indicated in red. P-values for comparison of natural hydrolases and design distributions were obtained by K-S test and are 0.09, 4.9e-7, 0.03, and 0.0001 for the apo, TI1, AEI, and TI2 states, respectively. Designs labeled "Richter" and "Rajagopalan" were reported in references 6 and 7, respectively. Natural hydrolases (n=13) were obtained from the PDB (see Methods, "Filtering"). Richter (n=4), Rajagopalan (n=4), and NTF2 (n=14) designs were all generated by matching followed by Rosetta sequence design, with the primary difference being the scaffold libraries (native proteins for Richter and Rajagopalan, and hallucinated NTF2s for the NTF2 set) and the input motifs (Cys-His dyads for the Richter set, organophosphate-targeted catalytic triads for the Rajagopalan set, and catalytic dyads and triads targeted toward organophosphates and esters for the NTF2 set) utilized at the matching stage.

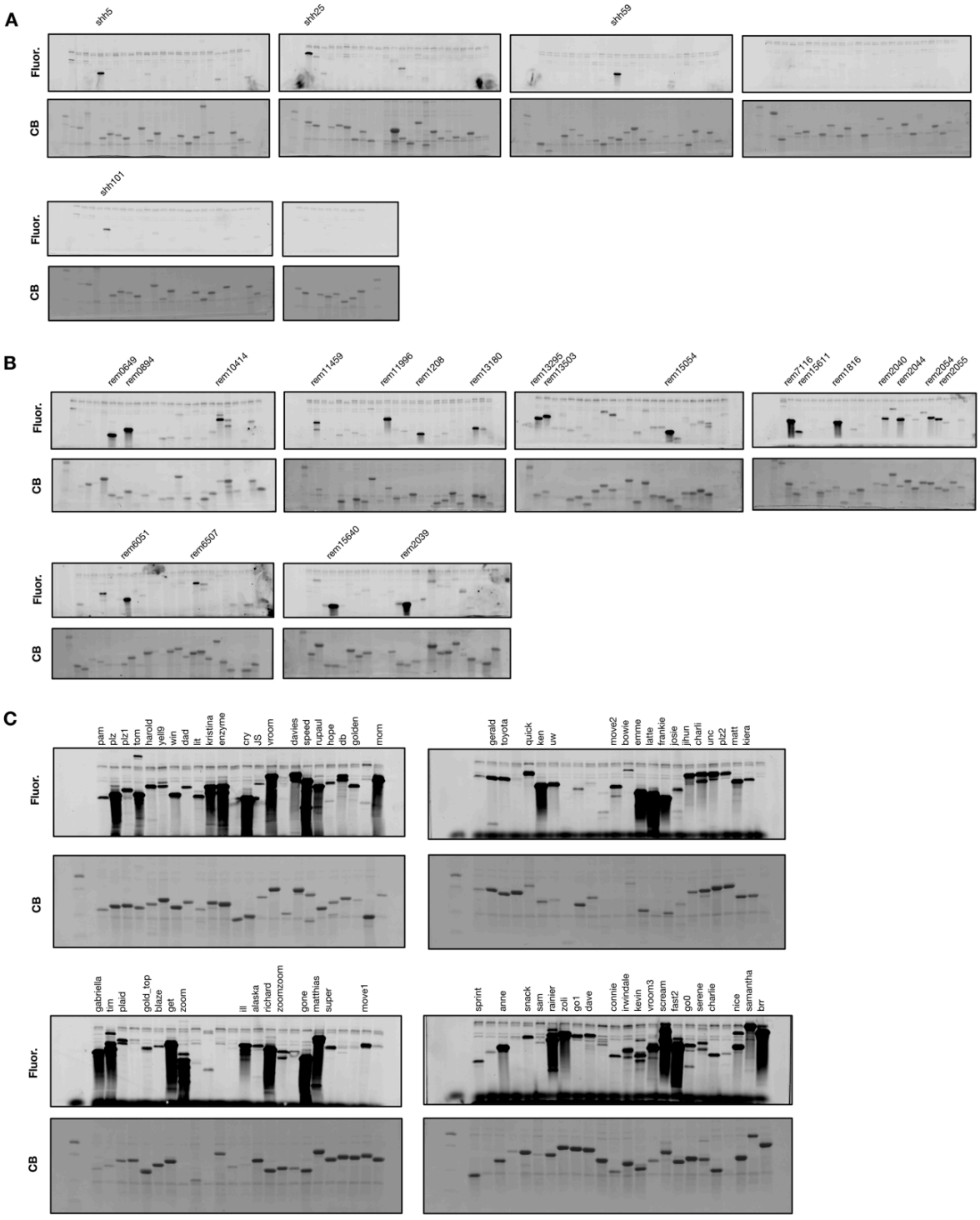

Figure S4. In-gel fluorescence imaging with fluorescently labeled fluorophosphonate activity-based probe. In-gel fluorescence of (**A**) round 1, (**B**) round 2, and (**C**) round 3 designs after 1 hour incubation of cell lysate with 1 μM FP-TAMRA. Fluor. stands for fluorescence and CB for Coomassie blue. Lanes with labels indicate designs that were purified and tested for esterase activity.

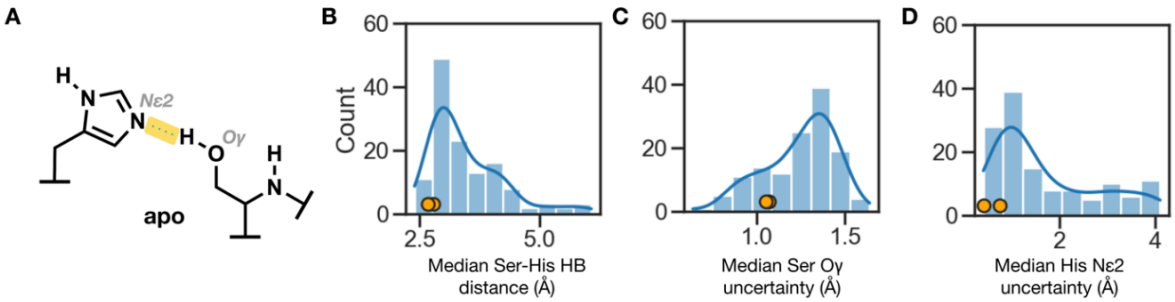

Figure S5. PLACER analysis suggests single-turnover designs are more preorganized. (**A**) Schematic depicting catalytic serine-histidine hydrogen bond interaction. (**B**) Median Ser-His H-bond distance, **(C)** median serine Oγ uncertainty, and (**D**) median histidine Nε2 uncertainty calculated from PLACER ensembles of the apo state for 130 round 1 designs. Yellow points represent the two single-turnover designs, **shh25** and **shh59.**

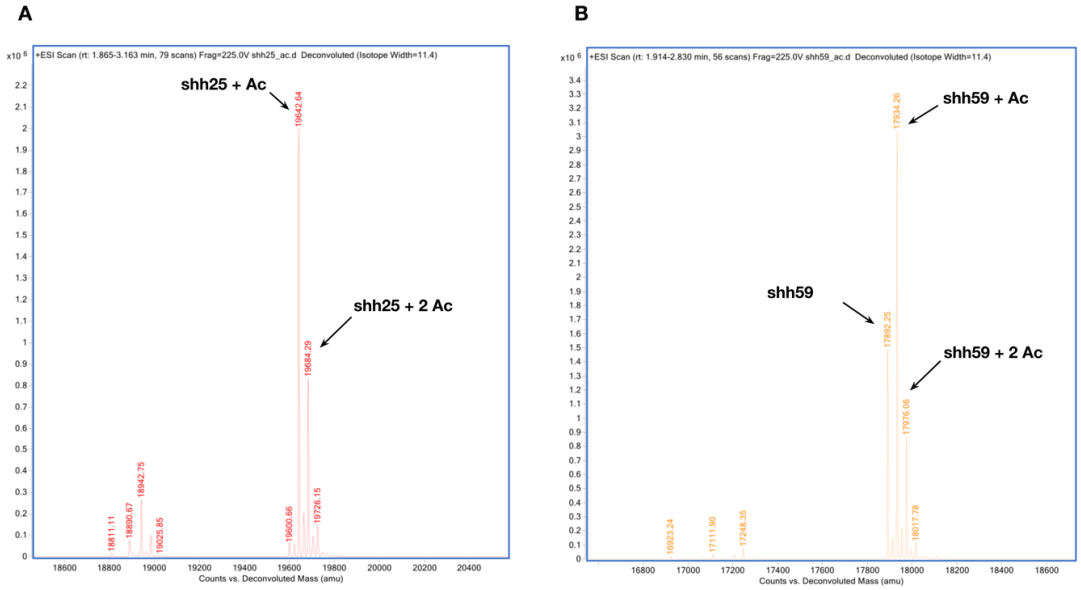

Figure S6. Mass spectra of single-turnover designs incubated with cognate ester substrates. Mass spectra of shh25 (**A**) and shh59 (**B**) after overnight incubation with 100 µM of 4MU-acetate in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO buffer.

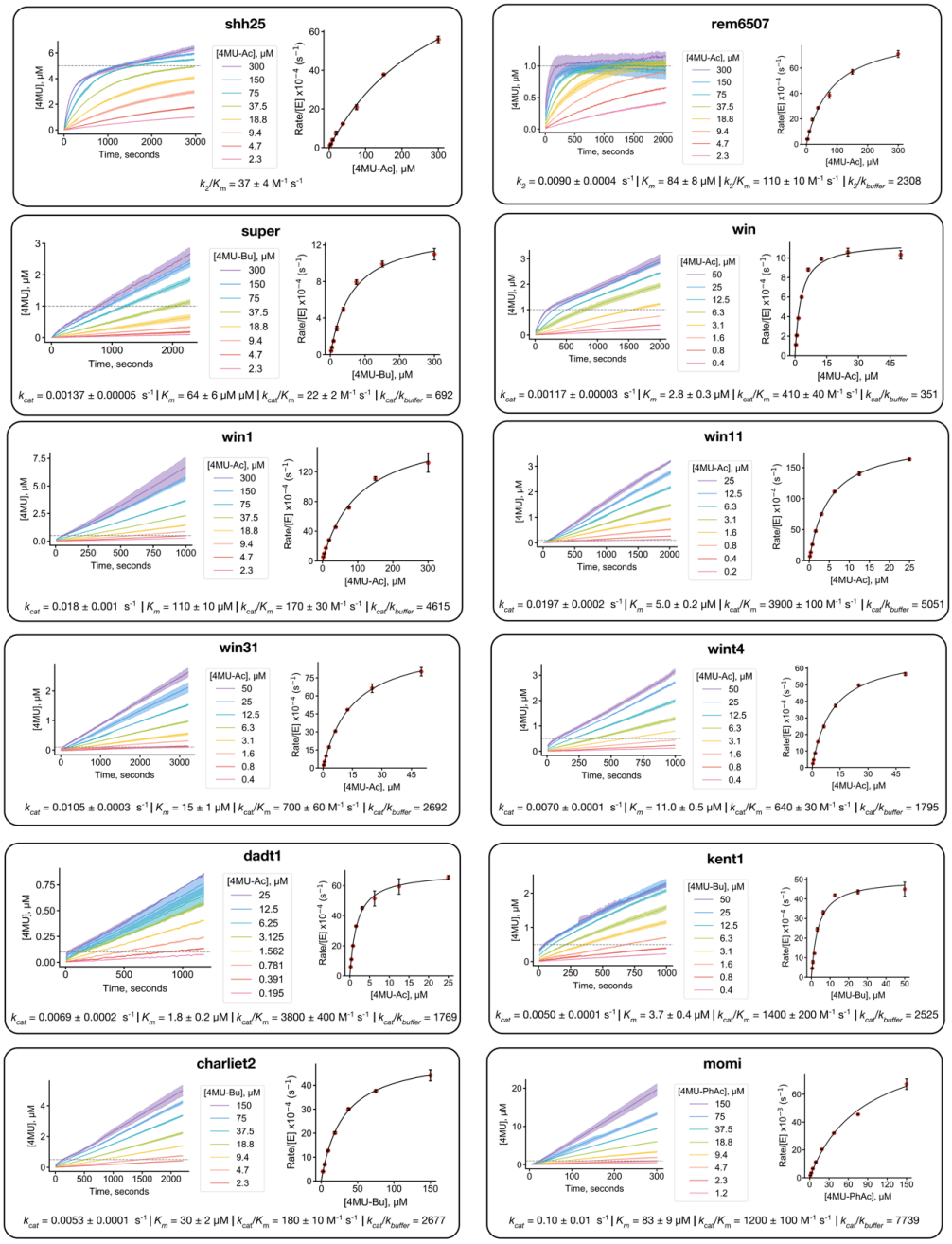

Figure S7. Michaelis-Menten kinetics of designed serine hydrolases. Reactions were performed at 30°C in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO assay buffer. Shaded area in progress curves and error bars in Michaelis-Menten plots represent standard deviation of three technical replicates. Horizontal dashed lines represent enzyme concentrations.

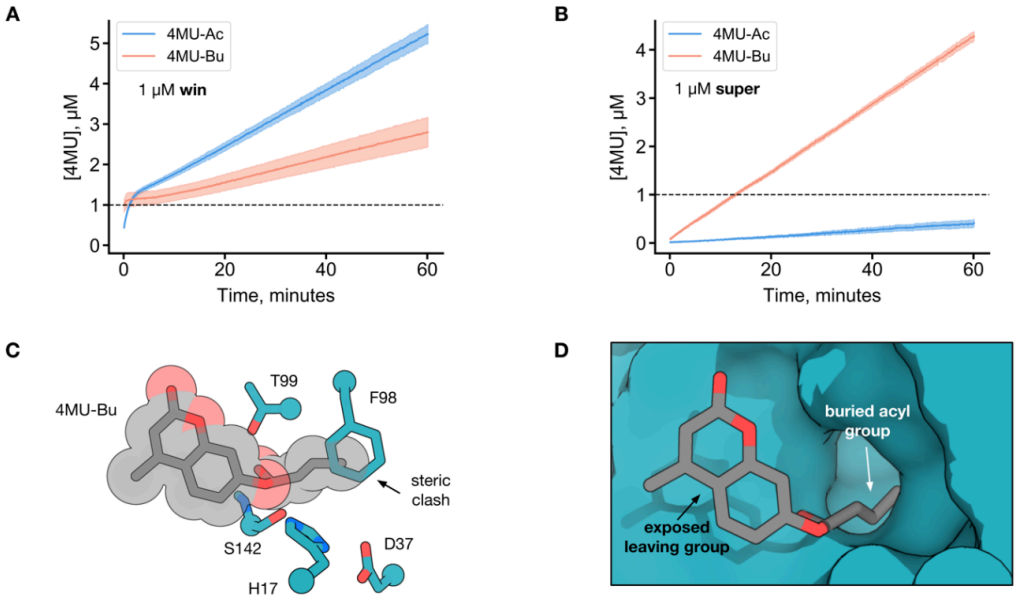

Figure S8. Substrate selectivity of win and super. Progress curves of (**A**) **win** and (**B**) **super** incubated with 50 μM of either 4MU-Ac or 4MU-Bu in 20 mM HEPES, 50 mM NaCl, pH 7.4 at 30°C. Dashed line indicates enzyme concentration and shaded area represents the standard deviation of three technical replicates. (**C**) **win** crystal structure with 4MU-Bu modeled in the active site. (**D**) Surface representation of **super** crystal structure with 4MU-Bu docked into the active site.

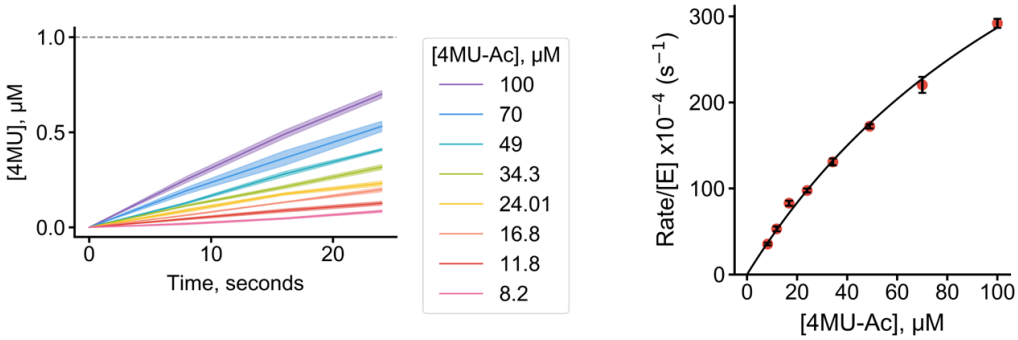

Figure S9. Burst phase kinetics of win. Progress curves (left) and Michaelis-Menten plot (right) of the burst phase of **win** (1 μM) incubated with up to 100 μM of 4MU-Ac does not reach a saturating velocity, suggesting $K_s >> K_m$ (K_m = 2.8 μM). Dashed line represents enzyme concentration and shaded area represents standard deviation of three technical replicates in the progress curve plot.

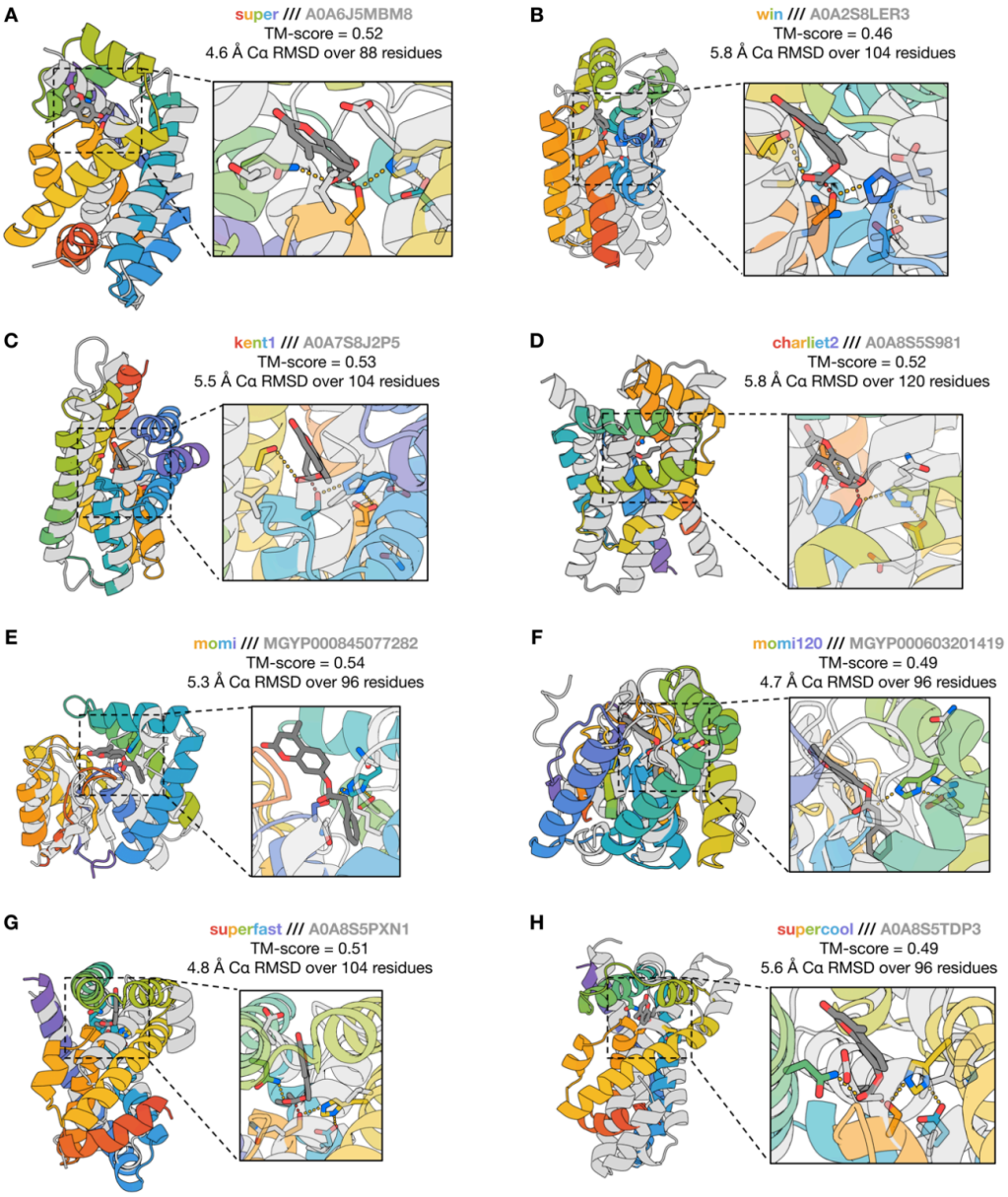

Figure S10. Structural novelty of designed esterases. Overlay of design models and most similar structures by TM-score from Foldseek searches for (**A**) **super**, (**B**) **win31**, (**C**) **kent1**, (**D**) **charliet2**, (**E**) **momi**, (**F**) **momi120**, (**G**) **superfast**, and (**H**) **supercool.**

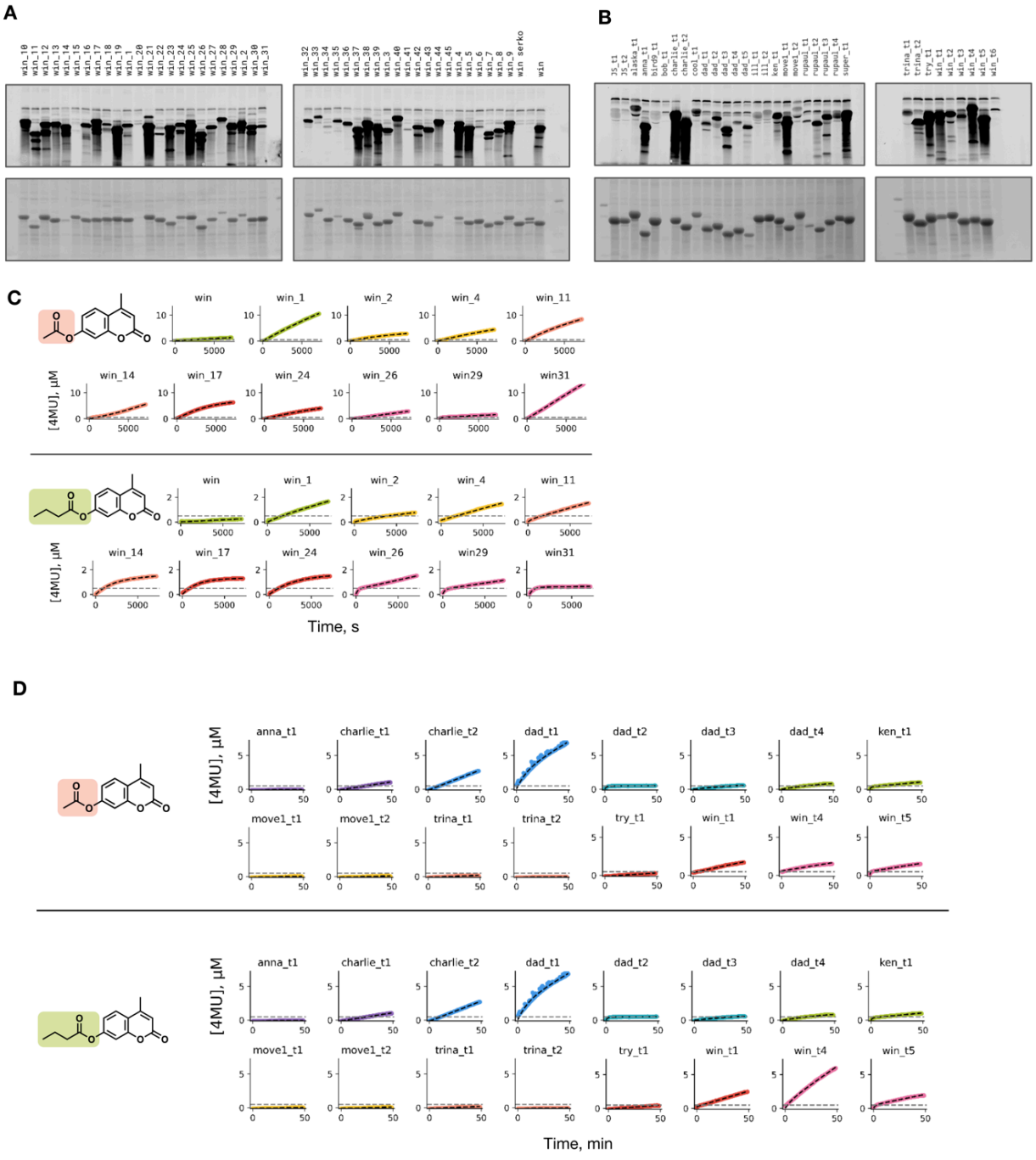

Figure S11. Screening and kinetic analysis of resampled designs. In-gel fluorescence imaging after incubation of cell lysates of (**A**) **r3-win** redesigns and (**B**) round 3 redesigns with 1 µM FP-TAMRA. Progress curves of (**C**) **r3-win** redesigns and (**D**) round 3 redesigns incubated at 30°C with 100 µM 4MU-Ac and 4MU-Bu in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO assay buffer. Dashed line represents enzyme concentration (0.5 µM).

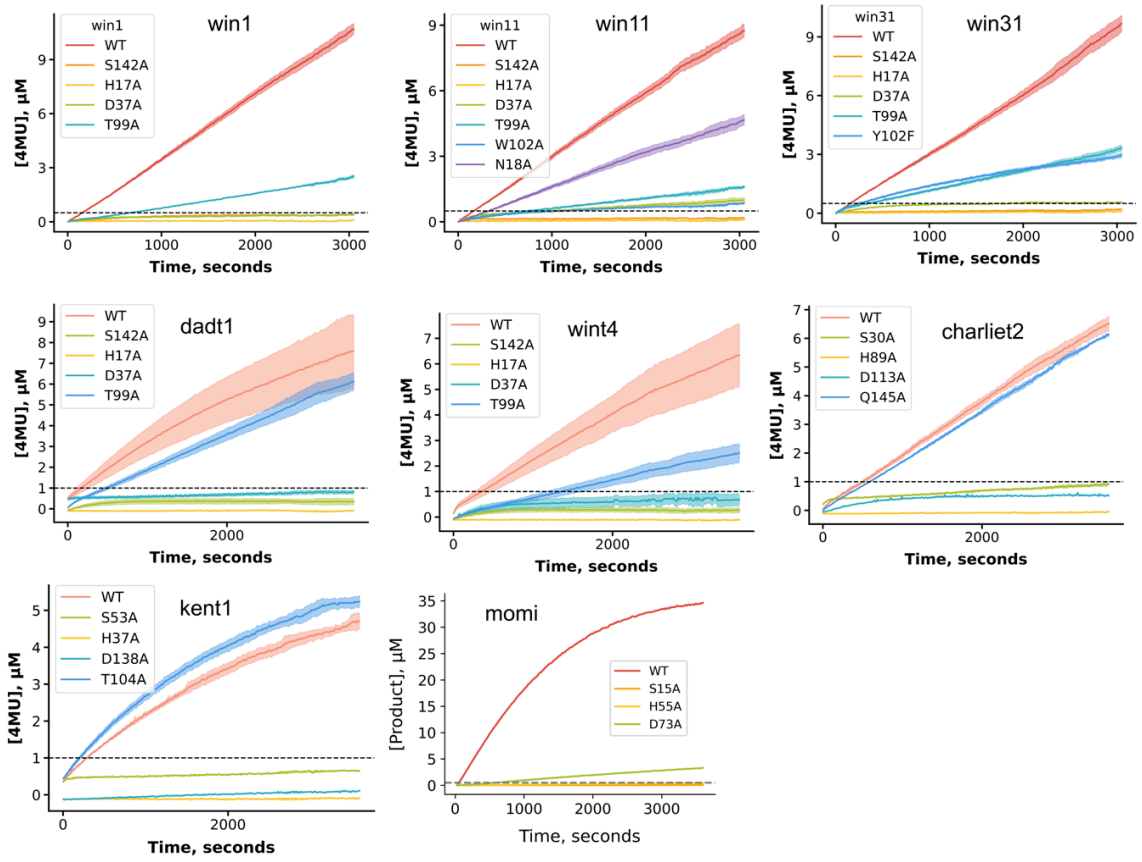

Figure S12. Progress curves for active site residue mutants of designed serine hydrolases. Reactions were performed at 30°C with 1 µM or 0.5 µM (**win1**, **win11**, **win31**) enzyme and 100 µM 4MU-Ac, 100 µM 4MU-Bu (**charliet2**, **kent1**), or 50 µM 4MU-PhAc (**momi**) in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO. Dashed lines indicate the enzyme concentration and shaded areas represent standard deviation of three technical replicates.

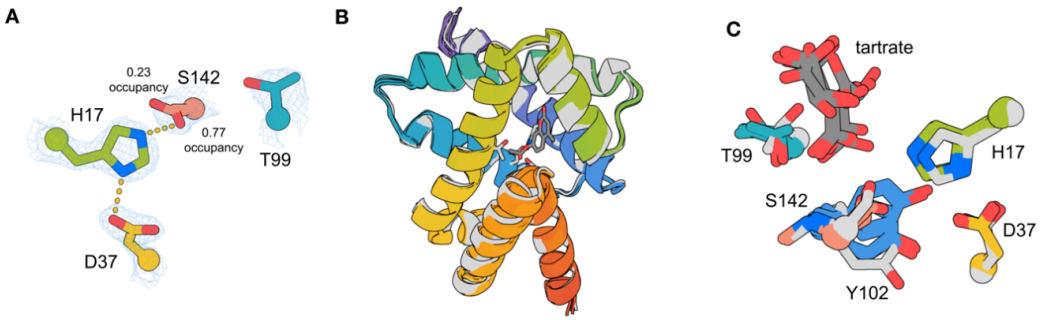

Figure S13. Analysis of win1 and win31 crystal structures. (**A**) The catalytic serine S142 in chain B of **win1** occupies two distinct rotameric states. 2Fo-Fc map in blue mesh shown at 1 σ. (**B**) Structural superposition of overall fold of all five chains in the asymmetric unit of **win31** (color) and **win31** design model (gray) (**C**) Active site zoom-in of structural superposition of the five chains in the **win31** asymmetric unit (color) overlaid with design model (gray).

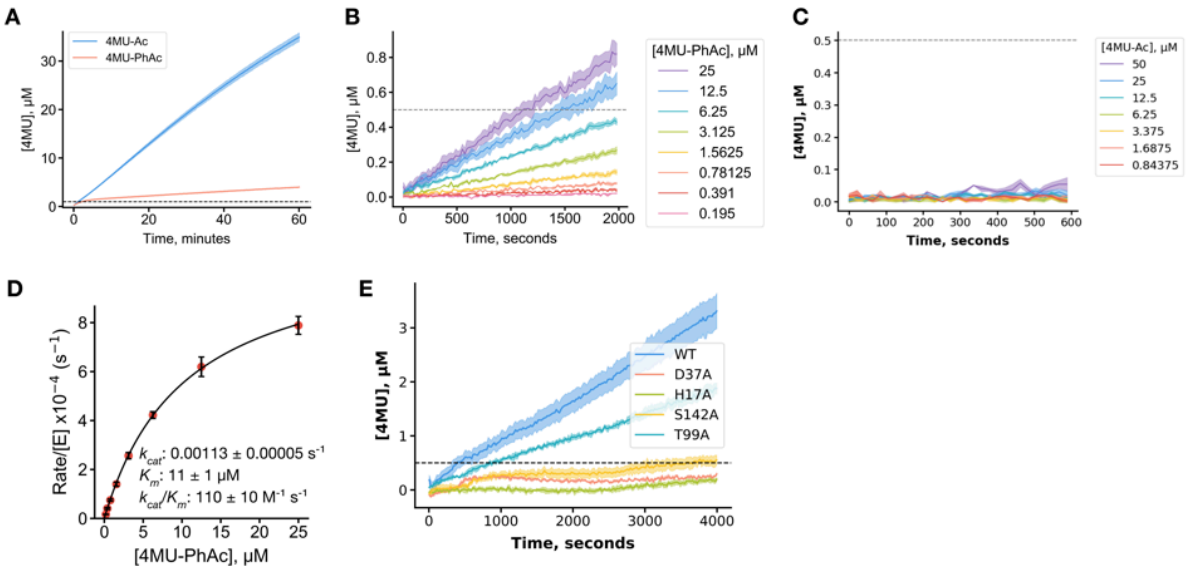

Figure S14. Characterization of redesigns of win1 for larger 4MU-PhAc substrate. (**A**) Progress curves of **win1** with 100 µM 4MU-Ac (blue) and 4MU-PhAc (red) indicate preference for 4MU-Ac. (**B**,**C**) Progress curves of **win1_b1** with 4MU-PhAc (**B**) and 4MU-Ac (**C**). (**D**) Michaelis-Menten plot of **win1_b1** with 4MU-PhAc. (**E**) Progress curves for catalytic residue knockout mutants of **win1_b1** with 100 µM 4MU-PhAc. Reactions were performed at 30°C in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO with 1 µM (**win1**) or 0.5 µM (**win1_b1**) enzyme. Dashed lines indicate the enzyme concentration and shaded areas represent standard deviation of three technical replicates.

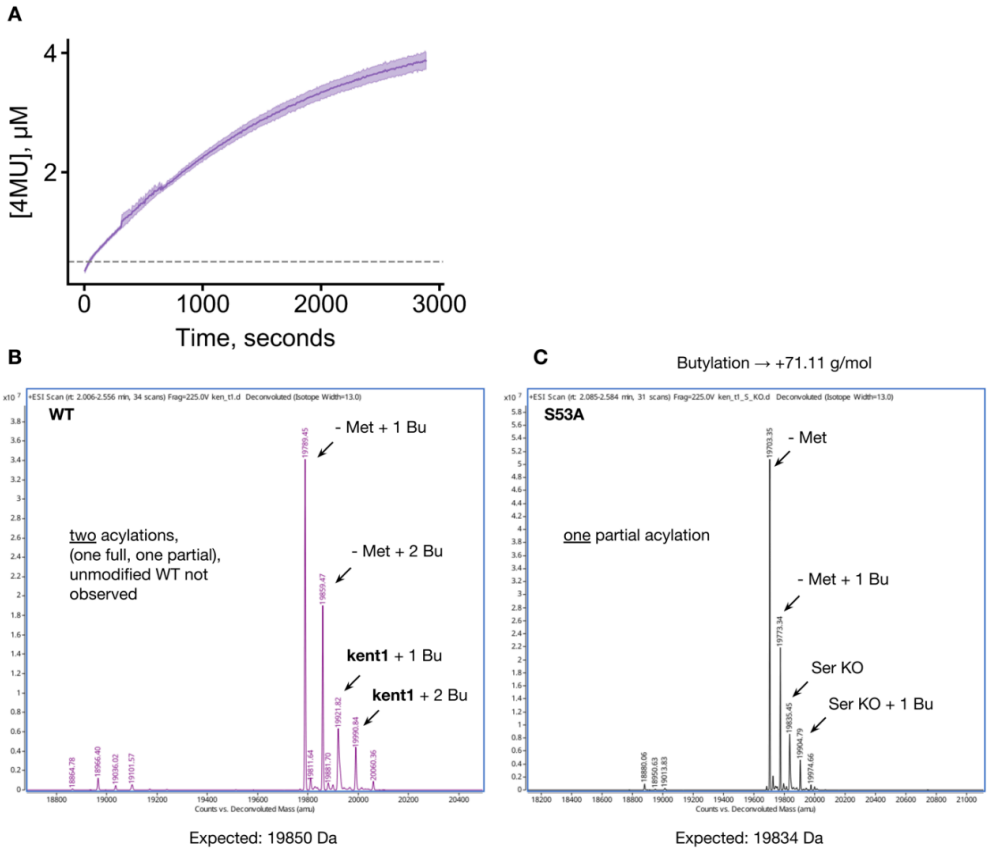

Figure S15. Kinetic and mass spectrometric analysis of kent1 reveal inactivation over time and stable acylated species. (**A**) Progress curve of 0.5 µM **kent1** incubated with 50 µM 4MU-Bu in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO assay buffer at 30°C. Dashed line represents the enzyme concentration and shaded area represents standard deviation of three technical replicates. (**B**) Mass spectra of WT and catalytic serine knockout (S53A) of 50 µM **kent1** after overnight incubation at room temperature with 100 µM 4MU-Bu.

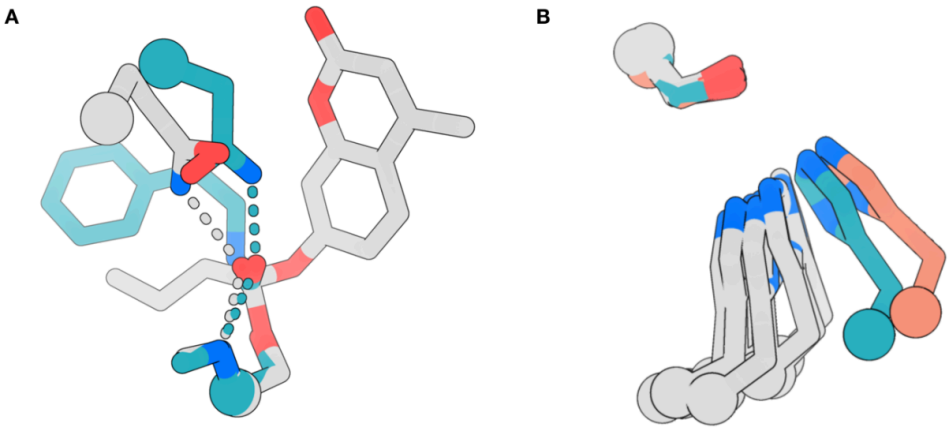

Figure S16. Comparison of catalytic geometries in designed and native serine hydrolases. (**A**) Structural alignment of **super** design model (gray) and crystal structure of subtilisin (teal) (PDB: 1scn) highlighting distinct oxyanion hole geometries of side chain glutamine (super) and asparagine (subtilisin) relative to the substrate carbonyl. **super's** oxyanion H-bond may stabilize the ground state and slow catalysis. (**B**) Comparison of His-Ser interactions found among native hydrolases (gray) and those in **super** (teal) and **win** (peach).

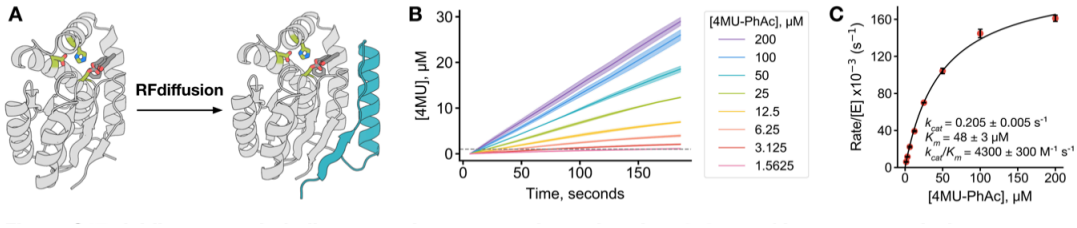

Figure S17. Adding second-shell structural support to the active site of momi improves catalysis. (**A**) Structural extension of N-terminus of **momi** with RFdiffusion to generate **momi120**. Original structure in gray, newly built region in teal. (**B**) Progress curves of **momi120** with increasing concentrations of 4MU-PhAc. Horizontal dashed lines indicate the enzyme concentration (1 μM) and shaded areas represent the standard deviation of three technical replicates. (**C**) Michaelis-Menten plot of **momi120** with 4MU-PhAc. Kinetic assays were performed in 20 mM HEPES, 50 mM NaCl, pH 7.4, 5% DMSO assay buffer at 30°C.

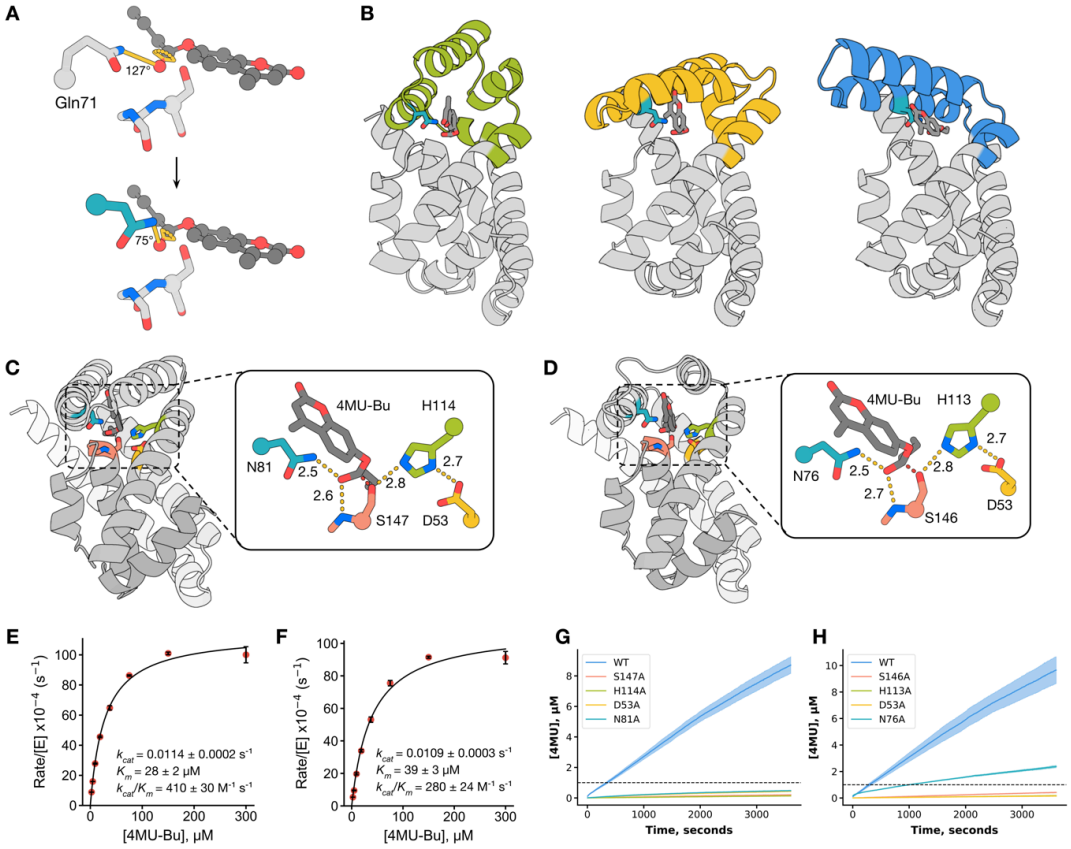

Figure S18. Adjusting super oxyanion hole geometry with RFdiffusion improves catalysis. (**A**) Repositioning of sidechain oxyanion hole Gln71 to a form a more out-of-plane H-bond with substrate carbonyl group (dihedral angle indicated in yellow between substrate ester and amino group shifted from 127 degrees in starting structure to 75 degrees). (**B**) Selected design models illustrating the diversity of the RFdiffusion-built region (green/yellow/blue) scaffolding the new sidechain oxyanion hole residue (teal). (**C**,**D**) Overall fold and active site zoom-in of **superfast** (**C**) and **supercool** (**D**) design model. (**E,F**) Michaelis-Menten plot of **superfast** (**E**) and **supercool** (**F**) with 4MU-Bu. (**G,H**) Progress curves of catalytic residue knockout mutants of **superfast** (**G**) and **supercool** (**H**). Dashed line represents the enzyme concentration (1 µM) and shaded area represents standard deviation of three replicates.

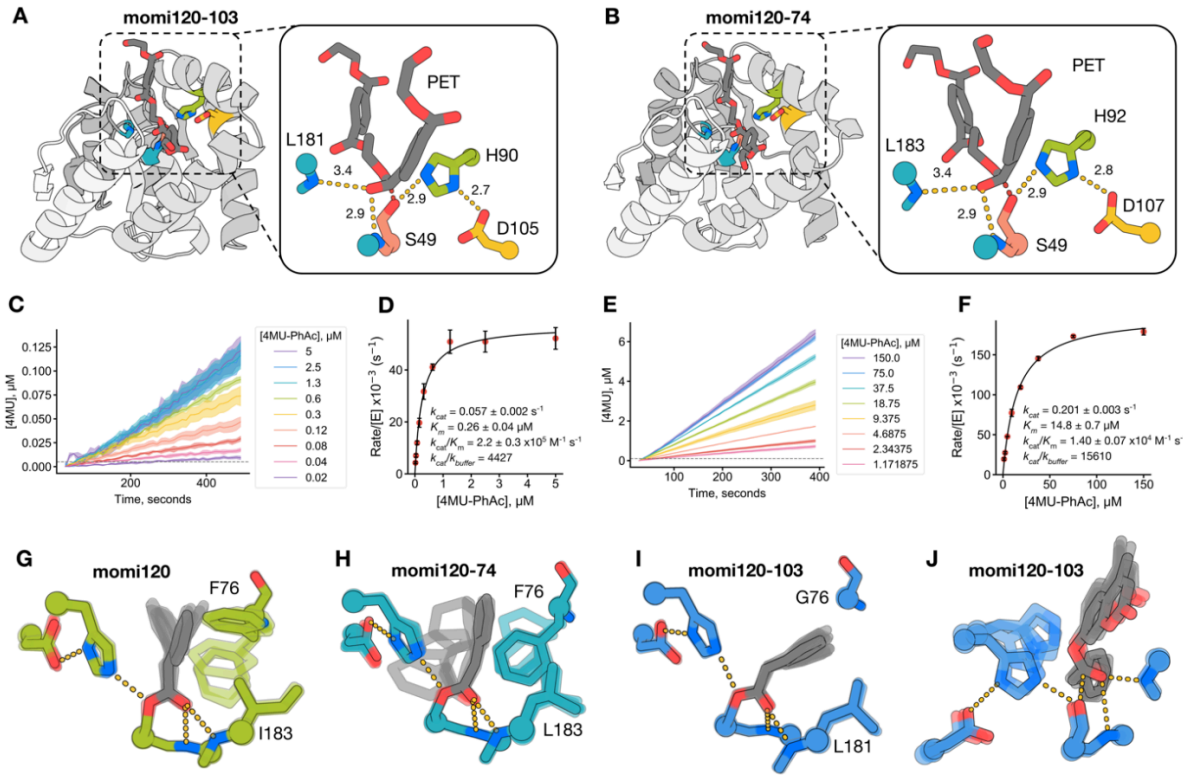

Figure S19. Structural modeling and kinetic characterization of momi120 redesigns. (**A,B**) Overall fold and active site zoom-in of momi120-103 (**A**) and momi120-74 (**B**) design models. (**C,E**) Progress curves of momi120-103 (**C**) and momi120-74 (**E**) incubated with 4MU-PhAc. (**D,F**) Michaelis-Menten plots of momi120-103 (**D**) and momi120-74 (**F**) with 4MU-PhAc. (**G-I**) PLACER ensembles of momi120 (**G**), momi120-74 (**H**), and momi120-103 (**I**) in the AEI state, highlighting active site differences involved in substrate binding at positions 76 and 181/183. (**J**) Superposition of five Chai-1 predictions of momi120-103 in complex with 4MU-PhAc.

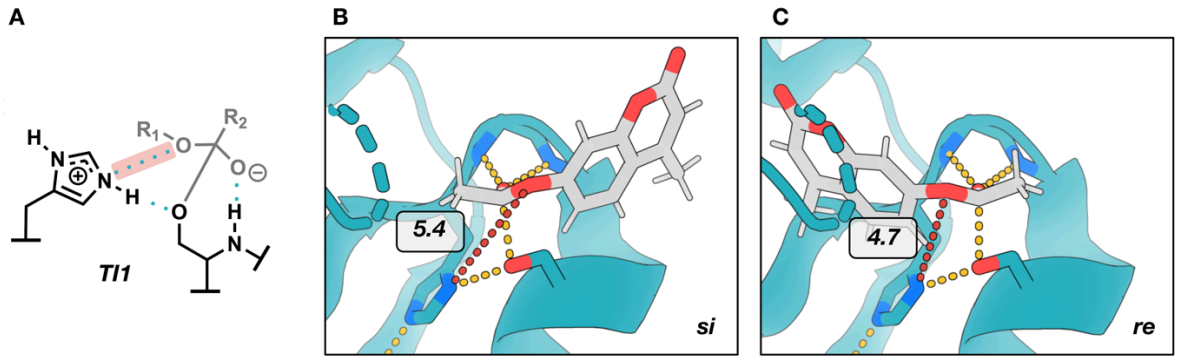

Figure S20. Catalytically incompatible geometries in previous serine hydrolase design OSH55. (A) Key interaction in tetrahedral intermediate 1 (TI1) between histidine Nε2 and substrate leaving group oxygen (red highlight). (**B, C**) Crystal structure of OSH55 design (E6D/L155R variant, PDB ID 4ess) shown in complex with computationally docked 4MU-Ac esterase substrate for both *re* and *si* faces of attack. The key histidine-leaving group interaction is depicted with red dashes. Distances in Å.

Table S1. Kinetically or structurally analyzed designs from each round.

Design round	Design names
Round 0	n8
Round 1	shh25
Round 2	rem6507
Round 3	super, win, josie
Round 4	win1, win11, win31, dadt1
Round 5	charliet2, kent1
Round 6	win1_b1
Round 7	momi
Round 8	momi120
Round 9	supercool, superfast

Round 10	momi120_74, momi120_103

Table S2. Data collection and refinement statistics. Statistics for the highest resolution shell are shown in parentheses.

	r0-n8	r3-super	r3-win	r4-win1	r4-win31	r4-dadt1
PDB ID	9DED	9DEE	9DEF	9DEG	9DEH	9MRB
Wavelength	0.97648	0.92010	1.00002	1.00002	0.97936	0.97934
Resolution range	40.00 - 1.77 (1.81 - 1.77)	32.12 - 1.21 (1.28 - 1.21)	46.02 - 1.75 (1.78 - 1.75)	43.27 - 2.11 (2.17 - 2.11)	58.33 - 2.21 (2.33 - 2.21)	51.19 - 2.001 (2.13 - 2.0)
Space group	P 1 21 1	R 3 :H	P 21 2 21	P 21 21 21	P 21 21 21	C 1 2 1
Unit cell	38.6, 31.6, 40.8; 90, 101.3, 90	68.4, 68.4, 76.4; 90, 90, 120	31.2, 46.0, 92.9; 90, 90, 90	35.8, 82.0, 101.9; 90, 90, 90	63.4, 107.0, 116.7; 90, 90, 90	42.4, 56.7, 51.6; 90, 97.23, 90
Total reflections	63121 (3594)	2592198 (38409)	153901 (8688)	192423 (15480)	457352 (66573)	35618 (6028)
Unique reflections	9440 (526)	40645 (5957)	14105 (749)	17989 (1414)	40598 (5849)	8217 (1357)
Multiplicity	6.7 (6.8)	6.4 (6.4)	10.9 (11.6)	10.7 (10.9)	11.3 (11.4)	4.3 (4.4)
Completeness (%)	98.8 (99.3)	99.86 (100.00)	99.80 (99.60)	99.9 (99.9)	100.00 (100.00)	98.77 (98.62)
Mean I/sigma(I)	11.4 (2.4)	12.6 (1.1)	14.0 (1.9)	10.3 (2.4)	6.4 (0.70)	4.04 (0.92)
Wilson B-factor	22	21	27	28	39	29
R-merge	0.090 (0.831)	0.048 (1.347)	0.084 (1.274)	0.151 (0.958)	0.284 (3.183)	0.185 (1.494)
R-pim	0.041 (0.368)	0.022 (0.639)	0.028 (0.402)	0.050 (0.314)	0.090 (1.024)	0.099 (0.787)
CC1/2	0.997 (0.798)	0.999 (0.471)	0.998 (0.752)	0.999 (0.918)	0.977 (0.424)	0.992 (0.566)
Reflections used in refinement	9422 (1332)	40645 (2937)	14057 (1366)	17872 (1321)	40457 (2855)	8173 (1355)
Reflections used for R-free	935 (138)	2012 (150)	1406 (137)	1788 (131)	2004 (138)	818 (136)

R-work	0.1896 (0.2479)	0.2057 (0.3380)	0.2361 (0.3217)	0.2039 (0.2557)	0.2381 (0.3796)	0.2131 (0.2667)
R-free	0.2216 (0.2904)	0.2414 (0.3917)	0.2726 (0.3556)	0.2590 (0.3229)	0.2697 (0.3863)	0.2575 (0.3123)
Number of non-hydrogen atoms	1044	1451	1284	2700	6346	1181
macromolecules	969	1292	1215	2535	6204	1131
ligands	0	0	5	6	83	13
solvent	75	159	64	159	59	37
Protein residues	114	160	151	314	790	141
RMS(bonds)	0.019	0.006	0.007	0.011	0.008	0.002
RMS(angles)	1.62	0.60	0.81	1.03	0.75	0.40
Ramachandran favored (%)	98.21	99.37	97.99	99.68	99.36	100.00
Ramachandran allowed (%)	1.79	0.63	2.01	0.32	0.64	0
Ramachandran outliers (%)	0.00	0.00	0.00	0.00	0.00	0
Average B-factor	26	27	40	33	49	31
macromolecules	26	26	40	33	49	31
ligands			34	51	56	44
solvent	315	35	35	36	40	34

Table S3. Template indicator, timestep and structural features input to both the diffusion and refinement models of CA RFdiffusion. Template #1 is the self-conditioning template, exactly as described in(*18*). For the new per token "is motif" indicator, this template receives a null value of -1 for all tokens, while still containing the encoding of the current discrete time step 1-T/t. The pairwise template features for template #1 contain the self-conditioning information (the distance/angle encoding of the previous prediction of X_0), and a null -1 indicator feature for all pairs. Template #2 contains null information (i.e., -1) for all indicator and timestep features, but the pairwise structure input contains the distance/angle encoding of the current noisy X_t structure being input to the network in 3D. Template #3

contains information about the perfect motif geometry and primary sequence location, to be scaffolded by the network. This template is the only one of the three which uses the "is motif" feature (0 if non-motif, 1 if motif token) and the "pair is constrained" indicator (1 if the pair is constrained, else 0).

	Per-token (1D) info	**Per-token-pair (2D) info**
Template #1 - Self-conditioning	**Timestep feature:** 1-T/t **Is motif feature:** -1	**Structural data:** Previous prediction of X_0 Pair constraint indicator: -1
Template #2 - Current X_t input	**Timestep feature:** -1 **Is motif feature:** -1	**Structural data:** Current X_t structure input Pair constraint indicator: -1
Template #3 - Perfect motif info	**Timestep feature:** -1 **Is motif feature:** 0 for False, 1 for True	**Structural data:** Perfect motif geometry **Pair constraint indicator:** 1 if token pair is geometrically constrained, else 0.

Table S4. Training dataset and masking strategy probabilities for diffusion model training. For protein-only monomers from the PDB, which is chosen 60% of the time, there are three possible masking strategies. (**A**) A "chunked" diffusion mask, in which 1-8 discontiguous fragments are designated as motif components. Each pair of discontiguous fragments has some nonzero probability of being unconstrained with respect to each other (see Algorithm 1 for details). (**B**) Multiple "triple contacts" are chosen, in which three residues that are close in spatial proximity but far from each other in primary sequence are selected (see Algorithm 2 for details). (**C**) Unconditional, in which there are no motifs and the task is to denoise the full structure. For the second dataset, which is PDB monomers in complex with small molecule ligands, there is a single motif masking strategy used every time. In this "small molecule contact mask" generation, the structure of the ligand is revealed in the template, along with either 0, 1, or 2 additional discontiguous protein fragments in close proximity to the ligand (see Algorithm 3 for details).

Dataset probabilities	**Masking probabilities**
Dataset #1: PDB monomers - 60% probability	(A) "Chunked diffusion mask" - 20% (B) Multi triple contact - 50% (C) Unconditional - 30%
Dataset #2: PDB monomers in complex with ligands - 40% probability	(A) Small molecule contact mask - 100%

Table S5. Training hyperparameters for CA RFdiffusion (diffusion model and refinement model).

Parameter	Description	Value
β_0	Variance at time t=0	0.01
β_T	Variance at time t=T=200	0.07
T	Total number of timesteps in noising process	200
Variance schedule type	NA	Linear
W_{disp}	CA displacement loss weight	0.5
$W_{dist\text{-}ang}$	Distogram and anglogram cross entropy loss weight	0.05
$W_{FAPE,prot}$	Intra-protein fragment FAPE loss weight	10.0
$W_{FAPE,prot\text{-}sm}$	Inter protein-ligand FAPE loss weight	10.0
$W_{FAPE,sm}$	Intra ligand FAPE loss weight	10.0
$W_{FAPE,prot(non\text{-}motif)}$	(Refinement model only) Non-motif protein FAPE loss weight	10.0
W_{pLDDT}	(Refinement model only) pLDDT loss weight	0.1
$A_{FAPE,prot}$	Normalizing constant for protein motif FAPE	5.0
$A_{FAPE,sm}$	…for intra-ligand FAPE	4.0
$A_{FAPE,prot\text{-}sm}$	…for inter protein-ligand FAPE	10
$A_{FAPE,prot(non\text{-}motif)}$	…for non-motif FAPE.	10
$D_{clamp,prot}$	Clamp upper bound for protein motif FAPE	5

$D_{clamp,sm}$	…for intra-ligand FAPE	4.0
$D_{clamp,prot-sm}$	…for inter protein-ligand FAPE	10
$D_{clamp,prot(non-motif)}$	…for non-motif FAPE	10
Batch size	Effective batch size on 8GPUs	16
Learning rate	NA	0.0005
p_show_motif_seq	Probability of showing the amino acid identity of non-ligand motif tokens	0.65
Diffusion coordinate scaling	Scaling factor applied to coordinates before the forward process, then inverted and applied after the forward process.	0.25
Refinement model 3D Gaussian variance	Variance for noising CA atoms in structures.	1.0 for epochs 1-4, 1.5 for epochs 5-7.
Epoch size	Number of examples per training epoch	25600

Table S6. Dataset proportions, mask sampling proportions, and 3D Gaussian noise variance for the two stage refinement model training procedure.

Refinement model training stage	Dataset probabilities and mask probabilities	3D Gaussian noise variance
1	Datasets: PDB monomer - 100% MasksTraining: “Chunked diffusion mask” - 20% Multi triple contact - 50%	1.0
2	Datasets: PDB monomer - 40% PDB monomer + ligand - 60% Masks (for PDB monomer set): “Chunked diffusion mask” - 20% Multi triple contact - 50% Unconditional - 30% Masks (for PDB monomer + ligand set): Small molecule contact mask - 100%	1.5

Algorithm 1. Small molecule contact mask generation. This function defines how 1D and 2D motif masks were generated for protein-ligand complex examples.

```
def _get_diffusion_mask_chunked(xyz, prop_low, prop_high, max_motif_chunks=6):
  """
  Masking strategy that creates discontiguous motifs, possibly unconstrained w.r.t each other.
Parameters:
xyz (torch.tensor, required): (L,14,3) tensor of coordinates
prop_low (float, required): lower bound on fraction of protein to mask
prop_high (float, required): upper bound on fraction of proteins to mask
  """
max_motif_chunks, chunks, is_motif, motif_ids, ij, ij_can_see = get_chunked_mask(xyz, prop_low, prop_high,
  chunk_starts = np.cumsum([0] + chunks[:-1])
  chunk_ends   = np.cumsum(chunks)
  # make 1D array designating which chunks are motif
  L = xyz.shape[0]
  mask = torch.zeros(L, L)
  is_motif = torch.zeros(L)
  for i in range(len(chunks)):
    is_motif[chunk_starts[i]:chunk_ends[i]] = chunk_is_motif[i]
  # 2D array designating which chunks can see each other
  for i in range(len(chunks)):
    for j in range(len(chunks)):
      i_is_motif = chunk_is_motif[i]
      j_is_motif = chunk_is_motif[j]
      if (i_is_motif and j_is_motif): # both are motif, so possibly reveal info
        ID_i = motif_ids[i]
        ID_j = motif_ids[j]
        assert ID_i != -1 and ID_j != -1, 'both motif but one has no ID'
        # always reveal self vs self
        if i == j:
          mask[chunk_starts[i]:chunk_ends[i], chunk_starts[j]:chunk_ends[j]] = 1
        else:
          # find out of this (i,j) are allowed to see each other
          ix = tuple(sorted([ID_i, ID_j]))
          can_see = ij_can_see[ij.index(ix)]
          if can_see:
            mask[chunk_starts[i]:chunk_ends[i], chunk_starts[j]:chunk_ends[j]] = 1
  return mask.bool(), is_motif.bool()



def get_chunked_mask(xyz, low_prop, high_prop, max_motif_chunks=8):
  """
  Produces a mask of discontiguous protein chunks that are revealed.
  Also produces a tensor indicating which chunks are given relative geom. info
  Parameters:
  -----------
  xyz (torch.tensor): (L, 14, 3) tensor of atomic coordinates
  low_prop (float): lower bound on proportion of protein that is masked
  high_prop (float): upper bound on proportion of protein that is masked
  """
```

```
    L = xyz.shape[0]
    # decide number of chunks
    n_motif_chunks = random.randint(2, max_motif_chunks)
    # decide what proportion of the protein is masked
    # prop cannot result in n_unmasked < n_motif_chunks --> clamp high prop
    max_prop = 1-(n_motif_chunks+1)/L       # add +1 to be safe
    high_prop = min(high_prop, max_prop)
    prop = random.uniform(low_prop, high_prop)
    n_masked   = int(L * prop)
    n_unmasked = L - n_masked
    # decide the length of each chunk by randomly sampling
    # positions to cut a line with n_chunks - 1 cuts
    cuts    = sorted(random.sample(range(1, n_unmasked), n_motif_chunks - 1))
    lengths = [cuts[0]] + [cuts[i] - cuts[i-1] for i in range(1, len(cuts))] + [n_unmasked -
    cuts[-1]]
    # decide which chunks are given relative geom. info
    # walk over all unique pairs
    motif_pairs = list(itertools.combinations(range(n_motif_chunks), 2))
    # 33% chance that a pair can see each other
    pairs_can_see = [random.choice([True, False, False]) for _ in range(len(motif_pairs))]
    # decide location of chunks within the protein
    # (1) decide order
    random.shuffle(lengths)
    # (2) split available space remaining into other chunks between
    #     the chunks that have been assigned a length
    ngap_low  = len(lengths) - 1
    ngap_high = ngap_low + 1
    ngap = random.randint(ngap_low, ngap_high)
    if ngap == (ngap_low): # there's no cterm/nterm gaps
       Nterm_gap = False
       Cterm_gap = False
    elif ngap == (ngap_low + 1): # there's either a cterm or nterm gap
       Nterm_gap = random.choice([True, False])
       Cterm_gap = not Nterm_gap
    else: # there's both a cterm and nterm gap
       Nterm_gap = True
       Cterm_gap = True
    gaps = sample_gaps(ngap, n_masked) # gaps between unmasked chunks
    random.shuffle(gaps)
    chunks = []
    is_motif = []
    motif_ids = []
    cur_motif_id = 0
    if Nterm_gap:
       chunks.append(gaps.pop())
       is_motif.append(False)
       motif_ids.append(-1)
    for i in range(len(lengths)):
       chunks.append(lengths[i])
       is_motif.append(True)
       motif_ids.append(cur_motif_id)
       cur_motif_id += 1
       if len(gaps) > 1:                       # more to spare
          chunks.append(gaps.pop())
          is_motif.append(False)
```

```
      motif_ids.append(-1)
    elif len(gaps) == 1 and not Cterm_gap:  # only one left, but no Cterm gap
      chunks.append(gaps.pop())
      is_motif.append(False)
      motif_ids.append(-1)
    else:
      pass # no more to spare
  if Cterm_gap:
    assert len(gaps) == 1
    chunks.append(gaps.pop())
    is_motif.append(False)
    motif_ids.append(-1)
  assert sum(chunks) == L, f'chunks sum to {sum(chunks)} but should sum to {L}'
  return chunks, is_motif, motif_ids, motif_pairs, pairs_can_see
def sample_gaps(n, M):
  """
  Samples n chunks that sum to M.
  https://stackoverflow.com/questions/2640053/getting-n-random-numbers-whose-sum-is-m/2640079#26400
  Parameters:
n (int, required): number of chunks
M (int, required): number that the chunk lengths should sum to
  """
  nums = np.random.dirichlet(np.ones(n))*M
  # now round to nearest integer, conserving the total sum
  rounded = []
  round_up = True
  for i in range(len(nums)):
    if round_up:
      rounded.append(np.ceil(nums[i]))
      round_up = False
    else:
      rounded.append(np.floor(nums[i]))
      round_up = True
  # ensure all > 0
  for i in range(len(rounded)):
    if rounded[i] < 1:
      rounded[i] = 1
  while sum(rounded) > M:
    ix = np.random.randint(0, len(rounded))
    if rounded[ix] < 2: # must be at least 1
      continue
    rounded[ix] -= 1
  while sum(rounded) != M:
    ix = np.random.randint(0, len(rounded))
    rounded[ix] += 1
  assert all([x >= 1 for x in rounded])
  return [int(x) for x in rounded]
```

Algorithm 2. Multi triple contact. These functions define how to retrieve a “multi triple contact” motif mask.

```
```

```
def _get_multi_triple_contact_3template(xyz,
                                        low_prop,
                                        high_prop,
                                        max_triples=2,
                                        xyz_less_than=6,
                                        seq_dist_greater_than=10,
                                        len_low=1,
                                        len_high=7,
                                        force_triples=None):
  """
  Gets 2d mask + 1d is motif for multiple triple contacts.
  Parameters:
     xyz (torch.tensor, required): (L, 14, 3) atomic coordinates
     low_prop (float, required): lower bound for proportion of protein to mask
     high_prop (float, required): upper bound for proportion of protein to mask
     max_triples (int, optional): maximum number of triples to find. Default is 2.
is 6. xyz_less_than (int, optional): maximum distance between atoms to consider a contact. Default
contacts. seq_dist_greater_than (int, optional): minimum sequence distance between atoms to consider a
     len_low (int, optional): minimum length of motif chunk. Default is 1.
     len_high (int, optional): maximum length of motif chunk. Default is 7.
None.  force_triples (int, optional): force the number of triples to be this number. Default is
  """
  contacts = get_contacts(xyz, xyz_less_than, seq_dist_greater_than)
  if not contacts.any():
     return _get_diffusion_mask_chunked(xyz, low_prop, high_prop, max_motif_chunks=6)
  is_motif_stack = []
  mask_2d_stack = []
  if force_triples is None:
     n_triples = random.randint(1, max_triples)
  else:
     n_triples = force_triples

  for i in range(n_triples):
     indices = find_third_contact(contacts)

     if (indices is None):
        if i == 0:
           # we found no triples at all, so just return a simple chunked diffusion mask
           return _get_diffusion_mask_chunked(xyz, low_prop, high_prop, max_motif_chunks=6)
        else:
           # we found i triples but couldn't find i+1 --> regenerate and return the i triples
           return _get_multi_triple_contact_3template(xyz, low_prop, high_prop, force_triples=i)
     L = xyz.shape[0]
     # 1d tensor describing which residues are motif
     tmp_is_motif = sample_around_contact(L, indices, len_low, len_high)
     # now get the 2d tensor describing which residues can see each other
     # For these, all motif chunks can see each other
     tmp_mask_2d = tmp_is_motif[:, None] * tmp_is_motif[None, :]
     is_motif_stack.append(tmp_is_motif)
     mask_2d_stack.append(tmp_mask_2d)
  is_motif = torch.stack(is_motif_stack, dim=0).bool()
  mask_2d = torch.stack(mask_2d_stack, dim=0).bool()
  is_motif = torch.any(is_motif, dim=0)
  mask_2d = torch.any(mask_2d, dim=0)
```

```
  return mask_2d, is_motif
def sample_around_contact(L, indices, len_low, len_high):
  """
  Given a list of indices, sample a revealed motif around each index.
  Parameters:
    L (int, required): length of protein
    indices (list, required): list of indices around which to sample motif
    len_low (int, optional): minimum length of motif.
    len_high (int, optional): maximum length of motif.
  """
  diffusion_mask = torch.zeros(L).bool()
  for anchor in indices:
    mask_length = int(np.floor(random.uniform(len_low, len_high)))
    l = anchor - mask_length // 2
    r = anchor + (mask_length - mask_length//2)
    l = max(0, l)
    r = min(r, L)
    diffusion_mask[l:r] = True
  return diffusion_mask
def get_Cb(xyz):
  """
  Compute Cb given N,Ca,C
  Parameters:
    xyz (torch.tensor, required): shape (batch, L, 3) Cartesian coordinates of atoms
  """
  N  = xyz[...,0,:]
  Ca = xyz[...,1,:]
  C  = xyz[...,2,:]
  b = Ca - N
  c = C - Ca
  a = torch.cross(b, c, dim=-1)
  return -0.58273431*a + 0.56802827*b - 0.54067466*c + Ca
def get_pair_dist(a, b):
  """
  calculate pair distances between two sets of points

  Parameters:
    a,b (torch.tensor, required): shape (batch, L, 3) Cartesian coordinates of atoms
  """
  dist = torch.cdist(a, b, p=2)
  return dist
def get_cb_distogram(xyz):
  Cb = get_Cb(xyz)
  dist = get_pair_dist(Cb, Cb)
  return dist
def get_contacts(xyz, xyz_less_than=5, seq_dist_greater_than=10):
  """
  Computes a 2D boolean mask of contacting residues. True if i,j are contacting, False otherwise.
  """
  L = xyz.shape[0]
  dist = get_cb_distogram(xyz)
  is_close_xyz = dist < xyz_less_than
  idx = torch.ones_like(dist).nonzero()
  seq_dist = torch.abs(torch.arange(L)[None] - torch.arange(L)[:,None])
```

```
  is_far_seq = torch.abs(seq_dist) > seq_dist_greater_than
  contacts = is_far_seq * is_close_xyz
  return contacts
def find_third_contact(contacts):
  """
  Finds a third contact for a pair of contacting residues
  Parameters:
    contacts (torch.tensor, required): (L, L) 2d mask of contacts between residues.
  """
  contact_idxs = contacts.nonzero()
  contact_idxs = contact_idxs[torch.randperm(len(contact_idxs))]
  for i,j in contact_idxs:
    if j < i:
      continue
    K = (contacts[i,:] * contacts[j,:]).nonzero()
    if len(K):
      K = K[torch.randperm(len(K))]
      for k in K:
        return torch.tensor([i,j,k])
  return None
```

Algorithm 3. Small molecule contact mask generation. The following function defines how 1D and 2D motif masks were generated for protein-ligand complex examples.

```
def _get_sm_contact_3template(xyz,
                  is_sm,
                  contact_cut=8,
                  chunk_size_min=1,
                  chunk_size_max=7,
                  min_seq_dist=9):
  """
reveProduces mask2d and is_motif for small molecule, possibly with contacting protein chunks
  Parameters:
    xyz (torch.Tensor): 3D coordinates of protein and small molecule (L, 14, 3)
    is_sm (torch.Tensor): binary tensor indicating which tokens are small molecule atoms (L,)
    contact_cut (float): distance cutoff for contact between protein and small molecule
    chunk_size_min (int): minimum size of revealed chunk
    chunk_size_max (int): maximum size of revealed chunk
    min_seq_dist (int): minimum sequence distance between revealed chunks
  """
  assert len(xyz.shape) == 3
  ca = xyz[~is_sm, 1,:]
  if ca.shape[0] == 0:
    sm_only = True
  else:
    sm_only = False
  sm_xyz  = xyz[is_sm, 1,:]
  dmap = torch.cdist(ca, sm_xyz)
  dmap = dmap < contact_cut
  protein_is_contacting = dmap.any(dim=-1) # which CA's are contacting sm
  where_is_contacting = protein_is_contacting.nonzero().squeeze()
```

```
  n_chunk_revealed = random.randint(0,4)
  if (n_chunk_revealed == 0) or (sm_only):
     is_motif = is_sm.clone()
     is_motif_2d = is_motif[:, None] * is_motif[None, :]
     return is_motif_2d, is_motif
  else:
     is_motif = is_sm.clone()
     cur_min_seq_dist = min_seq_dist # could possibly increment this if needed
     for i in range(n_chunk_revealed):
        chunk_size = torch.randint(chunk_size_min, chunk_size_max, size=(1,)).item()
        if len(where_is_contacting.shape) == 0:
           # ensures where_is_contacting is a 1d tensor
           where_is_contacting = where_is_contacting.unsqueeze(0)
        if (where_is_contacting.shape[0] == 0) and (i == 0):
           # no contacts, so sm is only motif
           is_motif = is_sm.clone()
           is_motif_2d = is_motif[:, None] * is_motif[None, :]
           return is_motif_2d, is_motif
        p = torch.ones_like(where_is_contacting)/len(where_is_contacting)
        chosen_idx = p.multinomial(num_samples=1, replacement=False)
        chosen_idx = chosen_idx.item()
        chosen_idx = where_is_contacting[chosen_idx]
        # find min and max indices for revealed chunk
        min_index = max(0, chosen_idx - chunk_size//2)
        max_index = min(protein_is_contacting.numel(), 1+chosen_idx + chunk_size//2)
        # reveal chunk
        is_motif[min_index:max_index] = True
        # update where_is_contacting
        start = max(0,min_index-cur_min_seq_dist)
        end = min(protein_is_contacting.numel(), max_index+cur_min_seq_dist)
        protein_is_contacting[start:end] = False # remove this option from where_is_contacting
        where_is_contacting = protein_is_contacting.nonzero().squeeze()
        if protein_is_contacting.sum() == 0:
           break # can't make any more chunks
     is_motif_2d = is_motif[:, None] * is_motif[None, :] # all tokens can "see" each other
     return is_motif_2d, is_motif
```

Sequence information

>**n8**
EEERFRAYYERYFAALAARDYETLLEILREFGVAKLVLNGREFASPEEAVQWARDTGLRFLRLIRE
RYEDGVYTVEDIVSRDDGRYYRHTSTLRRQPDGSYVQYSQLELL
>**shh_25**
LSEEEVEKAIKEIKEKFEKLAKEIEKLVAEGADRDTLVERLVKYAASLGAASASPETSTPEQYAAVR
EIFRALADAILDGRWEEAGERLVEATVRHTQALIDAARAAGRDELVPALRRLGLALAEAIFDILRE
YIAKKTGDAAAAERYKETYLARLRAAL
>**rem6507**
MSEEEVERRIREIHAMFEKLEEKIDELVAAGADVETLTSELVRFAASLGADSANPEKSTPEQYEAV
KEILRALAEAIINGKWEEAGKTLVDATVRHTGALIERARAAGREDEVPALRRLGEALLRSTLEILR
KYIEKKTGDKELADKYFETYLAEFRKKL
>**win**

MTEEELDQAVEDFLRVHSELVHRLAGDPPDELFQRLDRFVTDAIIEGNPERRDEIKADLARAARVF
GEALERDITTPEDFNAFLRELGPEAVELVSTFTQQFVDVIRGDPQAVAEHLNISLEDVARLAEAGEA
AIERGEGASLGVHRELRRIEARRNS

>**super**

SENEKLVEKVLEATRRIAREEAVKYKDAFLRAYRARDGAGLRRVITGLFSKVDSRLYKEVLTDVP
TIVALQRRAGVDITPEQAQEILDNYDNEKHTAAVMDETFALLARAAATQASYEELLAAAPSGSVI
LALEVLRVLLEINNLSWREVLPLLALAAAS

>**win1**

LSDRELEASLQAFFEVHTRLVHRLAGIEPDPRFEILDKYIFRQIVADNPEEREKIRLDYGRAAEIFRD
ALARDITTPEAFNAYLEALGPDAVRTVQDLTRRFVDVIRADPEAIAKLLNISKEDVQGLARAGEA
AIERGEGASLGVLRELRKIEKKRNE

>**win11**

MTQAELDDGVTRFIQVHNELVHRLAGVEPDERFIKLDVYVTNEIVNSDPTKAAERKQALARAAD
VFSEALARDIRTAEEFNAFLEELGPEAVELVADLTRAWVEVIESDPEKIAQLLNISVEEVEELAEAG
RRAIEEGRGASIGVLLKLREIEKIRSS

>**win31**

MTEEELEKGVEDFLVVHGKFVHRLAGIPPNAKFQALDKYITNQIVESDPSKEKEIKKAFGDAAKIL
RDALARNITTPEEAQAFLRDLGPWAVDLINTITRRYVDVIEKNPEGVAEILGISLEEVRELAEAGRR
AIEEGEGASLGILRKILELEAERAK

>**dadt1**

MTEEEKAQAVEDFHKYHTALVVRLAGVPPHERYERLDRWITEQLIYGDESKRDEWIDALAEAAG
IFKEALERNITTPEEFNAFLKEKGKRAVDLVATLTNAYIAVLEGDPEAIARFLGISLEEVQEIIDAAR
KAVKEGRGASIGIYTKLRELEERRAA

>**charliet2**

AAEAKARWKKAVEMFKKLSDKLLEAGAGVSPAFAVSIAYAAGIITEEQVYSTIEGVIAKVKADPE
RWQKEVAELFTTYKDDLDAFVEKHTELVKRLLGDSLPDEVFELFKEKDKEVLEKIPKELWEKVTS
LFEEGNYEEANKLIQKAYREYALEILEGEYNSL

>**kent1**

AEEEAVRAVEERDDEALGRALFGAADAGDLDGATRAHTAAVLKGRELGIDAGSLAIYDVVAEYV
RTGKKPPKEEAEKLVLGYAKRDVERRADGTPLERVLAEATLVFAKTFLDNYDEIFAELEKDPSLTP
DEAFAVADRFLWDVMAARGEWDKLLAAYRAAVG

>**win1_b1**

MTEAERQELLDRAVSAHQRFVFDRLGVERDPRYEALDRFTVDIIVGENKERRDEVYRDLALLRE
GIELFYKQNIPMKEITKEDIEKLPKEMYDAFTRLTDYFIEVVEERKEEIAKLLGFTLEEIDRIVEAGK
KAIEEGQGASIGVATAIREIAAEREA

>**momi**

SFSARRERVVVVGYSLGVFAGIIMFAANTTYEEALKLAKEIFKEALKNPELIARHLALHRNAETTG
KEEWRQDIETWIKIIEKRLGKPVDKSRIFIATTKEEAVRLAEEAVKLGRAVIYTPPHLLPVAGDEIVE
VLTKAGVTVLVKGGIGSGVPLKVYEA

>**superfast**

SEEEELINKIEEETKKLAKEVAEKYKDAFKAAFDAEDGKGLRDVLTEAWREVDIKIWKKFFLNPD
DPEGGKRNYALVTKVNRERAKELGIEEVAERVEKAISSAKNLEEIVDNHTEIVLELSFAQLVEAAK
EKADFDWIYENAPSISSRLFFIVLKELVEINNLSWVDVLPLLARAATA

>**supercool**

EKDKELHKKVFDFVDKVAREVVAGYVDALRAAVAARDGAGVRAVLTEAFREVDRELHKFLTSEE
NKELFDYVIEYNRRIYLEQNPELSELPEKYKELEKVIRDPDIDKYLEVHVPLVMEQAFGLLAEAAR
RRWDWDEVYAAAPSFSVRAFLRVLRVLLEQSDLDWVSTLPLIAEAATA

>**momi120**

MIKEYEFPAKKAKTVEEAEKNVEEEIELMKSSGVSFSARRERVVVVGYSLGVFAGIIMFAANTTY
EEALKLAKEIFKEALKNPELIARHLALHRNAETTGKEEWRQDIETWIKIIEKRLGKPVDKSRIFIAT
TKEEAVRLAEEAVKLGRAVIYTPPHLLPVAGDEIVEVLTKAGVTVLVKGGIGSGVPLKVYEA

>**momi120-103**

PVRRHRFPARKANNFEEAVANVERLIEEIRAAGVDFSARKERAVVVGYSLGVVTGMIMFATGTDF
IEALRKALEIGKKVVEEDPEFMERHRKIVTDGNRAEIREDIDYWIEVIEKETGHPVDRSRIFIAETV
EEAVELARRAVELGHAIIVLPPYLIGEAGEAVVEVLTAAGVDVLLMGGLGSGYPVTIYQA

>**momi120-74**

EVKVTTFPARKATTDEEAKANVEQLIEDIEKSGVTFSADVEKFVVIGYSLGIVTGMLMMYKKTNF
VEALREAMEIFDEIRADPANAPFLERHRRIHENGTKEEILEDIKFWIDHIEKKYGIPFDRSRIFIALTK
EEAVELAKKAVEYGRAIIYLPPSLIEEAGDEVVEVLANAGVQVLMMGGLGSGYPVRIYEA

Chapter 4 Supplement

Supplementary Figures

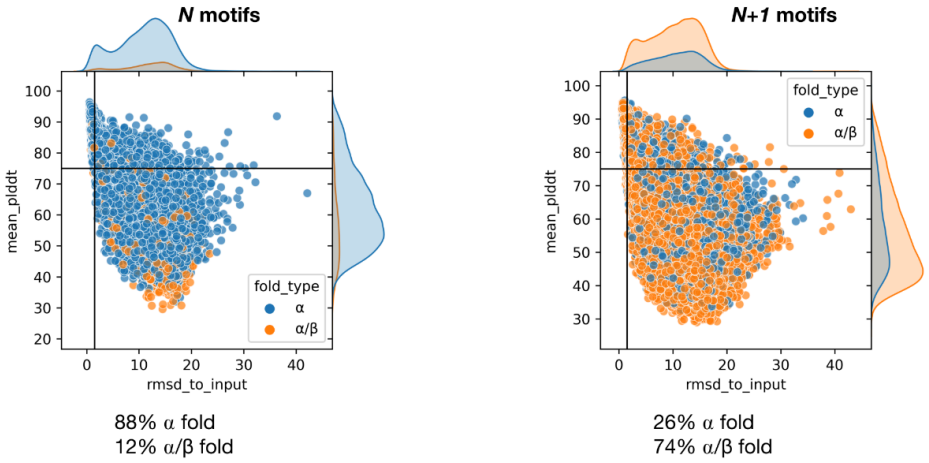

Figure S1. *N+1* motifs yield more complex folds. Designs generated by RFdiffusion using *N* motif theozymes were overwhelmingly composed of solely alpha-helical content. Utilization of the *N+1* oxyanion hole motif significantly increased the number of backbones with more complex, mixed alpha/beta secondary structures.

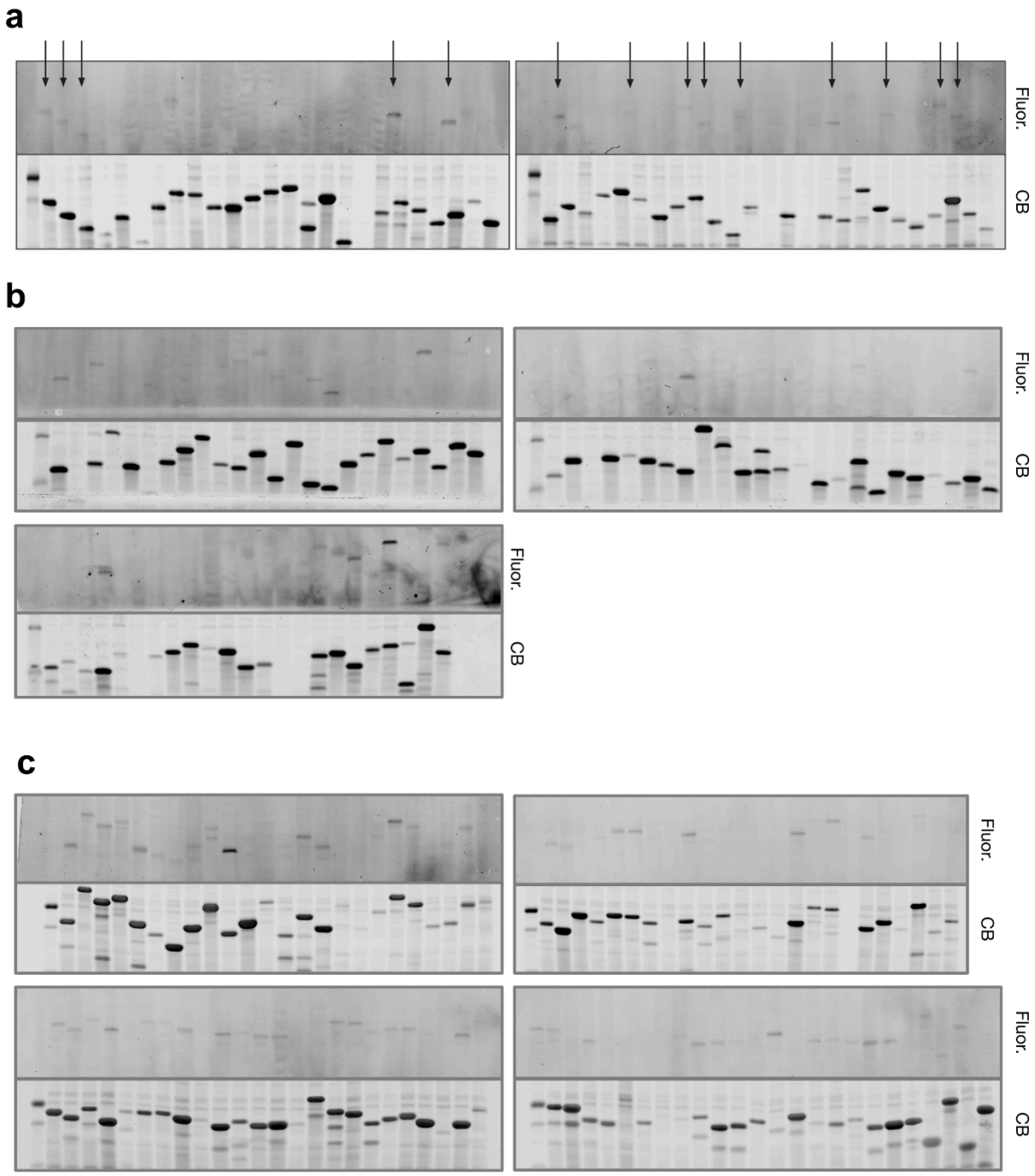

Figure S2. In-gel fluorescence imaging with fluorescently labeled fluorophosphonate activity-based probe. In-gel fluorescence of rounds (**A**) 1, (**B**) 2, and (**C**) 3 designs after 1 hour incubation of cell lysate with 1 μM PET-FP-TAMRA. Fluor. stands for fluorescence and CB for Coomassie blue. Lanes with labels indicate designs that were purified and tested for esterase activity.

Figure S3. Redesign of momi120 for PET enhances esterolysis (**A**) Reaction progress curves of momi120 redesigns incubated with 4MU-PhAc. (**B**) Purified designs incubated with dp0460. Three showed activity above background hydrolysis.

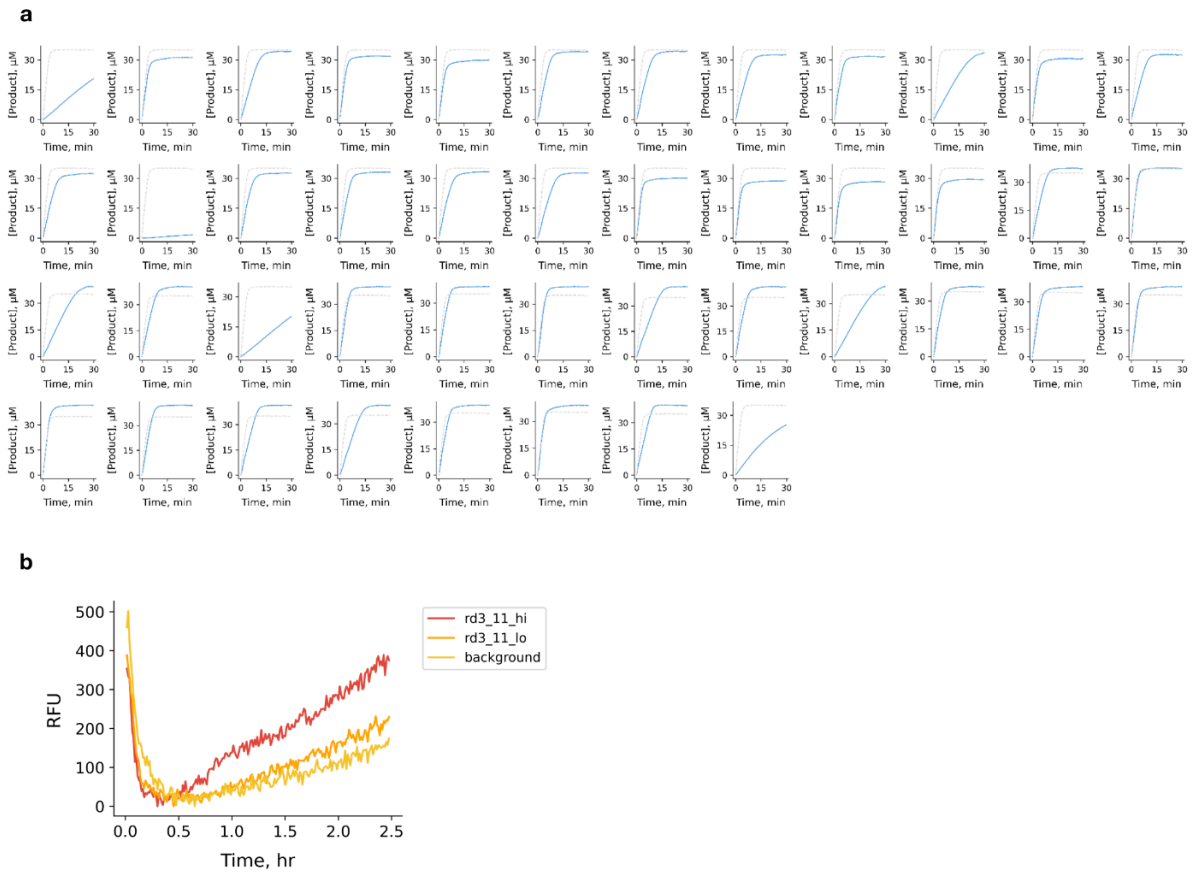

Figure S4. Redesign of momi120-103 for 4MU-PhAc enhances activity on PET. (**A**) Reaction progress curves of momi120-103 redesigns incubated with 4MU-PhAc. (**B**) Purified **rd3_11** incubated with dp0460 at two different enzyme concentrations (hi = 5 µM, lo = 1 µM).

References

1. C. C. Garcia, J. G. Potian, K. Hognason, B. Thyagarajan, L. G. Sultatos, N. Souayah, V. H. Routh, J. J. McArdle, Acetylcholinesterase deficiency contributes to neuromuscular junction dysfunction in type 1 diabetic neuropathy. *Am. J. Physiol. Endocrinol. Metab.* **303**, E551–61 (2012).

2. W. P. Jencks, *Catalysis in Chemistry and Enzymology* (Dover Publications, Mineola, NY, 1987).

3. A. J. Hutt, The development of single-isomer molecules: why and how. *CNS Spectr.* **7**, 14–22 (2002).

4. J. H. Kim, A. R. Scialli, Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicol. Sci.* **122**, 1–6 (2011).

5. G. Blaschke, H. P. Kraft, K. Fickentscher, F. Köhler, Chromatographic separation of racemic thalidomide and teratogenic activity of its enantiomers (author's transl). *Arzneimittelforschung* **29**, 1640–1642 (1979).

6. R. E. S. Thomson, S. E. Carrera-Pacheco, E. M. J. Gillam, Engineering functional thermostable proteins using ancestral sequence reconstruction. *J. Biol. Chem.* **298**, 102435 (2022).

7. C. N. Pace, G. R. Grimsley, J. A. Thomson, B. J. Barnett, Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.* **263**, 11820–11825 (1988).

8. C. E. Correnti, M. M. Gewe, C. Mehlin, A. D. Bandaranayake, W. A. Johnsen, P. B. Rupert, M.-Y. Brusniak, M. Clarke, S. E. Burke, W. De Van Der Schueren, K. Pilat, S. M. Turnbaugh, D. May, A. Watson, M. K. Chan, C. D. Bahl, J. M. Olson, R. K. Strong, Screening, large-scale production and structure-based classification of cystine-dense peptides. *Nat. Struct. Mol. Biol.* **25**, 270–278 (2018).

9. M. Zavodszky, C. W. Chen, J. K. Huang, M. Zolkiewski, L. Wen, R. Krishnamoorthi, Disulfide bond effects on protein stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Sci.* **10**, 149–160 (2001).

10. 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* **70**, 380 (2018).

11. F. H. Arnold, Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed Engl.* **57**, 4143–4148 (2018).

12. E. Bell, R. Smithson, S. Kilbride, J. Foster, F. Hardy, S. Ramachandran, A. Tedstone, S. Haigh, A. Garforth, P. Day, C. Levy, M. Shaver, A. Green, Directed evolution of an efficient and thermostable PET depolymerase, *Research Square* (2022). https://doi.org/10.21203/rs.3.rs-1350765/v1.

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

14. Crystallography: Protein data bank. *Nat. New Biol.* **233**, 223–223 (1971).

15. 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. *Nature* **596**, 583–589 (2021).

16. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. 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. C. 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* **373**, 871–876 (2021).

17. 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* **378**, 49–56 (2022).

18. 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. *Nature* **620**, 1089–1100 (2023).

19. S. Vázquez Torres, P. J. Y. Leung, P. Venkatesh, I. D. Lutz, F. Hink, H.-H. Huynh, J. Becker, A. H.-W. Yeh, D. Juergens, N. R. Bennett, A. N. Hoofnagle, E. Huang, M. J. MacCoss, M. Expòsit, G. R. Lee, A. K. Bera, A. Kang, J. De La Cruz, P. M. Levine, X. Li, M. Lamb, S. R. Gerben, A. Murray, P. Heine, E. N. Korkmaz, J. Nivala, L. Stewart, J. L. Watson, J. M. Rogers, D. Baker, De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).

20. M. Glögl, A. Krishnakumar, R. J. Ragotte, I. Goreshnik, B. Coventry, A. K. Bera, A. Kang, E. Joyce, G. Ahn, B. Huang, W. Yang, W. Chen, M. G. Sanchez, B. Koepnick, D. Baker, Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists. *Science* **386**, 1154–1161 (2024).

21. L. Cao, B. Coventry, I. Goreshnik, B. Huang, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, S. Halabiya, B. Hammerson, W. Yang, S. Benard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Robust de novo design of protein binding proteins from target structural information alone, *bioRxiv* (2021). https://doi.org/10.1101/2021.09.04.459002.

22. 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. *Nature* **453**, 190–195 (2008).

23. 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 3rd, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).

24. J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, D. Baker, Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309–313 (2010).

25. H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, S. L. Mayo, Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790–3795 (2012).

26. F. Richter, R. Blomberg, S. D. Khare, G. Kiss, A. P. Kuzin, A. J. T. Smith, J. Gallaher, Z. Pianowski, R. C. Helgeson, A. Grjasnow, R. Xiao, J. Seetharaman, M. Su, S. Vorobiev, S. Lew, F. Forouhar, G. J. Kornhaber, J. F. Hunt, G. T. Montelione, L. Tong, K. N. Houk, D. Hilvert, D. Baker, Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).

27. S. Rajagopalan, C. Wang, K. Yu, A. P. Kuzin, F. Richter, S. Lew, A. E. Miklos, M. L. Matthews, J. Seetharaman, M. Su, J. F. Hunt, B. F. Cravatt, D. Baker, Design of activated serine-containing catalytic triads with atomic-level accuracy. *Nat. Chem. Biol.* **10**, 386–391 (2014).

28. A. H.-W. Yeh, C. Norn, Y. Kipnis, D. Tischer, S. J. Pellock, D. Evans, P. Ma, G. R. Lee, J. Z. Zhang,

I. Anishchenko, B. Coventry, L. Cao, J. Dauparas, S. Halabiya, M. DeWitt, L. Carter, K. N. Houk, D. Baker, De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).

29. B. M. Beadle, B. K. Shoichet, Structural Bases of Stability–function Tradeoffs in Enzymes. *J. Mol. Biol.* **321**, 285–296 (2002).

30. T. J. Magliery, Protein stability: computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* **33**, 161–168 (2015).

31. A. Singh, V. Upadhyay, A. K. Upadhyay, S. M. Singh, A. K. Panda, Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microb. Cell Fact.* **14**, 41 (2015).

32. N. Rathore, R. S. Rajan, Current perspectives on stability of protein drug products during formulation, fill and finish operations. *Biotechnol. Prog.* **24**, 504–514 (2008).

33. R. E. Cobb, R. Chao, H. Zhao, Directed Evolution: Past, Present and Future. *AIChE J.* **59**, 1432–1440 (2013).

34. 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).

35. 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. *Nature* **604**, 662–667 (2022).

36. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, doi: 10.1038/s41587-022-01618-2 (2023).

37. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8852–8858 (2019).

38. 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* **378**, 56–61 (2022).

39. G. A. Ordway, D. J. Garry, Myoglobin: an essential hemoprotein in striated muscle. *J. Exp. Biol.* **207**, 3441–3446 (2004).

40. C. W. Hamm, Cardiac biomarkers for rapid evaluation of chest pain. *Circulation* **104**, 1454–1456 (2001).

41. M. Bordeaux, V. Tyagi, R. Fasan, Highly diastereoselective and enantioselective olefin cyclopropanation using engineered myoglobin-based catalysts. *Angew. Chem. Int. Ed Engl.* **54**, 1744–1748 (2015).

42. D. M. Carminati, J. Decaens, S. Couve-Bonnaire, P. Jubault, R. Fasan, Biocatalytic strategy for the highly stereoselective synthesis of CHF2 -containing trisubstituted cyclopropanes. *Angew. Chem. Int. Ed Engl.* **60**, 7072–7076 (2021).

43. O. F. Brandenberg, R. Fasan, F. H. Arnold, Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **47**, 102–111 (2017).

44. R. Simsa, J. Yuen, A. Stout, N. Rubio, P. Fogelstrand, D. L. Kaplan, Extracellular heme proteins influence bovine myosatellite cell proliferation and the color of cell-based meat. *Foods* **8**, 521 (2019).

45. J. Devaere, A. De Winne, L. Dewulf, I. Fraeye, I. Šoljić, E. Lauwers, A. de Jong, H. Sanctorum, Improving the aromatic profile of plant-based meat alternatives: Effect of myoglobin addition on volatiles. *Foods* **11**, 1985 (2022).

46. E. J. Moore, D. Zorine, W. A. Hansen, S. D. Khare, R. Fasan, Enzyme stabilization via computationally guided protein stapling. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 12472–12477 (2017).

47. J. A. Iannuzzelli, J.-P. Bacik, E. J. Moore, Z. Shen, E. M. Irving, D. A. Vargas, S. D. Khare, N. Ando, R. Fasan, Tuning enzyme thermostability via computationally guided covalent stapling and structural basis of enhanced stabilization. *Biochemistry* **61**, 1041–1054 (2022).

48. S. R. Hubbard, W. A. Hendrickson, D. G. Lambright, S. G. Boxer, X-ray crystal structure of a recombinant human myoglobin mutant at 2·8 Å resolution. *J. Mol. Biol.* **213**, 215–218 (1990).

49. A. Keppner, D. Maric, M. Correia, T. W. Koay, I. M. C. Orlando, S. N. Vinogradov, D. Hoogewijs, Lessons from the post-genomic era: Globin diversity beyond oxygen binding and transport. *Redox Biol* **37**, 101687 (2020).

50. O. H. Kapp, L. Moens, J. Vanfleteren, C. N. Trotman, T. Suzuki, S. N. Vinogradov, Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. *Protein Sci.* **4**, 2179–2190 (1995).

51. D. A. Gell, Structure and function of haemoglobins. *Blood Cells Mol. Dis.* **70**, 13–42 (2018).

52. 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* **377**, 387–394 (2022).

53. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

54. P. G. Blommel, B. G. Fox, A combined approach to improving large-scale production of tobacco etch virus protease. *Protein Expr. Purif.* **55**, 53–68 (2007).

55. J. Phan, A. Zdanov, A. G. Evdokimov, J. E. Tropea, H. K. Peters 3rd, R. B. Kapust, M. Li, A. Wlodawer, D. S. Waugh, Structural basis for the substrate specificity of tobacco etch virus protease. *J. Biol. Chem.* **277**, 50564–50572 (2002).

56. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).

57. J. Breidenbach, U. Bartz, M. Gütschow, Coumarin as a structural component of substrates and probes for serine and cysteine proteases. *Biochim. Biophys. Acta: Proteins Proteomics* **1868**, 140445 (2020).

58. R. B. Kapust, J. Tözsér, J. D. Fox, D. E. Anderson, S. Cherry, T. D. Copeland, D. S. Waugh, Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng.* **14**, 993–1000 (2001).

59. M. I. Sanchez, A. Y. Ting, Directed evolution improves the catalytic efficiency of TEV protease. *Nat. Methods* **17**, 167–174 (2020).

60. R. Otten, R. A. P. Pádua, H. A. Bunzel, V. Nguyen, W. Pitsawong, M. Patterson, S. Sui, S. L. Perry, A. E. Cohen, D. Hilvert, D. Kern, How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).

61. G. Jiménez-Osés, S. Osuna, X. Gao, M. R. Sawaya, L. Gilson, S. J. Collier, G. W. Huisman, T. O. Yeates, Y. Tang, K. N. Houk, The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **10**, 431–436 (2014).

62. T. Ishida, Effects of point mutation on enzymatic activity: correlation between protein electronic structure and motion in chorismate mutase reaction. *J. Am. Chem. Soc.* **132**, 7104–7118 (2010).

63. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).

64. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019).

65. E. A. Berry, B. L. Trumpower, Simultaneous determination of hemes a, b, and c from pyridine hemochrome spectra. *Anal. Biochem.* **161**, 1–15 (1987).

66. M. Reichlin, Enzyme Proteins: Hemoglobin and Myoglobin in Their Reactions with Ligands . Eraldo Antonini and Maurizio Brunori. North-Holland, Amsterdam, 1971 (U.S. distributor, Elsevier, New York). xx, 436 pp., illus. $30. Frontiers of Biology, vol. 21. *Science* **178**, 296–296 (1972).

67. D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, Amber 2020: University of california. *San Franc.*

68. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

69. C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).

70. M. J. Frish, J. W. Trucks, H. B. Schlegel, G. E. Scuseria, Gaussian 16, Revision C. 01; Gaussian. *Inc.: Wallingford, CT, USA*.

71. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

72. H. C. Andersen, Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* (1980).

73. T. A. Andrea, W. C. Swope, H. C. Andersen, The role of long ranged forces in determining the structure and properties of liquid water. *J. Chem. Phys.* **79**, 4576–4584 (1983).

74. S. Miyamoto, P. A. Kollman, Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).

75. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N · log(N) method for Ewald sums in large

systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

76. S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).

77. R. V. Rakotoharisoa, B. Seifinoferest, N. Zarifi, J. D. M. Miller, J. M. Rodriguez, M. C. Thompson, R. A. Chica, Design of efficient artificial enzymes using crystallographically enhanced conformational sampling. *J. Am. Chem. Soc.* **146**, 10001–10013 (2024).

78. A. Broom, R. V. Rakotoharisoa, M. C. Thompson, N. Zarifi, E. Nguyen, N. Mukhametzhanov, L. Liu, J. S. Fraser, R. A. Chica, Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).

79. G. Kiss, D. Röthlisberger, D. Baker, K. N. Houk, Evaluation and ranking of enzyme designs. *Protein Sci.* **19**, 1760–1773 (2010).

80. S. J. Fleishman, S. D. Khare, N. Koga, D. Baker, Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **20**, 753–757 (2011).

81. H. A. Bunzel, J. L. R. Anderson, D. Hilvert, V. L. Arcus, M. W. van der Kamp, A. J. Mulholland, Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **13**, 1017–1022 (2021).

82. M. P. Frushicheva, J. Cao, Z. T. Chu, A. Warshel, Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16869–16874 (2010).

83. A. J. Burton, A. R. Thomson, W. M. Dawson, R. L. Brady, D. N. Woolfson, Installing hydrolytic activity into a completely de novo protein framework. *Nat. Chem.* **8**, 837–844 (2016).

84. D. N. Bolon, S. L. Mayo, Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14274–14279 (2001).

85. A. J. Burke, S. L. Lovelock, A. Frese, R. Crawshaw, M. Ortmayer, M. Dunstan, C. Levy, A. P. Green, Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* **570**, 219–223 (2019).

86. S. Studer, D. A. Hansen, Z. L. Pianowski, P. R. E. Mittl, A. Debon, S. L. Guffy, B. S. Der, B. Kuhlman, D. Hilvert, Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science* **362**, 1285–1288 (2018).

87. B. S. Der, D. R. Edwards, B. Kuhlman, Catalysis by a de novo zinc-mediated protein interface: implications for natural enzyme evolution and rational enzyme engineering. *Biochemistry* **51**, 3933–3940 (2012).

88. Y. S. Moroz, T. T. Dunston, O. V. Makhlynets, O. V. Moroz, Y. Wu, J. H. Yoon, A. B. Olsen, J. M. McLaughlin, K. L. Mack, P. M. Gosavi, N. A. J. van Nuland, I. V. Korendovych, New tricks for old proteins: Single mutations in a nonenzymatic protein give rise to various enzymatic activities. *J. Am. Chem. Soc.* **137**, 14905–14911 (2015).

89. V. Tournier, C. M. Topham, A. Gilles, B. David, C. Folgoas, E. Moya-Leclair, E. Kamionka, M.-L. Desrousseaux, H. Texier, S. Gavalda, M. Cot, E. Guémard, M. Dalibey, J. Nomme, G. Cioci, S. Barbe, M. Chateau, I. André, S. Duquesne, A. Marty, An engineered PET depolymerase to break

down and recycle plastic bottles. *Nature* **580**, 216–219 (2020).

90. 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). [Preprint] (2016). https://doi.org/10.1126/science.aad6359.

91. D. M. Blow, Structure and mechanism of chymotrypsin. *Acc. Chem. Res.* **9**, 145–152 (1976).

92. P. Carter, J. A. Wells, Dissecting the catalytic triad of a serine protease. *Nature* **332**, 564–568 (1988).

93. P. Carter, J. A. Wells, Functional interaction among catalytic residues in subtilisin BPN'. *Proteins* **7**, 335–342 (1990).

94. L. Polgár, The catalytic triad of serine peptidases. *Cell. Mol. Life Sci.* **62**, 2161–2172 (2005).

95. P. Bryan, M. W. Pantoliano, S. G. Quill, H. Y. Hsiao, T. Poulos, Site-directed mutagenesis and the role of the oxyanion hole in subtilisin. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3743–3745 (1986).

96. D. R. Corey, C. S. Craik, An investigation into the minimum requirements for peptide hydrolysis by mutation of the catalytic triad of trypsin. *J. Am. Chem. Soc.* **114**, 1784–1790 (1992).

97. B. Zerner, R. P. M. Bond, M. L. Bender, Kinetic Evidence for the Formation of Acyl-Enzyme Intermediates in the α-Chymotrypsin-Catalyzed Hydrolyses of Specific Substrates. *J. Am. Chem. Soc.* **86**, 3674–3679 (1964).

98. J. Kraut, Serine proteases: structure and mechanism of catalysis. *Annu. Rev. Biochem.* **46**, 331–358 (1977).

99. L. Hedstrom, Serine protease mechanism and specificity. *Chem. Rev.* **102**, 4501–4524 (2002).

100. A. J. T. Smith, R. Müller, M. D. Toscano, P. Kast, H. W. Hellinga, D. Hilvert, K. N. Houk, Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J. Am. Chem. Soc.* **130**, 15361–15373 (2008).

101. S. Du, R. C. Kretsch, J. Parres-Gold, E. Pieri, V. W. D. Cruzeiro, M. Zhu, M. M. Pinney, F. Yabukarski, J. P. Schwans, T. J. Martínez, D. Herschlag, Conformational ensembles reveal the origins of serine protease catalysis. *Science* **387**, eado5068 (2025).

102. E. S. Radisky, J. M. Lee, C.-J. K. Lu, D. E. Koshland Jr, Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6835–6840 (2006).

103. F. Praetorius, P. J. Y. Leung, M. H. Tessmer, A. Broerman, C. Demakis, A. F. Dishman, A. Pillai, A. Idris, D. Juergens, J. Dauparas, X. Li, P. M. Levine, M. Lamb, R. K. Ballard, S. R. Gerben, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, B. F. Volkman, J. Nivala, S. Stoll, D. Baker, Design of stimulus-responsive two-state hinge proteins. *Science* **381**, 754–760 (2023).

104. A. Pillai, A. Idris, A. Philomin, C. Weidle, R. Skotheim, P. J. Y. Leung, A. Broerman, C. Demakis, A. J. Borst, F. Praetorius, D. Baker, De novo design of allosterically switchable protein assemblies. *Nature* **632**, 911–920 (2024).

105. M. L. Zastrow, A. F. A. Peacock, J. A. Stuckey, V. L. Pecoraro, Hydrolytic catalysis and structural

stabilization in a designed metalloprotein. *Nat. Chem.* **4**, 118–123 (2011).

106. 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. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).

107. 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). https://doi.org/10.1101/2024.08.02.606416.

108. 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. *bioRxivorg*, doi: 10.1101/2024.09.25.614868 (2024).

109. P. A. Frey, A. D. Hegeman, *Enzymatic Reaction Mechanisms* (Oxford University Press, 2007).

110. C. Walsh, *Enzymatic Reaction Mechanisms* (W. H. Freeman, 1979).

111. W. P. Jencks, *Catalysis in Chemistry and Enzymology* (Courier Corporation, 1987).

112. J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, D. Baker, Atomic context-conditioned protein sequence design using LigandMPNN, *bioRxiv* (2023)p. 2023.12.22.573103.

113. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).

114. 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.* **14**, 2625 (2023).

115. S. Du, R. C. Kretsch, J. Parres-Gold, E. Pieri, V. W. D. Cruzeiro, M. Zhu, M. M. Pinney, F. Yabukarski, J. P. Schwans, T. J. Martínez, D. Herschlag, Conformational Ensembles Reveal the Origins of Serine Protease Catalysis, *bioRxiv* (2024)p. 2024.02.28.582624.

116. L. Simón, J. M. Goodman, Hydrogen-bond stabilization in oxyanion holes: grand jeté to three dimensions. *Org. Biomol. Chem.* **10**, 1905–1913 (2012).

117. L. Simón, J. M. Goodman, Enzyme catalysis by hydrogen bonds: the balance between transition state binding and substrate binding in oxyanion holes. *J. Org. Chem.* **75**, 1831–1840 (2010).

118. A. R. Buller, C. A. Townsend, Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E653–61 (2013).

119. E. Zakharova, M. P. Horvath, D. P. Goldenberg, Structure of a serine protease poised to resynthesize a peptide bond. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11034–11039 (2009).

120. G. Dodson, A. Wlodawer, Catalytic triads and their relatives. *Trends Biochem. Sci.* **23**, 347–352 (1998).

121. R. Wolfenden, Y. Yuan, The “neutral” hydrolysis of simple carboxylic esters in water and the rate enhancements produced by acetylcholinesterase and other carboxylic acid esterases. *J. Am. Chem.*

Soc. **133**, 13821–13823 (2011).

122. A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, R. Milo, The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).

123. F. J. Kezdy, M. L. Bender, The kinetics of the α-chymotrypsin-catalyzed hydrolysis of p-nitrophenyl acetate*. *Biochemistry* **1**, 1097–1106 (1962).

124. L. Polgar, M. L. Bender, The reactivity of thiol-subtilisin, an enzyme containing a synthetic functional group. *Biochemistry* **6**, 610–620 (1967).

125. J. L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, I. Silman, Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein. *Science* **253**, 872–879 (1991).

126. A. Warshel, S. Russell, Theoretical correlation of structure and energetics in the catalytic reaction of trypsin. *J. Am. Chem. Soc.* **108**, 6569–6579 (1986).

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

128. M. D. Tyka, D. A. Keedy, I. André, F. Dimaio, Y. Song, D. C. Richardson, J. S. Richardson, D. Baker, Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).

129. F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic, D. Baker, De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).

130. D. Kim, S. M. Woodbury, W. Ahern, I. Kalvet, N. Hanikel, S. Salike, S. J. Pellock, A. Lauko, D. Hilvert, D. Baker, Computational Design of Metallohydrolases, *bioRxiv* (2024). https://doi.org/10.1101/2024.11.13.623507.

131. F. W. Studier, Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

132. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019).

133. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

134. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861–877 (2019).

135. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

136. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, M.

Steinegger, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

137. Plastic Overshoot Day 2024: Global waste crisis surpasses management capacity, *SAFE - Safe Food Advocacy Europe* (2024). https://www.safefoodadvocacy.eu/plastic-overshoot-day-2024-global-waste-crisis-surpasses-management-capacity/.

138. A. J. Nihart, M. A. Garcia, E. El Hayek, R. Liu, M. Olewine, J. D. Kingston, E. F. Castillo, R. R. Gullapalli, T. Howard, B. Bleske, J. Scott, J. Gonzalez-Estrella, J. M. Gross, M. Spilde, N. L. Adolphi, D. F. Gallego, H. S. Jarrell, G. Dvorscak, M. E. Zuluaga-Ruiz, A. B. West, M. J. Campen, Bioaccumulation of microplastics in decedent human brains. *Nat. Med.* **31**, 1114–1119 (2025).

139. G. G. N. Thushari, J. D. M. Senevirathna, Plastic pollution in the marine environment. *Heliyon* **6**, e04709 (2020).

140. Z. O. G. Schyns, M. P. Shaver, Mechanical recycling of packaging plastics: A review. *Macromol. Rapid Commun.* **42**, e2000415 (2021).

141. H. Chen, P. Cebe, Vitrification and devitrification of rigid amorphous fraction of PET during quasi-isothermal cooling and heating. *Macromolecules* **42**, 288–292 (2009).

142. CARBIOS celebrates groundbreaking of first plant, *CARBIOS celebrates groundbreaking of first plant* (2024). https://www.carbios.com/newsroom/en/carbios-celebrates-the-groundbreaking-of-its-pet-biorecycling-plant/.

143. E. Erickson, J. E. Gado, L. Avilán, F. Bratti, R. K. Brizendine, P. A. Cox, R. Gill, R. Graham, D.-J. Kim, G. König, W. E. Michener, S. Poudel, K. J. Ramirez, T. J. Shakespeare, M. Zahn, E. S. Boyd, C. M. Payne, J. L. DuBois, A. R. Pickford, G. T. Beckham, J. E. McGeehan, Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. *Nat. Commun.* **13**, 7850 (2022).

144. D. Danso, C. Schmeisser, J. Chow, W. Zimmermann, R. Wei, C. Leggewie, X. Li, T. Hazen, W. R. Streit, New insights into the function and global distribution of Polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. *Appl. Environ. Microbiol.* **84** (2018).

145. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

146. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

147. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

148. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).