Modularly designed protein fragments combine into thousands of active and structurally diverse enzymes

Rosalie Lipsh-Sokolik[1], Olga Khersonsky[1], Sybrin P. Schröder[2], Casper de Boer[2], Shlomo-Yakir Hoch[1], Gideon J. Davies[3], Hermen S. Overkleeft[2] and Sarel J. Fleishmam[1]*

[1]Department of Biomolecular Sciences, Weizmann Institute of Science, 7610001, Rehovot, Israel

[2]Leiden Institute of Chemistry, Leiden University, Einsteinweg 55, 2300 RA Leiden, The Netherlands

[3]York Structural Biology Laboratory, Department of Chemistry, The University of York, Heslington, York, YO10 5DD, UK

*Corresponding author. Email: sarel@weizmann.ac.il

Abstract

Functional diversity in enzymes often requires significant backbone changes, but even subtle mutations in an enzyme may negatively impact its activity. We introduce CADENZ, an atomistic and machine-learning strategy to design modular enzyme fragments that combine all-against-all to generate diverse, low-energy structures with stable catalytic constellations. Applied to endoxylanases, CADENZ designed nearly a million unique proteins, and activity-based protein profiling with an endoxylanase-specific probe rapidly recovered 3,114 active enzymes that adopt 756 distinct backbones. Functional designs exhibit higher active-site preorganization than nonfunctional ones and more stable and compact packing outside the active site. Implementing these lessons into CADENZ led to a tenfold improved hit rate. This design-test-learn loop can be applied, in principle, to any modular enzyme or binding protein family, yielding huge diversity and general lessons on protein design principles.

Introduction

Innovation in many areas of engineering relies on the combination of preexisting modular parts(*1*). For instance, in electrical engineering, standard modular parts, such as transistors or processing units, are combined in new ways to assemble novel devices(*2*). Similarly, in a hypothetical, entirely modular protein, fragments could be combined to generate stable, well-folded, and potentially functional domains(*3*). In practice, however, protein domains exhibit a high density of conserved molecular interactions that are necessary for accurate native-state folding. Furthermore, mutations may be epistatic, such that they can only be incorporated against the background of other mutations, severely limiting options for fragment combination(*4*, *5*). Therefore, whereas recombination is an important source of novelty in natural and laboratory evolution(*6*–*8*) and the design of *de novo* backbones(*9*), due to epistasis, evolution is typically restricted to recombining fragments from only a few high-homology proteins.

Despite these challenges, immune-system antibodies present a remarkable example in which modularity enables extremely rapid and effective innovation through the combination of a small set of genetic fragments (V, (D), and J genes)(*10*). The result of this process is an enormous diversity of binding proteins that can counter, in principle, any pathogen. Nature has no equivalent strategy to generate structural and functional diversity in enzymes, but some protein folds, such as TIM barrels, β propellers, and repeat proteins, have evolved through the duplication, recombination, and mutation of modular fragments and are therefore prominent candidates for fragment combination. Moreover, these folds comprise some of the most structurally and functionally versatile enzymes and binding proteins in nature(*11*).

Here, we ask whether enzymes could be generated, similar to antibodies, from combinable fragments? We develop a method, called CADENZ, for Combinatorial Assembly and Design

of ENZymes, to design and select protein fragments that, when combined all-against-all, give rise to vast repertoires of low-energy proteins that exhibit high sequence and structural diversity. Isolating active enzymes in such vast protein libraries would be a daunting task when using contemporary (fluorogenic) substrate processing enzymes, but it can be readily and accurately achieved using activity-based protein profiling (ABPP). ABPP makes use of mechanism-based, covalent and irreversible inhibitors composed of a chemical scaffold that emulates structural features of the target substrate together with an enzyme active-site electrophile and a fluorophore or affinity tag. To exploit ABPP, we focused on glycoside hydrolase family 10 (GH10) xylanases(*12–14*) (Enzyme Classification: 3.2.1.8) as a model system and a dedicated GH10 xylanase-specific activity-based probe (ABP) as the principal enzyme activity readout. We found that CADENZ generated thousands of functional enzymes adopting more than 700 diverse backbones. We then trained a machine-learning model to rank designs based on their structure and energy features. Applying the learned model, we designed a second-generation library that demonstrated an order of magnitude increase in success rate in obtaining functional enzymes.

Design of modular and combinable protein fragments

For a protein fold to be a candidate for modular assembly and design, its secondary structure elements should be conserved among homologs while its loop regions exhibit diverse conformations including insertions and deletions(*15–17*). In such cases, the secondary structure elements typically provide robustness, whereas the loop regions encode functional differences. The TIM-barrel fold is a prime example of such modularity in which eight β/α segments comprise an inner β barrel surrounded by α helices(*18*, *19*). The catalytic pocket is located at the top of the barrel with critical contributions from all β/α loops. Evolutionary analysis indicates that TIM barrel proteins arose by dual duplication of an ancestral β/α-β/α segment, suggesting that modern TIM barrel enzymes can be segmented into four parts (Fig.

1A)(*20–22*). Nevertheless, in the course of evolution, each protein accumulated mutations to adapt the inter-segment interactions for specific functional and stability requirements. Therefore, recombining fragments from existing proteins mostly produces unstable and dysfunctional proteins that require further mutational optimization to turn into stable and active ones(*23*). To address this problem, the CADENZ design objective is to compute a spanning set of backbone fragments that give rise to folded and active proteins when combined all-against-all without requiring further optimization. The primary challenge CADENZ addresses is designing mutually compatible (modular) fragments, *i.e.*, ones among which epistasis is minimal.

CADENZ starts by aligning homologous but structurally diverse enzymes (in this case, 81 unique structures of GH10 xylanases) and fragmenting them along points that are structurally highly conserved (within the core β segments, see Fig. 1B for a visual guide to the algorithm)(*16*, *24*). Next, the fragments are designed to increase stability while holding the active site fixed. All design calculations take place within a single arbitrarily chosen template (PDB entry 3W24) to provide a realistic structural context for design and to promote compatibility between fragments. In practice, each fragment replaces the corresponding one in the template structure, and we use the PROSS stability-design algorithm(*27*) to implement stabilizing mutations within the fragment (8-42 mutations in each fragment; up to 28%) (Fig. S1A). In GH10 xylanases, the active site interacts with the xylan substrate through more than a dozen residues from all β/α loops(*25*, *26*) posing a significant challenge for modular design (Fig. 1C). To maintain catalytic activity, in all design calculations the sidechains of four key catalytic amino acids are restrained to their crystallographically observed conformations. At the end of this process, we obtain a set of fragments that are internally stabilized within a common template and designed to support the catalytically competent constellation of active-site residues.

Due to epistasis, however, combining the designed fragments would likely result in a majority of high-energy structures that are unlikely to fold into their intended conformation (Fig. 1D). To address this problem, we enumerate all possible full-length proteins by combining the designed fragments all-against-all and ranking them by Rosetta all-atom energy (Fig. 1B, step I). This process yields hundreds of thousands of unique structures, most of which exhibit unfavorable energies, as expected. To find mutually compatible (modular) fragments, we present a machine-learning-based approach, called EpiNNet (Epistasis Neural Network), which ranks fragments according to their probability to form low-energy full-length structures (Fig. 2A). EpiNNet is trained to predict whether a combination of fragments exhibits favorable Rosetta energy based on its constituent fragments. The trained network weights are then used to nominate fragments to generate the enzyme library. For the following steps, we used the top 6-7 fragments from each segment, assembling into 1,764 structures.

To add active-site diversity and increase the chances of favorable fragment combination, we next design several sequence variants for each of the backbone fragments. We use the FuncLib design method(*28*) to generate low-energy amino acid constellations at positions in the active-site and in the interfaces between designed β/α fragments while fixing the conformations of the key catalytic residues as observed in experimentally determined structures (Fig. 1C). We then use EpiNNet again, this time to find the single-point mutations that are most likely to form low-energy full-length proteins in combination with other mutations (Fig. S2A).

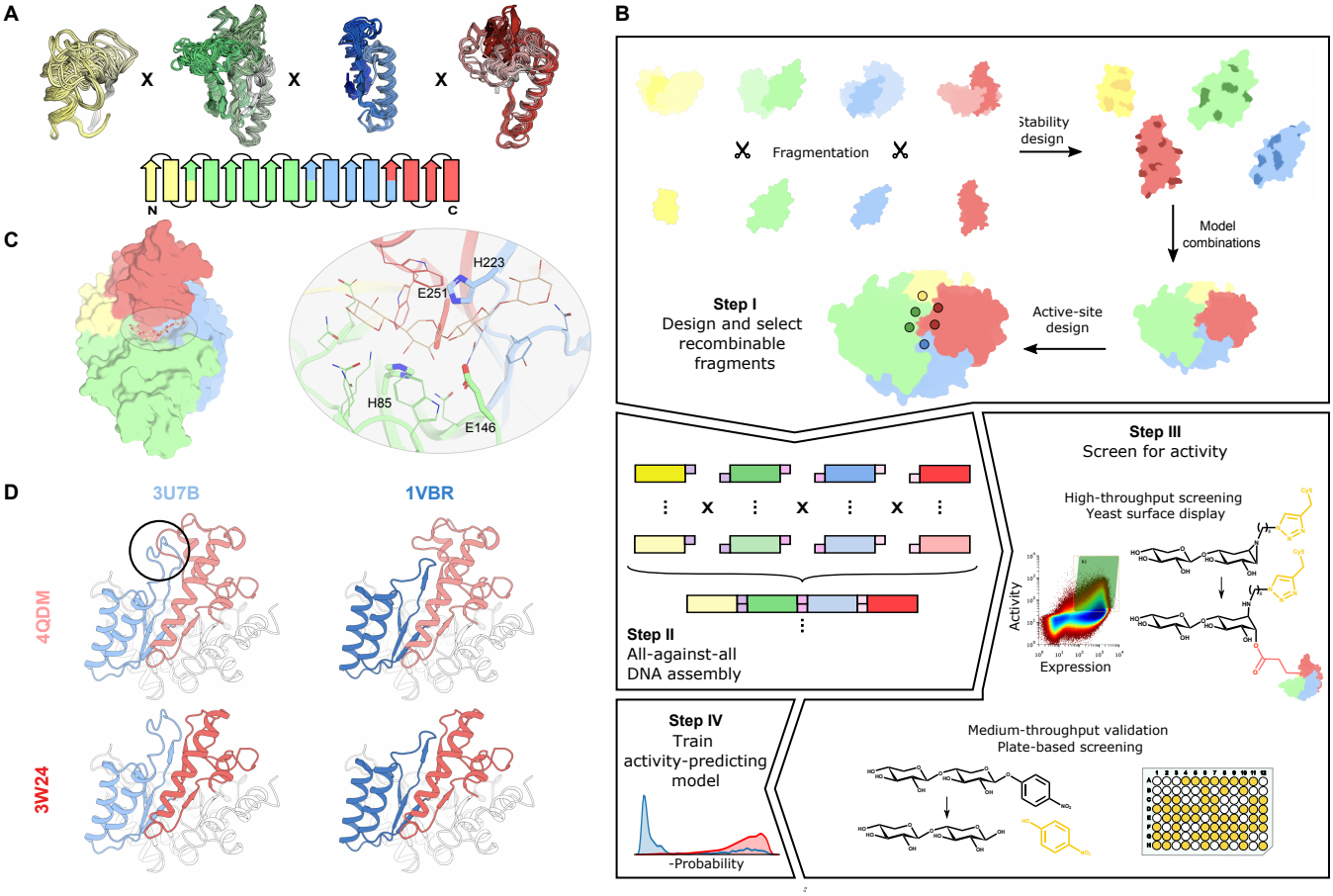

Figure 1. Key steps in the CADENZ workflow. (A) (*top*) cartoon representation of selected fragments. (*bottom*) segmentation scheme for GH10 xylanases (color scheme is consistent in all structural figures). **(B)** The design pipeline. **Step I:** Design maximizes internal stability and compatibility with other fragments and diversifies active-site positions that are not directly involved in the catalytic step. **Step II:** DNA oligos encoding fragments are ligated all-against-all using Golden Gate Assembly(29) to generate DNA molecules encoding the full-length designs. **Step III:** Designs are sorted using a xylobiose-emulating activity-based probe(30) that labels the nucleophilic Glu (red lines) of yeast-displayed functional enzymes. Activity is confirmed on a subset of the selected enzymes in a plate-based chromogenic assay. **Step IV:** An activity predictor is trained based on features that distinguish active and inactive designs. **(C)** Four catalytic residues are restrained throughout design calculations (sticks, numbering corresponds to PDB entry: 3W24). **(D)** Fragments can assemble into low or high-energy structures, depending on other fragments. (*top left*) segments 3 (blue) and 4 (red) are incompatible (overlap marked by black circle), resulting in extremely high energy (+1,529 Rosetta energy units; R.e.u.). The other designs exhibit low energies (<-950 R.e.u.).

To summarize, unlike current protein design methods which generate sets of specific full-length proteins, CADENZ designs a combination of fragments that can be combined to generate mostly low-energy structures that stabilize the catalytic constellation. CADENZ does not necessarily select the lowest-energy combinations but rather mitigates the risk of combining incompatible fragments. The consequences of inter-segment epistasis are striking: whereas the energies of the fully enumerated set of designed GH10s can be as high as +2,500 Rosetta energy units (R.e.u.), following EpiNNet fragment selection, the energies are <-890 R.e.u. (Fig. 2B). As a reference, we also generated the distribution of energies obtained by combining the sequence of the fragments selected by EpiNNet prior to any of the design steps. This reference simulates recombination of natural GH10 genes and exhibits a substantially worse energy distribution than the combination of PROSS-stabilized fragments (>100 R.e.u. difference on average), underscoring the dramatic impact of the design process (Fig. 2B and S1B). Furthermore, EpiNNet alleviates inter-segment epistasis by discarding backbone fragments and designed single-point mutations that are incompatible with neighboring segments (Fig. 2C-E). This analysis highlights the significant challenge that epistasis poses for effective fragment combination while underscoring the strengths of the EpiNNet selection strategy. Although EpiNNet eliminates many fragments, the designed library is extremely diverse, comprising 6-7 unique fragments per segment for a total of 952,000 sequences adopting 1,764 different backbones.

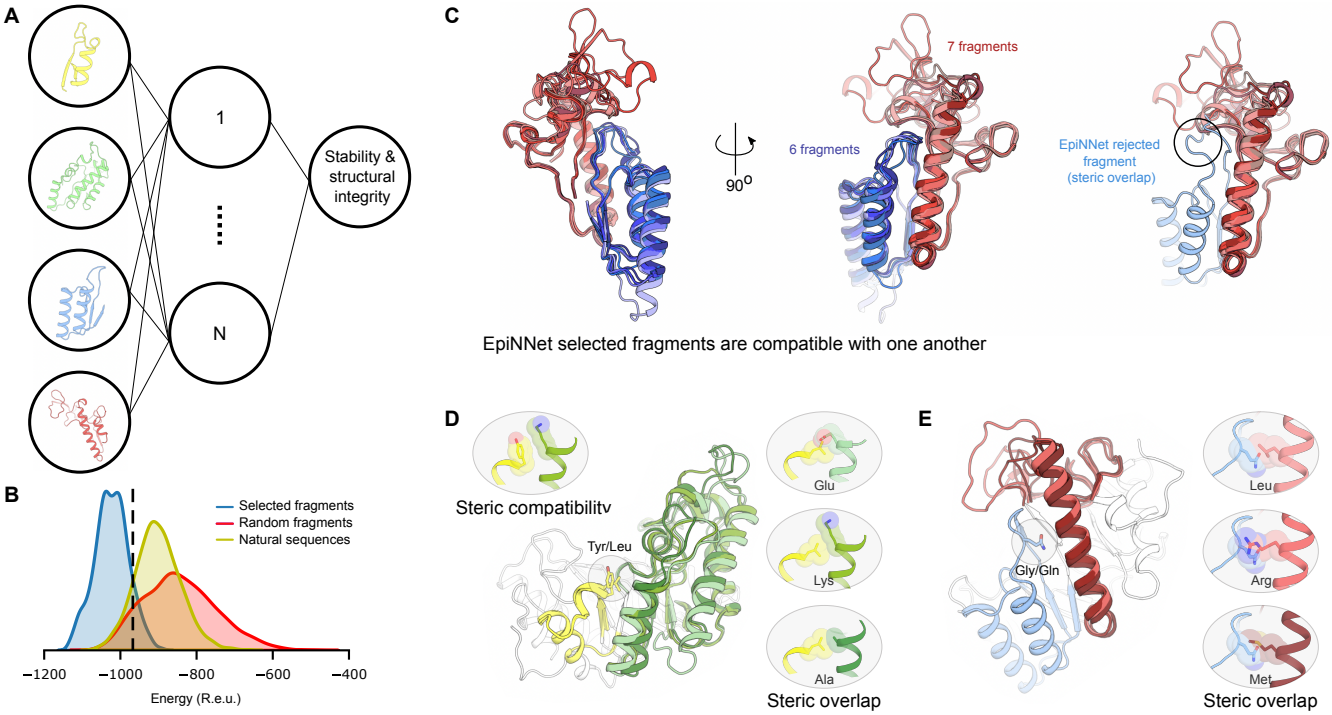

Figure 2. EpiNNet selects fragments that assemble to low-energy structures. (A) EpiNNet is trained to predict whether the input fragments assemble into low-energy full-length structures with catalytically competent active sites according to Rosetta atomistic calculations. Following training, EpiNNet selects fragments with a high probability of producing low-energy combinations. The input layer can represent backbone fragments (depicted here) or designed single-point mutations. **(B)** Rosetta energies of all 1,764 backbones assembled by a randomly selected set of fragments (red), fragments selected by EpiNNet with the sequences prior to design (yellow), or these fragments using their PROSS-designed sequences (blue). The majority (89%) of the EpiNNet-selected designs exhibit low energy (<-967 R.e.u., dashed line, see Methods). **(C)** EpiNNet removes incompatible fragments. (*left*) all fragments selected for segment 3 (blue) and 4 (red). (*right*) discarded fragment with a β/*α* loop that is incompatible with the other fragments. **(D-E)** Examples of mutations selected by EpiNNet (taken from the second-generation library). **(D)** Segment 1 of PDB entry 3W24 (yellow) faces segment 2 (green); EpiNNet prioritizes a Tyr over Leu which cannot be accommodated with neighboring fragments. **(E)** Segment 3 of PDB entry 1VBR (blue) faces segment 4 (red); EpiNNet prioritizes the small Gly over the large Gln.

CADENZ generates thousands of structurally diverse and active enzymes

We used Golden Gate Assembly to synthesize full-length genes from the designed fragments(*29*)(Fig. 1B Step II) and transformed the library into yeast cells for functional screening using cell-surface display(*31*) (Fig 1B, step III). To probe enzyme activity, we incubated the library with a xylobiose ABP(*30*), which reacts within the enzyme active site to form a covalent and irreversible ester linkage with the glutamic acid nucleophile(*32*). We then used fluorescence-activated cell sorting (FACS) to collect the population of yeast cells expressing active designs (Fig. S3A). ABP labeling depends on the nucleophilicity of the catalytic Glu, the ability of the active-site catalytic acid-base residue to enhance the electrophilicity of the ABP-epoxide by protonation, and the integrity of the xylan molecular recognition elements within the active-site pocket. Therefore, this is a sensitive probe for design accuracy in the active site (which comprises elements from all β/α units). Retaining glycosidase ABPs report on the first steps of substrate processing, namely ligand binding to the active site followed by nucleophilic attack. To confirm that selected proteins exhibit the complete catalytic cycle(*33*), we transformed *E. coli* cells with DNA from the sorted population and randomly selected 186 colonies for screening in 96-well plates using the chromogenic substrate 4-nitrophenyl β-xylobioside (O-PNPX_2)(*34*). This screen demonstrated that 58% processed the substrate (Fig. S3B), indicating that the majority of the designs selected by the ABP were true enzymes.

We next applied single-molecule real-time (SMRT) long-read sequencing(*35*) to the sorted population. Encouragingly, sequencing showed that the sorted population included a large number of structurally diverse designs: specifically, 3,114 unique designs based on 756 different backbones were active (Fig. 3A), compared to only 376 GH10 xylanase entries in the UniProt database(*36*). The active designs exhibited many insertions and deletions relative to one another, with sequence lengths varying from 317 to 395 amino acids and diverse β/α

loop conformations, with 62% sequence identity to one another on average. In all models, residues responsible for the catalytic steps are held in place by construction, but the pocket exhibits high geometric and electrostatic diversity (Fig. 3A bottom). Strikingly, the designs exhibit as many as 169 mutations and 48-73% sequence identity (Fig. 3B) to their nearest natural homolog (in the nr sequence database(*37*)), and most designs source fragments from four different structures (Fig. 3C).

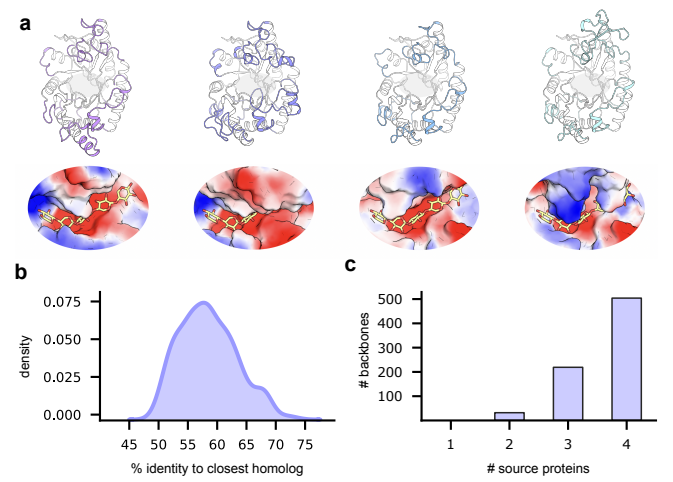

Figure 3. CADENZ generates functional enzymes with high structure and sequence diversity. (A) (*top*) representative model structures of active enzymes designed by CADENZ. High-diversity regions are highlighted in colors. (*bottom*) active-site electrostatic potential surfaces of the representative designs exhibit marked differences (putative ligand-bound conformation marked in yellow sticks based on PDB entry 4PUD). **(B)** Distribution of sequence identity to nearest natural homologs of active backbones. **(C)** The number of unique structures from which fragments are sourced. Most active designs incorporate fragments from four different sources.

Active designs are compact and preorganized for activity

The deep sequencing analysis provides an unprecedented dataset for improving enzyme-design methodology. For each design, we computed 85 structure and energy metrics, some relating to the entire protein, and others restricted to the active site. We avoided using the designed mutations or fragment identities as features for learning so that we might infer general lessons that apply to other enzymes. We tested the differences between the active and inactive sets using an independent two-sample *t*-test, finding that 63 metrics exhibited *p*-values $<10^{-10}$. To select the most meaningful metrics, we visually inspected the distributions of these metrics and focused on ten (Fig. 4A and S4) that do not exhibit significant correlations with one another (Methods). We then trained a logistic regression model based on these ten metrics to predict whether an enzyme is active (Fig. 4B and tables S1-S3).

We were encouraged to find that the ten dominant predictive features relate to essential aspects of enzyme catalysis. The most dominant feature is atomic density, which gauges protein compactness and correlates with stable packing. Another dominant feature is the compatibility of the amino acid identity and the local backbone conformation, a key determinant of protein foldability(*38*). By contrast, this feature is disfavored within the active-site pocket, presumably because active-site residues are selected for their impact on activity rather than stability. We also find that hydrogen-bond energies are highly discriminating, reflecting the high propensity of buried long-range hydrogen-bond networks in large proteins of a complex fold such as TIM barrels(*39*, *40*) (fig. S5). Rosetta system energy, however, makes a small contribution to predicting activity, but this is likely since all designs exhibit low energy by construction (Fig. 2B).

Within the active site, the model assigns almost equal importance to atomic density and van der Waals energy, two features that promote precise catalytic residue placement but penalize

overly packed constellations, respectively. The resulting dense yet relaxed packing arrangements are likely to be key to promoting active-site preorganization. Focusing on the four catalytic residues only, the model includes a feature that penalizes high repulsive energy, further underlining the importance of a relaxed and preorganized active site. Our analysis highlights prerequisites of catalytic activity that were not observed in previous high-throughput studies of design methods which focused on the kinetic stability of designed miniproteins and binders(*41*, *42*). They also highlight that the design objective function is substantially different within the active site and in the remainder of the protein.

Recently, the AlphaFold2 *ab initio* structure prediction method(*43*) has been shown to discriminate between correctly and incorrectly folded *de novo* designed binders(*44*). Applying AlphaFold2 to our set, however, showed no discernable difference between active and inactive designs in either the root-mean-square deviation (rmsd) between predicted and designed models or in the AlphaFold2 confidence scores (pLDDT; fig. S6). This result suggests that despite the high mutational load and the sequence and structure diversity in our designs, CADENZ generates sequences with native-like characteristics.

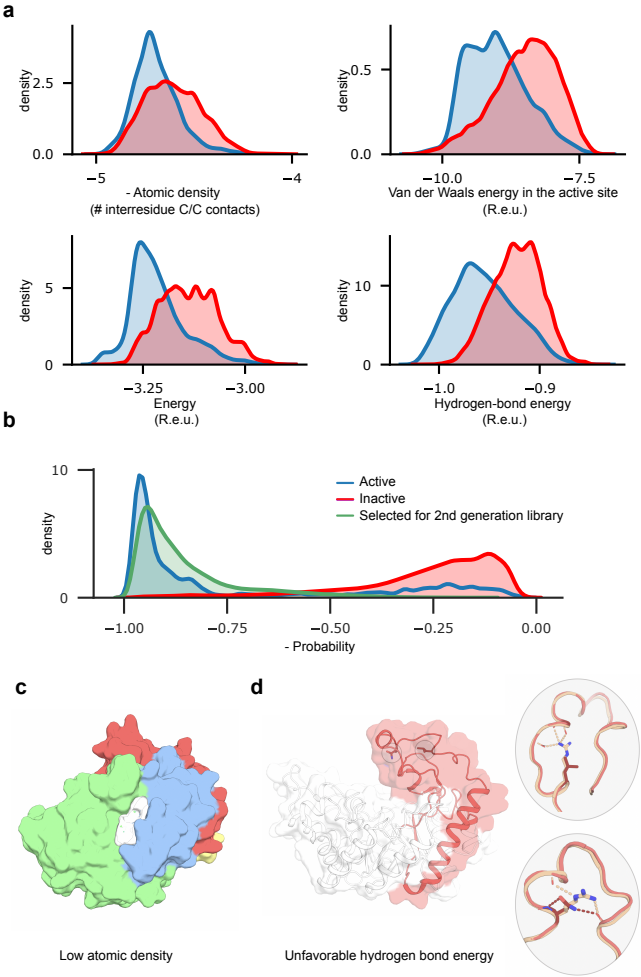

Figure 4. Energy and structure features discriminate between active and inactive designs. (A) Representative features included in the activity predictor. The features show a statistically significant difference between active and inactive designs, but none is individually an effective discriminator. All features are normalized by the protein length. Low values are favorable. **(B)** Separation of active and inactive designs by an activity predictor based on a logistic regression model. In green, probability distribution for designs assembled by the fragments selected for the second-generation library. **(C-D)** Examples of backbone fragments eliminated by the activity predictor in the second-generation library. Fragment color scheme as in Figure 1. **(C)** 2WYS in segment 2 (green) was selected for the first-generation library but discarded in the second due to low atomic density. The interface with segment 3 (blue) is poorly packed, leaving a gap between the segments (white). **(D)** 1UQZ in segment 4 (red) was selected for the first-generation library but discarded in the second due to unfavorable hydrogen-bond energy. Close inspection revealed two mutations introduced during sequence design, Arg282Asn and Arg289Leu (residue numbering refers to PDB entry: 1UQZ), eliminating hydrogen bonds that are crucial for β/α loop backbone stabilization. Mutations in red.

Order of magnitude increase in design success in second-generation library

We next asked whether the lessons we learned from the first-generation library could improve design success rate. We used the same set of combinatorial designs from the first library, but this time, instead of ranking them based on Rosetta energies, we ranked them according to the activity predictor (Fig. 4B and S2B). Then, we applied EpiNNet to nominate fragments that are likely to be mutually compatible. As in the first library, we designed several sequence variants for each backbone fragment. Here too, sidechain conformations of the core catalytic residues were held fixed in all design calculations. This library included three unique backbone fragments for each of the four segments and up to 11 sequence variants per fragment, resulting in 334,125 designed full-length xylanases using 81 different backbones.

The backbone fragments chosen in the first library but discarded in the second are predicted to assemble into full-length proteins with unfavorable properties in some of the features that were strongly discriminating. For instance, atomic density (Fig. 4C) and hydrogen-bond energy (Fig. 4D) are unfavorable in many discarded fragments. By contrast, designs assembled using fragments selected by the activity predictor have favorable values in features associated with foldability and high-efficiency catalysis and lower the risk of negative epistasis among segments.

We synthesized and screened the second-generation library as before (Fig. S7). Remarkably, sequencing confirmed 9,859 active designs, an order of magnitude increase in the rate of positive hits compared to the first library (Fig. 5A,B). We also screened the library using an ABP that is based on cellobiose, the disaccharide repeat moiety in cellulose rather than in xylan(*45*). We found 2,778 designs that reacted with the cellobiose ABP but were not sequenced in the library sorted with the xylan ABP, for a total of more than 12,637 active designs (3.8% of the population). These results demonstrate that CADENZ may generate

enzymes with varying substrate selectivity. To verify that the ABPs selected designs that exhibited the full catalytic cycle, we used plate-based validation with O-PNPX_2 and cellPNP confirming 85% and 60% of active clones in the xylobiose and cellobiose labeled populations, respectively (Fig. S7C,D).

Although this library is more focused than the first one, it is nevertheless extremely diverse. The active designs adopt 79 of the 81 backbones implemented in the library and present diverse sequence lengths: 312-347 amino acids (Fig. 5C). Although the activity predictor was blind to the identities of the designed fragments and mutations, we were concerned that it might have focused the second library on a set of fragments identified in functional enzymes in the first library. We analyzed the source of the active designs in the second-generation library, finding that 75% incorporate backbone fragments that were not encoded in the first library and verifying that the learned energy and structure features generalized to fragments not included in the training data. Moreover, the active designs are as divergent from natural GH10 enzymes as in the first library, exhibiting 50-73% identity to the most similar sequences in the nr database (Fig. 5D) and up to 140 mutations and eight unaligned regions (Fig. 5E). Furthermore, the second-generation library incorporates more active-site mutations (Fig. 5F), increasing the potential for altered substrate specificities. We also analyzed the distribution of energy and structure metrics among active and inactive designs in this second-generation library. The discrimination we observed, however, was significantly lower than in the first library, suggesting that the specific learning process we implemented converged.

As an independent test, we applied the learned activity predictor to select a small set of individually designed GH10 enzymes(*16*). Here, we used Rosetta atomistic modeling to enumerate all fragment combinations and design them as full-length enzymes instead of designing the fragments independently of one another. We then selected designs for

experimental testing using the activity predictor. Thus, in this implementation, the design and selection process did not favor modularity but rather optimal structure and energy properties. Twenty-seven designs were selected for experimental characterization with up to 143 mutations and 51-74% sequence identity to their nearest natural homolog. Although these designs were generated using a different process than the one used to train the activity predictor, remarkably, 25 (93%) of the designs were active in hydrolyzing O-PNPX$_2$ (Table S4), compared to fewer than half in a previous application of AbDesign to GH10 enzymes(*16*). We further characterized the kinetics of the seven most promising designs with various substrates (Table S5). Among these designs, several exhibited catalytic efficiencies (k_{cat}/K_M) comparable to natural GH10 xylanases from thermophiles, including against natural wood and wheat xylan (Fig. 5G,H), despite incorporating >80 mutations from any known natural protein sequence. These results are a marked improvement in the success of backbone design in enzymes and underscore that the lessons we learned from high-throughput screening can be applied to generate a diverse and highly active set of designs, whether for high- or low-throughput screening.

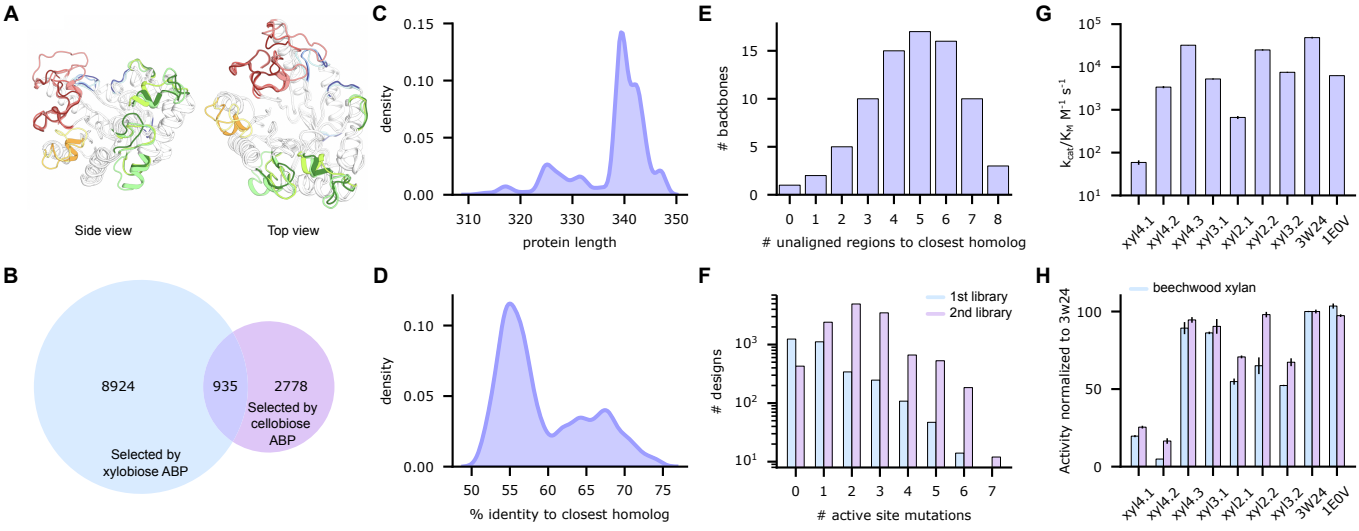

Figure 5. Activity predictor significantly increases design success rate. (A) Backbone fragments selected for the second-generation library (colors as in Fig. 1, low diversity regions in white). **(B)** Number of sequences in the population selected by the xylobiose and cellobiose ABPs (blue and purple, respectively). In the middle, designs selected by both ABPs. **(C)** Protein sequence length of active designs in the second-generation library. **(D)** Distribution of sequence identity to nearest natural homolog of active backbones in the second-generation library. **(E)** Number of unaligned regions to nearest natural homologous sequence of active backbones in the second-generation library. **(F)** Active designs in the second-generation library incorporate more active-site mutations than in the first-generation library. **(G)** Catalytic efficiency of seven xylanases from the small-set design and two representative natural ones (right, names correspond to PDB entry. 3W24 is a xylanase from the thermophilic organism *Thermoanaerobacterium saccharolyticum*). The first number in the design name indicates the number of different proteins from which the fragments were sourced. **(H)** Normalized activity with wheat arabinoxylan and beechwood xylan. Data are the means ± standard deviation of duplicate measurements.

Discussion

Modularity is a prerequisite for innovation in numerous engineering disciplines, but protein domains exhibit high epistasis, severely limiting their modularity. CADENZ addresses this conflict by designing a spanning set of low-energy and mutually compatible protein fragments that can be assembled into thousands of diverse and functional proteins. Thus, CADENZ massively increases the number and diversity of functional enzymes that can be interrogated relative to the natural diversity. The approach is general, requiring only a diverse set of homologous structures, and can be applied, in principle, to numerous binder and enzyme families. In a companion paper, we show that the EpiNNet strategy, developed here to select modular fragments, can be applied to design vast sequence libraries in the densely packed GFP chromophore-binding pocket, resulting in thousands of unique active designs that exhibit diverse spectral properties. Our design strategy can therefore be applied, in principle, to any protein to access unexplored and active structure and sequence spaces.

Combining CADENZ and ABPP enabled us to implement effective design-test-learn cycles on an unprecedented number of enzyme designs. Such cycles have led to deeper understanding of the design principles of *de novo* designed miniproteins(*41*, *42*) but have not yet been implemented at this scale nor to sophisticated enzymes. The rules we learned increased the design success rate by an order of magnitude and may provide a guide to improving macromolecular energy functions. Remarkably, these lessons were directly transferable to automated small-scale design, leading to >90% success rate in the design of active GH10 xylanases that exhibit more than 100 mutations from any natural protein. Some of these designs exhibited stability and activity levels on par with naturally occurring enzymes in thermophiles and some exhibited promiscuous cellulase activity that may provide a starting point for optimization. Furthermore, the functional data on a large number of

homologous yet structurally diverse enzymes may advance efforts to develop AI-based enzyme design methods.

Current methods for optimizing and diversifying proteins rely on cycles of mutation, recombination, and screening(*46*). Due to high epistasis, however, these methods explore a small fraction of sequence and structure space. By contrast, in the current study, we isolated more than 10,000 unique and active enzymes. Finally, CADENZ and ABPP are highly compatible methodologies, as shown here. The combined strategy can be extended to other substrates(*47*), whether ones catalyzed by natural enzymes or ones that are not. CADENZ can also be combined with other sophisticated high-throughput approaches, such as microfluidics-based screening, to offer a new and effective paradigm for enzyme discovery and optimization.

Acknowledgments

We thank Nir London, Birte Höcker, Shiran Barber-Zucker, and Dina Listov for discussions and Shira Warzsawski and Korina Goldin for technical help. RL-S is supported by a fellowship from the Arianne de Rothschild Women Doctoral Program. Research in the Fleishman lab was supported by the Volkswagen Foundation (94747), the Israel Science Foundation (1844), the European Research Council through a Consolidator Award (815379), the Dr. Barry Sherman Institute for Medicinal Chemistry, and a donation in memory of Sam Switzer. GJD thanks the Royal Society for the Ken Murray Research Professorship. HSO thanks the European Research Council (ERC-2011-AdG-290836 'Chembiosphing') and the Netherlands Organization for Scientific Research (NWO TOP grant 2018-714.018.002 'Endoglycoprobe') for funding.

Methods

Source code. Source code is available at https://github.com/Fleishman-Lab/CADENZ and in (*24*).

Template structure. We selected PDB entry 3W24 as the template structure. Its sequence was designed using PROSS(*27*) using ref2015 as the energy function(*48*) with default parameters and $\Delta\Delta G$ cutoff=-0.75 R.e.u. The 3W24 design incorporated 49 mutations. This structure was subsequently idealized and refined (see below). In all following design calculations, the four key catalytic amino acids (residue number: 85, 146, 223 and 251; all numbering with respect to PDB entry 3W24) were restrained to their crystallographically observed side chain conformations. The Rosetta model can be found at <github>/models/3w24_template.pdb.gz

GH10 xylanase source structures. We downloaded all GH10 structures from the Pfam database(*49*) (see <github>/data/gh10_pdb_ids) (143 structures). The structures were split into their constituent chains, then idealized and refined using Rosetta (see <github>/rosetta_scripts/idealize_n_refine.xml). A Position-specific Scoring Matrix (PSSM) was generated for each source structure using PROSS(*27*).

Source structure segmentation. All source structures were aligned to the template using PyMOL(*50*) and segmented into four fragments. The segmentation points are selected by eye to maximize structural conservation and seamless assembly of different fragments into full length proteins(*24*). The structures were first only crudely segmented, then further refined to match perfectly the template fragments using <github>/scripts/find_best_match.py. The segment borders relative to the template (PDB entry 3W24): S1: 19-45, S2: 46-186, S3: 187-249, S4: 250-320. The PSSMs were segmented accordingly. Example files of the segmented template structure and PSSM can be found at <github>/models/s*_3w24_template.pdb and <github>/pssms/template_s*.pssm.

Fragment design. Using AbDesign(*16*, *24*, *40*), each fragment is modeled in the context of the template, and the mainchain dihedral angles are computed and stored for the backbone assembly step (see <github>/rosetta_scripts/splice_out_library.xml for modeling and <github>/models/fragment_design_4pud_s4.pdb.gz for an output example of segment 4 from PDB entry: 4PUD modeled in the context of the template). The segments are two residues shorter than the previous step to allow two constant residues between segments (S1: 20-44, S2: 47-185, S3: 188-248, S4: 251-319). Next, the fragment's sequence is designed for stability using a custom version of PROSS (<github>/rosetta_scripts/fragment_only_filterscan.xml). In both the modeling and sequence design steps, only the sequence of the fragment is designed, while the template sequence is kept constant.

Fragment clustering. To enable full enumeration of all possible fragment combinations and enforce structural diversity, the fragments of each segment were clustered using MaxCluster(*51*) (performing RMSD fit, sequence-independent mode, maximum linkage clustering with the initial threshold set to 0.1 and maximum threshold set to 0.76) resulting in 19-26 clusters per segment. The fragment with the lowest Rosetta score per cluster was selected for the next modeling steps.

Modeling all backbone combinations. AbDesign is used to model and refine all backbone combinations (261,326 in total) according to the torsion angles databases and designed sequences from the previous step (<github>/rosetta_scripts/repertoire_splice_in.xml).

Select best backbone fragments. The Rosetta scores of the modeling steps are used to train a neural network. In the first library, the designs are labeled "1" if packstat > 0.6(*52*) and Rosetta score is within the top 10% among the designs (< -967.3 R.e.u.) . In the second library, we use the whole-structure activity predictor to label "1" designs with predicted

probability > 0.88 and packstat >= 0.58 (for labeling with activity predictor: <github>/notebooks/label_using_activity_predictor.ipynb). The network is generated using sklearn MLPClassifier with a single, 4-neuron, hidden layer. We used 4 neurons in the internal-layer to represent the 4 segments of the combined structures. The number of neurons can be determined empirically or in relation to the number of combined segments, as we have done here. Post-training, the fragments are ranked according to their activations and the top-ranked ones are selected for sequence design. For GH10, we used 6-7 and 3 fragments per segment for the first and second libraries, respectively. Code, training data and trained model of the first library can be found at <github>/notebooks/select_bb_fragments.ipynb.

Computational validation of selected backbone fragment, first library. We compared the energies of the 1,764 selected backbones to: (1) designs generated by randomly selected backbone fragments, where each segment had the same number of fragments as in the first-generation library, including the template's fragment (for fair comparison). (2) designs generated by the same fragments as in the library, but with the wild type sequence (PROSS sequence for the template). The selected and the random populations had 89% and 9.5% of designs with score better than the "1" labeling cutoff. In the designed backbone with natural fragment sequences, only 2 of the 1,764 designs had better energies than the designed sequence (fig. S1B).

Sequence design of selected backbone fragment. For each fragment, we selected 2-11 residues for diversification. We first selected active site residues by adding a xylan substrate from PDB entry 4PUD to all fragments and selecting positions within 3.5 Å from the substrate. For segments 1 and 3 we selected positions on β/*α* loop pointing towards the adjacent segment. We used a custom version of FuncLib(*28*) to calculate the allowed sequence space at the diversified positions for each fragment in the context of the template structure, similarly to the above "fragment design"

(<github>/rosetta_scripts/funclib_sequence_space_fragments.xml). In the first library, we used $\Delta\Delta G$<2.5 R.e.u. and PSSM>0 cutoffs. In the second library there were fewer fragments and thus we considered a larger sequence space: PSSM>-2, $\Delta\Delta G$<3.5 R.e.u. for segments 1 and 3, and $\Delta\Delta G$<6 R.e.u. for segments 2 and 4.

To model each fragment's multipoint mutants in all structural contexts, we model each multipoint mutant on all backbones that can be assembled using the fragments selected in the previous step. Overall, we modeled 1,460,382 and 9,438,696 designs in the first and second library, respectively.

Select multipoint mutants. We train a neural network to predict design quality based on the point mutations introduced to the assembled structure. The input layer is a one-hot encoding of the mutations. In the first library, we labeled "1" designs with Rosetta score difference of up to +5 R.e.u. from the no-mutations assembled structures. In the second library, we used the activity predictor with active site features to rank the designs, and labeled "1" the top 10% (*i.e.*, probability > 0.63). The network has the same architecture as the network from the best fragment selection step, besides having a hidden layer size matching the number of diversified positions (143 and 65 in the first and second libraries, respectively). Post-training, the multipoint mutants were ranked according to their activations. In the first library, the top-ranked mutants are selected (up to 7 multipoint mutants per fragment). In the second library we observed similar multipoint mutants; thus, to increase diversity, we (1) clustered the multipoint mutants for fragments with >40 multipoint mutants (2) removed backbone fragments without representatives in the top 40% of clustered multipoint mutants, leaving only 3 backbone fragments per segment (3) selected the best (up to 11) multipoint mutants per fragment.

Computational validation of selected fragment, first library. To confirm that multipoint mutant combinations give rise to low-energy designs, we randomly modeled 100,000 designs

generated by selected multipoint mutants and another set of 100,000 designs from randomly selected multipoint mutants (fig. S2A). While the randomly-selected population had only 62% designs labebeld good (defined in the above section), the selected population had 96% good designs.

Computational validation of selected fragments, second library. To confirm that combinations of mutated fragments give rise to designs with good predicted probabilities by the activity-predictor, we randomly modeled 50,000 designs generated by selected multipoint mutants and another set of 50,000 designs from randomly selected multipoint mutants (fig. S2B). While the randomly selected population had only 3.9% of designs labeled "1", the selected population had 95.6% such designs.

Design of DNA encoding backbone fragments. We start by designing the DNA to be codon-optimized for *E. coli* expression. To ensure the correct assembly of fragments into full-length enzyme-encoding DNA molecules in the test tube (*i.e.*, number and order of fragments), the DNA molecules are flanked with a unique sequence coding for a recognition site for the *BsaI* Type IIs restriction enzyme. This enzyme digests the DNA outside its recognition site, thus leaving a designable 4-bp overhang. DNA molecules encoding for the same segment are designed to have identical overhangs. The DNA sequence of the two-residues constant region between segments is used as overhangs for Golden Gate assembly(*29*). Overhangs are selected to be compatible for one-pot assembly (i.e., overhangs are not reverse complements of one another and at least 2 base pair different from each other). Full DNA sequences can be found here <github>/data/rnd*_fragments_dna.csv

Golden Gate Assembly of fragment DNA into full-length enzyme coding molecules. Synthetic dsDNA encoding the fragments were custom synthesized as linear fragments by TWIST Bioscience. All fragments of each segment were pooled. To increase efficiency, we

split the Golden gate reaction into three: (1) assembly of fragments from segments 1 & 2 pools; (2) assembly of fragments from segments 3 & 4 pools (3) assembly into final enzyme-encoding DNA by pooling the above two reactions. The assembly is done according to the *Golden Gold (24 fragment) assembly protocol* from NEB(*53*). To avoid non-specific amplification, the assembly product was amplified in two steps: an initial 8-cycles amplification, splitting into 16 reactions, then an additional 8-cycles amplification. Elongation was shortened to 30sec to prevent non-specific amplification.

Yeast surface display. The xylanase libraries were cloned into pCTCon2 vector using homologous recombination in yeast, and the active xylanases were selected by yeast-display. The libraries were transformed into EBY100 cells by electroporation(*54*), the cells were grown overnight at 30°C in SDCAA and then resuspended in 10ml induction medium (SGCAA) and induced for 20h at 20°C(*31*). Around 10^7 cells were taken for labeling, the cells were first labeled with primary antibody for expression monitoring (mouse monoclonal IgG1 anti-cMyc (9E10), Santa Cruz Biotechnology), diluted 1:100 in PBS, for 30 minutes in RT. Then, the cells were labeled with a mixture of secondary antibody for expression monitoring (AlexaFluor488 goat-anti-mouse IgG1 (Life Technologies), diluted 1:100) and 0.03μM activity-based probe (xyl-Cy5 or cel-Cy5, diluted from 10mM stock in DMSO), in PBS, for 30 minutes in RT. ABP concentration was chosen by labeling the positive controls (PDB entry: 1TA3 and xyl3.1 and xyl8.3 from(*16*)) with various ABP concentrations, and choosing a concentration at which high signal was obtained while non-specific binding (of a non-expressing population) was minimized. Selections were performed using Fluorescence Activated Cell Sorting (FACS), with BD FACS Melody instrument for the first round library and S3e Cell Sorter (Bio-Rad) for the second round library. Cells were collected based on a combination of three selection gates: yeast cells (FSC vs SSC plot), singlet cells, and 2-5% of top binders. Three sorts were performed for xyl-Cy5 and four sorts for cel-Cy5.

O-PNPX$_2$ and cellPNP plate-based activity assay. The plasmids from sorted libraries were extracted using Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research). The xylanase inserts were amplified by PCR, re-cloned into pETMBPH vector(*55*) using EcoRI and PstI restriction sites, and transformed into BL21 DE3 cells, which were plated on LB plates with kanamycin. Individual colonies were cultured in 96-well plates with 300 µl of 2YT medium supplemented with kanamycin, the cultures were grown overnight at 37°C and used to inoculate 500µl of 2YT medium supplemented with kanamycin. The cultures were grown for 2 hours at 37°C, induced with 0.2mM IPTG, grown overnight at 20°C, pelleted and frozen at -20°C. The cells were lysed by mixing for 1 hour at 37°C in 300µl lysis buffer (Tris pH 7, supplemented with 100 mM NaCl, 0.1mg/ml lysozyme and benzonase) and pelleted. Xylanase activity was assayed using 100µl lysate and 0.5mM O-PNPX$_2$ substrate (4-nitrophenyl beta-D-xylobioside, Megazyme) or cellPNP substrate (4-nitrophenyl beta-cellobioside, Sigma Aldrich), at 37°C, by monitoring the absorbance of the leaving group at 405nm in a BioTek plate reader. Designed xylanases xyl3.2, xyl3.1 and xyl3.3(*16*) were used as positive controls. Xylanases were labeled active if the initial reaction velocity (mOD/min) was > 0.3 for O-PNPX$_2$ and > 0.4 for cellPNP.

Deep sequencing. DNA of the active population was extracted from yeast cells using Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research) and amplified in two steps, as described above for the Golden Gate reaction product. In the second library, each population was amplified with a unique barcoded primer. The *Amplicon Template Preparation and Sequencing* from PacBio(*56*) was used to prepare the DNA for long-read sequencing, and the Circular Consensus Sequence (CCS) analysis with default parameters was used to extract the high-quality reads. Each read was run in BLAST against all possible designs to assign the design it is encoding. In the first library, reads were filtered out if mutated on the protein level relative to their matched design. In the second library, reads were filtered out if #mutations >

2. Reads with #mutations <= 2 were filtered out if mutations were in key positions (i.e., active site and fragment interfaces). The list of active designs can be found here (<github>/data/active_designs_rnd1 and <github>/data/activity_rnd2.pickle).

Sequence identity analysis. Sequence identity to closest natural enzymes was calculated using BLASTP(*57*) with the NCBI non-redundant (nr) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/).

AlphaFold2 analysis of designed backbones. The structures of the backbones (without multipoint mutants) in the first designed library were predicted using ColabFold(*58*) with default parameters. RMSD between Rosetta models and AlphaFold2 predictions were done using the RMSD filter in RosettaScripts(*52*).

Metric calculations and filtering. All designs in the library were modeled in Rosetta (using <github>/rosetta_scripts/funclib_sequence_space_fragments.xml) and the different metrics were calculated (see <github>/rosetta_scripts/features.xml and <github>/rosetta_scripts/features_repack_activesite.xml for calculations, <github>/data/rnd*_all_seqs.fa for sequences and <github>/data/rnd*_features.csv for metrics data on all designs of the first and second library). All metrics depending on protein length were normalized (*i.e.*, metric / #residues); features relating to the entire protein were normalized by the protein's length, whereas features restricted to the active site were normalized by the number of residues included in the active site (12-18 residues, depending on the backbone fragments included in the protein). Following visual inspection of the distribution of the metrics in the active and inactive populations, 18 metrics were considered as features for the activity predictor: fa_rep_catres, vdw, vdw_activesite, fa_sol_catres, fa_elec, fa_elec_activesite, hbond, hbond_activesite, p_aa_pp, p_aa_pp_activesite, p_aa_pp_catres, fa_dun_catres, yhh_planarity, yhh_planarity_activesite, degree_activesite,

contacts, contacts_activesite, and total_score. The VDW metric is defined as fa_atr + 0.55 * fa_rep (as weighted in Rosetta ref15 scoring function(*48*). The hbond metric is defined as the sum of all hydrogen bond energies (hbond_lr_bb + hbond_sr_bb + hbond_bb_sc + hbond_sc). The features were standardized by removing the mean and scaling to unit variance.

To eliminate correlated features, we used feature selection procedures. The data was split into training and test sets (80 and 20 percent of the data, respectively). We used stepwise selection: at each iteration, there is a forward and a backward step. Forward step: add the feature having the lowest *p*-value when added to the logistic regression model, stop when *p*-value > 0.05 or the feature's coefficient < 0.1. Backward step: train the model with the newly added feature, and drop features with an increased *p*-values (> 0.05). The stepwise selection procedure was repeated five times, each with a different under-sampled population from the train set. We used the 10 features that were selected in all five repeats. To make sure all features contribute to the model performance, we performed a likelihood ratio test for each feature where we compare the model to a restricted model lacking the tested feature (*i.e.*, restricted model). We compute $\chi^2 = -2log(\frac{L_{null}}{L_{alternative}})$. The null hypothesis, where the restricted model is better, is rejected if χ^2 is greater than a χ^2 percentile with *k* degrees of freedom (*k* is the difference in parameters between the models; in this case *k*=1). We trained the models using the python package statsmodels, which also computes the log-likelihood of the model. All features significantly rejected the null hypothesis (*p*-value → 0). The features included in the models and their coefficients are summarized in Table S1.

Logistic regression model training. A few resampling procedures were tested as the data is highly unbalanced: oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE), with and without Tomek links (all were performed using the python

package imbalanced-learn). All methods gave similar results; thus, we chose random undersampling as it is less prone to overfitting and is faster to train on. We trained two logistic regression models; one including the 10 selected features described above (Logit model from python's statsmodels, regularized fit with α=0.1). The second model includes only features describing all residues of the protein. The features of the second model were selected as described above and are listed in Table S1. The trained models can be found at <github>/ml_models/logit_xylanases*.pickle

Design of a small set of xylanases. Designs were generated as described(*24*). Briefly, we used the same structures and PSSMs from the library design. Fragments were extracted and modeled in the context of the template to compute the mainchain dihedral angles using AbDesign (<github>/rosetta_scripts/small_set_splice_out.xml). The fragments of each segment were clustered, and the lowest Rosetta energy fragment was selected for the backbone recombination and design step (25, 20, 17, and 21 fragments for segments 1,2,3, and 4, respectively, see <github>/rosetta_scripts/small_set_splice_in.xml). During design, the active site residues (both core catalytic and binding substrate residues) were biased towards their crystallographically observed conformations. The assembled and designed structures were filtered for: (1) correct positioning of Trp292 and Trp300 in the active site, i.e., 1.5Å RMSD from the conformation in the template structure (2) packstat >= 0.58. The filtered 29,451 designs (18.5% of designed chimeras) were labeled using both activity predictors and 1170 designs ranked among the top 7% of both predictors were selected for PROSS stability design. To control the number of different source proteins in the chimeras, 21 and 72 designs from only 2 or 3 sources (template and either segment 2 or 4 from a different source) were also selected for PROSS. Segments 2 and 4 were selected for diversification as they harbor most active site residues and are more structurally diverse than segments 1 and 3.

We applied PROSS to design the selected designs using a sequence space of mutations with PSSM probability>0 and $\Delta\Delta G$ < -0.75 R.e.u., incorporating 4.6 mutations on average. We filtered the designs again using packstat and Trp residues positioning as described above, and scored the remaining 1,155 designs using both activity predictors. To increase diversity in the experimentally tested set, the designs were structurally clustered (see Fragments clustering above), and a representative with the highest average probability of both activity predictors was selected. We filtered out: (1) representatives with predicted probabilities lower than the mean probability (2) representatives with segments 2 and 4 sourced from the same proteins. We similarly selected designs for the sets of designs where we explicitly select for only 2 and 3-source, filtering out for probabilities lower than the 25th percentile instead of the 50th. Overall, we selected 27 designs; 19, 3 and 5 from general selection, 2-source and 3-source designs, respectively.

We used pSUFER(*59*) to detect poorly designed positions, flagging positions with more than 4 alternative amino acids with $\Delta\Delta G$ < 0, located outside the active site. Those positions were designed by FuncLib (*28*), using mutations with PSSM > -2 and $\Delta\Delta G$ < 6 R.e.u. but present in the natural diversity of xylanases. The designs were computationally validated using AlphaFold2 (see above AlphaFold2 analysis of designed backbones) with 96.12 < plDDT < 98.13. The sequences can be found at <github>/data/small_set_xylanase.fa

Experimental validation of a small set of xylanases. The genes of the 27 designed xylanases were synthesized by TWIST Bioscience and cloned into pETMBPH vector using restriction sites *EcoRI* and *PstI*. The genes were transformed into BL21 DE3 cells and expressed in 10ml culture as described before. After centrifugation and storage at –20°C, the pellets were resuspended in the lysis buffer and lysed by sonication. Xylanase activity was assayed using 100 µl lysate and 0.5mM O-PNPX_2 and cellPNP substrates, as described above, and seven most active designs were taken for further characterization (Table S4).

The proteins were expressed in 50 ml culture, lysed by sonication and the cleared lysate was bound to amylose resin (NEB), washed with 50mM Tris pH 7 with 100mM NaCl, and the proteins were eluted with wash buffer containing 10mM maltose. Elution fraction was analyzed for purity by SDS-PAGE and used for kinetic measurements. For stability measurements, the MBP fusion tag was cleaved by TEV (1:20 TEV, Tris 50mM with 100mM NaCl and 1mM DTT, 24-48h at room temperature), the MBP fusion tag was removed by binding to Ni-Nta resin.

Specific activity (μM product per minute per mg protein) was measured with 0.5mM O-PNPX$_2$ and cellPNP at pH 6.5 and 37°C by monitoring the absorbance of the leaving group at 405 nm. For determination of kinetic parameters, activity with O-PNPX$_2$ was measured with a range of substrate concentrations, and the kinetic parameters were obtained by fitting the data to the Michaelis-Menten equation [$v_0 = k_{cat}[E]_0[S]_0/([S]_0 + K_M)$] using Prism 7. In cases where solubility limited substrate concentrations, data were fitted to the linear regime of the Michaelis-Menten model ($v_0 = [S]_0[E]_0 k_{cat}/K_M$) and k_{cat}/K_M values were deduced from the slope. The reported values represent the means ± S.D. of at least two independent measurements.

Activity with natural xylans. Xylanase activity with beechwood xylan and wheat arabinoxylan was determined by measuring the reducing sugars released from xylan by the dinitrosalicylic acid (DNS) method(*60*). 50μl protein was mixed with 50μl of 0.1M sodium citrate buffer pH 6, and added to 100μl of 2% xylan dispersed in water. Reaction mixture was incubated 0.5 hour at 50°C and then quenched for 10 minutes on ice. Then, 200μl of DNS reagent was added, the mixture was vortexed, incubated for 10 minutes at 95°C and cooled on ice. The mixture was diluted 1 to 4 in water, and absorbance at 540 nm was measured, compared to a blank sample (cell lysate expressing MBP). The reported values represent the means ± S.D. of two independent measurements.

T_M measurements. Melting temperatures of the tagless xylanases were measured by nanoDSF (nanoscale differential scanning fluorimetry) on Prometheus NT.Plex instrument (NanoTemper Technologies)(*61*). The temperature was increased from 20°C to 95°C at 1°C/min ramp, and the melting temperatures were calculated at inflection point.

References

1. M. E. Csete, J. C. Doyle, Reverse engineering of biological complexity. *Science*. **295**, 1664–1669 (2002).
2. Y. Lazebnik, Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer Cell*. **2**, 179–182 (2002).
3. M. Parter, N. Kashtan, U. Alon, Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLoS Comput. Biol.* **4**, e1000206 (2008).
4. D. M. Weinreich, R. A. Watson, L. Chao, Perspective: Sign epistasis and genetic costraint on evolutionary trajectories. *Evolution*. **59**, 1165–1174 (2005).
5. D. A. Kondrashov, F. A. Kondrashov, Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
6. C. A. Voigt, C. Martinez, Z.-G. Wang, S. L. Mayo, F. H. Arnold, Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002).
7. S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, Global view of the protein universe. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11691–11696 (2014).
8. N. Ferruz, F. Lobos, D. Lemm, S. Toledo-Patino, J. A. Farías-Rico, S. Schmidt, B. Höcker, Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. *J. Mol. Biol.* **432**, 3898–3914 (2020).
9. T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski, B. Kuhlman, Design of structurally distinct proteins using strategies inspired by evolution. *Science*. **352** (2016), pp. 687–690.
10. K. Murphy, C. Weaver, *Janeway's Immunobiology* (Garland Science, 2016).
11. E. Dellus-Gur, A. Toth-Petroczy, M. Elias, D. S. Tawfik, What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J. Mol. Biol.* **425**, 2609–2621 (2013).
12. S. G. Withers, D. Dombroski, L. A. Berven, D. G. Kilburn, R. C. Miller Jr, R. A. Warren, N. R. Gilkes, Direct 1H n.m.r. determination of the stereochemical course of hydrolyses catalysed by glucanase components of the cellulase complex. *Biochem. Biophys. Res. Commun.* **139**, 487–494 (1986).
13. E. Drula, M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat, N. Terrapon, The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).
14. CAZypedia Consortium, Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology*. **28**, 3–8 (2018).
15. A. Tóth-Petróczy, D. S. Tawfik, The robustness and innovability of protein folds. *Curr. Opin. Struct. Biol.* **26**, 131–138 (2014).
16. G. Lapidoth, O. Khersonsky, R. Lipsh, O. Dym, S. Albeck, S. Rogotner, S. J. Fleishman, Highly active enzymes by automated combinatorial backbone assembly and sequence design. *Nat. Commun.* **9**, 2780 (2018).
17. R. Netzer, D. Listov, R. Lipsh, O. Dym, S. Albeck, O. Knop, C. Kleanthous, S. J. Fleishman, Ultrahigh specificity in a network of computationally designed protein-interaction pairs. *Nat. Commun.* **9**, 5286 (2018).
18. N. Nagano, C. A. Orengo, J. M. Thornton, One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765 (2002).
19. R. Sterner, B. Höcker, Catalytic versatility, stability, and evolution of the (betaalpha)8-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055 (2005).
20. B. Höcker, J. Claren, R. Sterner, Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 16448–16453 (2004).

21. B. Höcker, S. Beismann-Driemeyer, S. Hettwer, A. Lustig, R. Sterner, Dissection of a (betaalpha)8-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36 (2001).

22. M. Richter, M. Bosnali, L. Carstensen, T. Seitz, H. Durchschlag, S. Blanquart, R. Merkl, R. Sterner, Computational and experimental evidence for the evolution of a (beta alpha)8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *J. Mol. Biol.* **398**, 763–773 (2010).

23. D. L. Trudeau, M. A. Smith, F. H. Arnold, Innovation by homologous recombination. *Curr. Opin. Chem. Biol.* **17**, 902–909 (2013).

24. R. Lipsh-Sokolik, D. Listov, S. J. Fleishman, The AbDesign computational pipeline for modular backbone assembly and design of binders and enzymes. *Protein Sci.* **30**, 151–159 (2021).

25. S. R. Andrews, S. J. Charnock, J. H. Lakey, G. J. Davies, M. Claeyssens, W. Nerinckx, M. Underwood, M. L. Sinnott, R. A. J. Warren, H. J. Gilbert, Substrate specificity in glycoside hydrolase family 10: tyrosine 87 and leucine 314 play a pivotal role in discriminating between glucose and xylose binding in the proximal active site of Pseudomonas cellulosa xylanase 10A. *J. Biol. Chem.* **275**, 23027–23033 (2000).

26. B. Henrissat, I. Callebaut, S. Fabrega, P. Lehn, J. P. Mornon, G. Davies, Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 7090–7094 (1995).

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

28. O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D. L. Trudeau, J. Prilusky, P. Amengual-Rigo, V. Guallar, D. S. Tawfik, S. J. Fleishman, Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* (2018), doi:10.1016/j.molcel.2018.08.033.

29. C. Engler, R. Kandzia, S. Marillonnet, A one pot, one step, precision cloning method with high throughput capability. *PLoS One*. **3**, e3647 (2008).

30. S. P. Schröder, C. de Boer, N. G. S. McGregor, R. J. Rowland, O. Moroz, E. Blagova, J. Reijngoud, M. Arentshorst, D. Osborn, M. D. Morant, E. Abbate, M. A. Stringer, K. B. R. M. Krogh, L. Raich, C. Rovira, J.-G. Berrin, G. P. van Wezel, A. F. J. Ram, B. I. Florea, G. A. van der Marel, J. D. C. Codée, K. S. Wilson, L. Wu, G. J. Davies, H. S. Overkleeft, Dynamic and Functional Profiling of Xylan-Degrading Enzymes in Aspergillus Secretomes Using Activity-Based Probes. *ACS Cent. Sci.* **5**, 1067–1078 (2019).

31. G. Chao, W. L. Lau, B. J. Hackel, S. L. Sazinsky, S. M. Lippow, K. D. Wittrup, Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).

32. D. Tull, S. G. Withers, N. R. Gilkes, D. G. Kilburn, R. A. Warren, R. Aebersold, Glutamic acid 274 is the nucleophile in the active site of a "retaining" exoglucanase from Cellulomonas fimi. *J. Biol. Chem.* **266**, 15621–15625 (1991).

33. D. Tull, S. G. Withers, Mechanisms of cellulases and xylanases: a detailed kinetic study of the exo-beta-1,4-glycanase from Cellulomonas fimi. *Biochemistry*. **33**, 6363–6370 (1994).

34. H. Taguchi, T. Hamasaki, T. Akamatsu, H. Okada, A Simple Assay for Xylanase Using o-Nitrophenyl-β-D-xylobioside. *Biosci. Biotechnol. Biochem.* **60**, 983–985 (1996).

35. A. Rhoads, K. F. Au, PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. **13**, 278–289 (2015).

36. UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

37. E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, S. T. Sherry, Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

38. A. Goldenzweig, S. J. Fleishman, Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **87**, 105–129 (2018).

39. O. Khersonsky, S. J. Fleishman, What Have We Learned from Design of Function in Large Proteins? *BioDesign Research*. **2022**, 1–11 (2022).

40. D. Baran, M. G. Pszolla, G. D. Lapidoth, C. Norn, O. Dym, T. Unger, S. Albeck, M. D. Tyka, S. J. Fleishman, Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10900–10905 (2017).

41. G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*. **357**, 168–175 (2017).

42. 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, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein binding proteins from target structure alone. *Nature* (2022), doi:10.1038/s41586-022-04654-9.

43. 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).

44. N. 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. *bioRxiv* (2022), p. 2022.06.15.495993.

45. C. de Boer, N. G. S. McGregor, E. Peterse, S. P. Schröder, B. I. Florea, J. Jiang, J. Reijngoud, A. F. J. Ram, G. P. van Wezel, G. A. van der Marel, J. D. C. Codée, H. S. Overkleeft, G. J. Davies, Glycosylated cyclophellitol-derived activity-based probes and inhibitors for cellulases. *RSC Chem Biol*. **1**, 148–155 (2020).

46. 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).

47. L. Wu, Z. Armstrong, S. P. Schröder, C. de Boer, M. Artola, J. M. Aerts, H. S. Overkleeft, G. J. Davies, An overview of activity-based probes for glycosidases. *Curr. Opin. Chem. Biol.* **53**, 25–36 (2019).

48. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack Jr, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

49. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

50. W. L. DeLano, Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* (2002) (available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.5879&rep=rep1&type=pdf#page=44).

51. MaxCluster, (available at http://www.sbg.bio.ic.ac.uk/~maxcluster/index.html).

52. S. J. Fleishman, A. Leaver-Fay, J. E. Corn, E.-M. Strauch, S. D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One*. **6**, e20161 (2011).

53. New England Biolabs, Golden Gate (24 Fragment) Assembly Protocol, (available at https://international.neb.com/protocols/2018/06/05/golden-gate-24-fragment-assembly-protocol).

54. L. Benatuil, J. M. Perez, J. Belk, C.-M. Hsieh, An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23**, 155–159 (2010).

55. Y. Peleg, T. Unger, Application of high-throughput methodologies to the expression of recombinant proteins in E. coli. *Methods Mol. Biol.* **426**, 197–208 (2008).

56. K. J. Travers, C.-S. Chin, D. R. Rank, J. S. Eid, S. W. Turner, A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).

57. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST : architecture and applications. *BMC Bioinformatics*. **10** (2009), , doi:10.1186/1471-2105-10-421.

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

59. D. Listov, R. Lipsh-Sokolik, S. Rosset, C. Yang, B. E. Correia, S. J. Fleishman, Assessing and enhancing foldability in designed proteins. *Protein Sci.* **31** (2022), doi:10.1002/pro.4400.

60. G. L. Miller, Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31**, 426–428 (1959).

61. Strutz, Exploring protein stability by NanoDSF. *Biophys. J.* (available at https://www.cell.com/biophysj/pdf/S0006-3495(15)03309-3.pdf).