Corrected 8 August 2023. See full text.

ENZYME DESIGN

Combinatorial assembly and design of enzymes

R. Lipsh-Sokolik[1], O. Khersonsky[1], S. P. Schröder[2], C. de Boer[2]†, S.-Y. Hoch[1], G. J. Davies[3], H. S. Overkleeft[2], S. J. Fleishman[1]*

The design of structurally diverse enzymes is constrained by long-range interactions that are necessary for accurate folding. We introduce an atomistic and machine learning strategy for the combinatorial assembly and design of enzymes (CADENZ) to design fragments that combine with one another to generate diverse, low-energy structures with stable catalytic constellations. We applied CADENZ to endoxylanases and used activity-based protein profiling to recover thousands of structurally diverse enzymes. Functional designs exhibit high active-site preorganization and more stable and compact packing outside the active site. Implementing these lessons into CADENZ led to a 10-fold improved hit rate and more than 10,000 recovered enzymes. This design-test-learn loop can be applied, in principle, to any modular protein family, yielding huge diversity and general lessons on protein design principles.

Innovation in many areas of engineering relies on the combination of preexisting modular parts (*1*). For example, in electrical engineering, standard modular parts, such as transistors or processing units, are combined to assemble devices (*2*). Similarly, in a hypothetical and entirely modular protein, fragments could be combined to generate stable, well-folded, and potentially functional domains (*3*). However, in practice, 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*). Recombination is an important source of protein diversity in natural and laboratory evolution (*6*–*8*) and the design of de novo backbones (*9*); however, because of epistasis, evolution is typically restricted to recombining fragments from only a few high-homology proteins (*6*).

Despite these challenges, immune system antibodies present a notable 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, in principle, counter 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 constitute some of the most structurally and functionally versatile enzymes and binding proteins in nature (*11*).

In this study, we ask whether enzymes could be generated from combinable fragments. We develop a method called CADENZ (combinatorial assembly and design of enzymes) to design and select protein fragments that, when freely combined, give rise to vast repertoires of low-energy proteins that exhibit high sequence and structural diversity. Isolating active enzymes in such vast protein libraries requires high-throughput screening methods (*12*, *13*) but can be readily and accurately achieved by using activity-based protein profiling (ABPP). ABPP uses mechanism-based covalent and irreversible inhibitors composed of a chemical scaffold that emulates structural features of the target substrate with an enzyme active site electrophile and a fluorophore or affinity tag. To exploit ABPP, we focused on glycoside hydrolase family 10 (GH10) xylanases (Enzyme Classification: 3.2.1.8) (*14*–*16*) 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 with more than 700 diverse backbones. We then trained a machine learning model to rank designs on the basis of their structure and energy features. Applying the learned model, we designed a second-generation library that demonstrated an order of magnitude increase in the success rate of 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, but loop regions should exhibit diverse conformations, including insertions and deletions (*17*–*19*). 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 (*20*, *21*). 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 through dual duplication of an ancestral β/α–β/α segment, suggesting that modern TIM-barrel enzymes can be segmented into four parts (Fig. 1A) (*22*–*24*). Nevertheless, during evolution, each protein accumulated mutations to adapt their intersegment 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 become stable and active (*25*). To address this problem, the CADENZ design objective is to compute a spanning set of backbone fragments that produce folded and active proteins when freely combined and without requiring further optimization. The primary challenge CADENZ addresses is designing mutually compatible (modular) fragments among which epistasis is minimal.

The first step of CADENZ is the alignment of homologous but structurally diverse enzymes (in the case of this study, 81 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) (*18*, *26*). 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 [Protein Data Bank (PDB) entry 3W24 (*27*)] to provide a realistic structural context and to promote compatibility between fragments. In practice, each fragment replaces the corresponding one in the template structure, and we used the PROSS stability-design algorithm (*28*) to implement stabilizing mutations within the fragment (8 to 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 (*29*, *30*), posing a challenge for modular design (Fig. 1C). To maintain catalytic activity, in all design calculations the side chains of four key catalytic amino acids are restrained to their crystallographically observed conformations. At the end of this process, we obtained a set of fragments that are internally stabilized within a common template and designed to support the catalytically competent constellation of active site residues.

However, because of epistasis, combining the designed fragments would likely result in mostly high-energy structures that are unlikely

[1]Department of Biomolecular Sciences, Weizmann Institute of Science, 7610001 Rehovot, Israel. [2]Leiden Institute of Chemistry, Leiden University, Einsteinweg 55, 2300 RA Leiden, 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
†Present address: DSM Nutritional Products Ltd, Wurmisweg 576, 4303 Kaiseraugst, Switzerland.

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

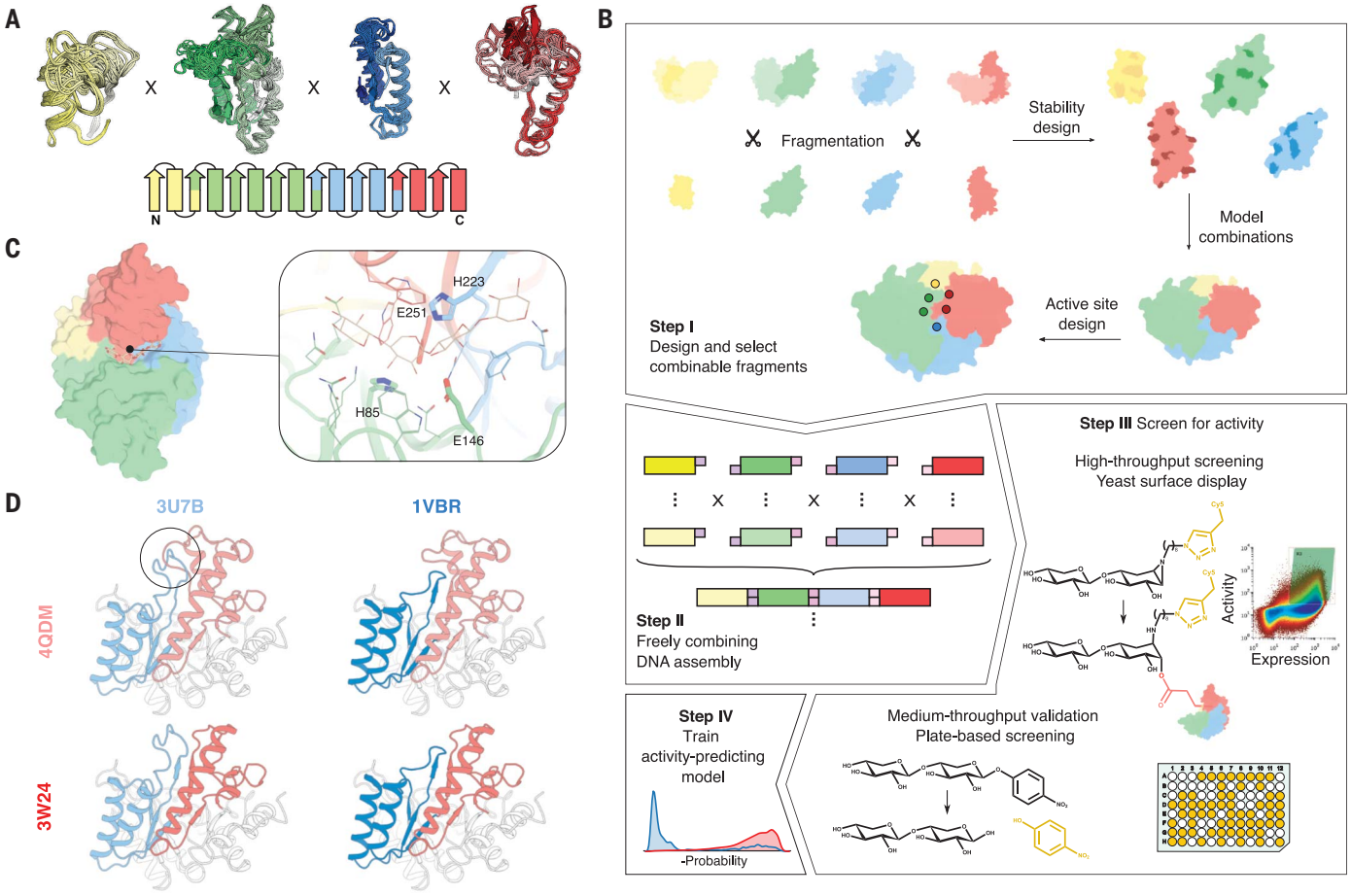

Fig. 1. Key steps in the CADENZ workflow. (**A**) (Top) Cartoon representation of selected fragments. (Bottom) Segmentation scheme for GH10 xylanases (color scheme is consistent in all structural figures). (**B**) The design pipeline. Step I: Design maximizes internal stability and compatibility with other fragments and diversifies active site positions that are not directly involved in the catalytic step. Step II: DNA oligos encoding fragments are freely ligated with Golden Gate Assembly (*32*) to generate DNA molecules encoding the full-length designs. Step III: Designs are sorted with a xylobiose-emulating activity-based probe (*34*) 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 on the basis of features that distinguish presumed active and inactive designs. (**C**) Four catalytic residues are restrained throughout design calculations (in sticks, numbering correspond 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 REU). The other designs exhibit low energies (≤ 950 REU).

to fold into their intended conformation or support the catalytic constellation (Fig. 1D). To address this problem, we enumerated all possible combinations of designed fragments and ranked them by Rosetta all-atom energy (Fig. 1B, step I). This process yields hundreds of thousands of distinct 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 of forming low-energy full-length structures (Fig. 2A). EpiNNet is trained to predict whether a combination of fragments exhibits favorable Rosetta energy on the basis of its constituent fragments. The trained network weights are then used to nominate fragments to generate the enzyme library. For the next design steps, we used the top six to seven fragments from each segment, which assembled into 1764 structures.

To add active site diversity and increase the chances of favorable fragment combination, we designed several sequence variants for each of the backbone fragments. We used the FuncLib design method (*31*) to generate low-energy amino acid constellations at positions in the active site and in the interfaces between β/α fragments while fixing the conformations of the key catalytic residues as observed in experimentally determined structures (Fig. 1C). We then used 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).

The CADENZ strategy does not necessarily select the lowest-energy fragment combinations, but rather mitigates the risk of combining incompatible ones. The consequences of intersegment epistasis are notable: Whereas the energies in the fully enumerated set of designed GH10s can be as high as +2500 Rosetta energy units (REU), after EpiNNet fragment selection, the energies are < –890 REU (Fig. 2B). As a reference, we also generated the distribution of energies obtained by combining the sequence of the fragments selected by EpiNNet before any of the design steps. This reference simulates the recombination of natural GH10 genes and exhibits a less favorable energy distribution than the combination of PROSS-stabilized fragments (> 100 REU difference, on average), underscoring the impact

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

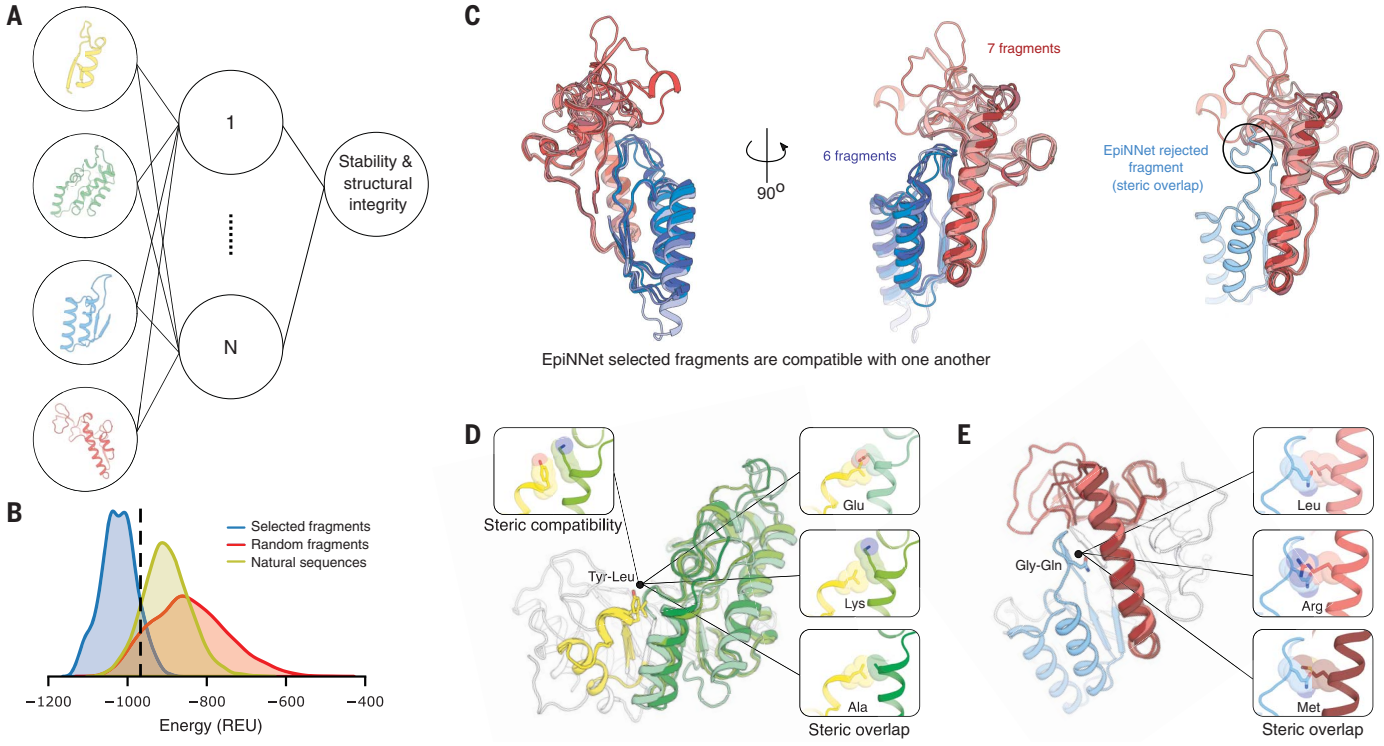

Fig. 2. EpiNNet selects fragments that assemble to low-energy structures. (**A**) Schematic representation of the EpiNNet architecture. (**B**) Most (89%) of the EpiNNet-selected designs exhibit low energy (< -967 REU, dashed line) (see materials and methods) relative to proteins generated by assembling randomly selected or natural fragments. (**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** and **E**) Examples of mutations selected by EpiNNet (taken from the second-generation library). (D) Segment 1 from PDB entry 3W24 (*27*) (yellow) faces segment 2 (green). EpiNNet prioritizes tyrosine (Tyr) over leucine (Leu) which cannot be accommodated with neighboring fragments. (E) Segment 3 from PDB entry 1VBR (*55*) (blue) faces segment 4 (red). EpiNNet prioritizes the small Gly over the large Gln.

of the design process (Fig. 2B and fig. S1B). Furthermore, EpiNNet alleviates intersegment epistasis by discarding backbone fragments and designed single-point mutations that are incompatible with neighboring segments (Fig. 2, C to E). This analysis highlights the challenge that epistasis poses for effective fragment combination while underscoring the strengths of the EpiNNet selection strategy. Although EpiNNet eliminates more than 60% of the fragments, the designed library exhibits high diversity and includes a total of 952,000 sequences adopting 1764 different backbones.

CADENZ generates thousands of structurally diverse and active enzymes

We used Golden Gate Assembly to combine designed fragments into full-length genes (Fig. 1B, step II) (*32*) and transformed the library into yeast cells for functional screening with cell-surface display (Fig. 1B, step III) (see materials and methods) (*33*). To probe enzyme activity, we incubated the library with a xylobiose ABP (*34*), which reacts within the enzyme active site to form a covalent and irreversible ester linkage with the glutamic acid

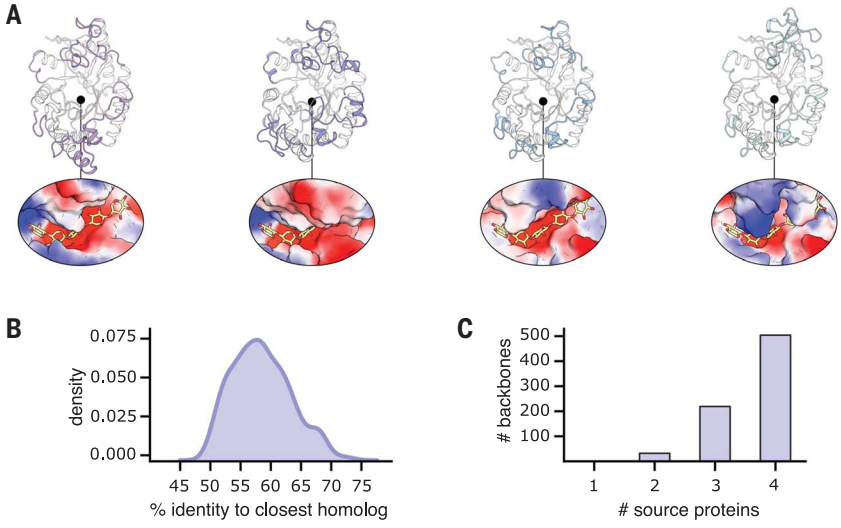

Fig. 3. CADENZ generates functional enzymes with high structure and sequence diversity. (**A**) Representative model structures of recovered enzymes designed by CADENZ. Regions that vary among the four designs are highlighted in colors (top). Active-site electrostatic potential surfaces of the representative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

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

nucleophile (*35*). 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 glutamic acid (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, ABP labeling acts as 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 and subsequent nucleophilic attack. To confirm that selected proteins exhibit the complete catalytic cycle (*36*), we transformed *Escherichia coli* cells with DNA from the sorted population and randomly selected 186 colonies for screening in 96-well plates with the chromogenic substrate 4-nitrophenyl β-xylobioside (O-PNPX_2) (*37*). Of the selected enzymes, 58% processed the substrate (fig. S3B), indicating that most designs selected by the ABP exhibited catalytic activity for this reaction.

We next applied single-molecule real-time (SMRT) long-read sequencing (*38*) to the sorted population. Encouragingly, sequencing showed that the sorted population included a large number of structurally diverse designs: specifically, 3114 distinct designs based on 756 different backbones (Fig. 3A), compared with only 376 GH10 xylanase entries in the UniProt database (*39*). The recovered designs exhibited many insertions and deletions relative to one another, with sequence lengths varying from 317 to 395 amino acids and 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 active-site pocket exhibits high geometric and electrostatic differences (Fig. 3A, bottom) because of loop conformation diversity. The designs exhibit as many as 169 mutations and 48 to 73% sequence identity (Fig. 3B) to their nearest natural homolog [in the nonredundant (nr) sequence database (*40*)], and most designs source fragments from four different structures (Fig. 3C).

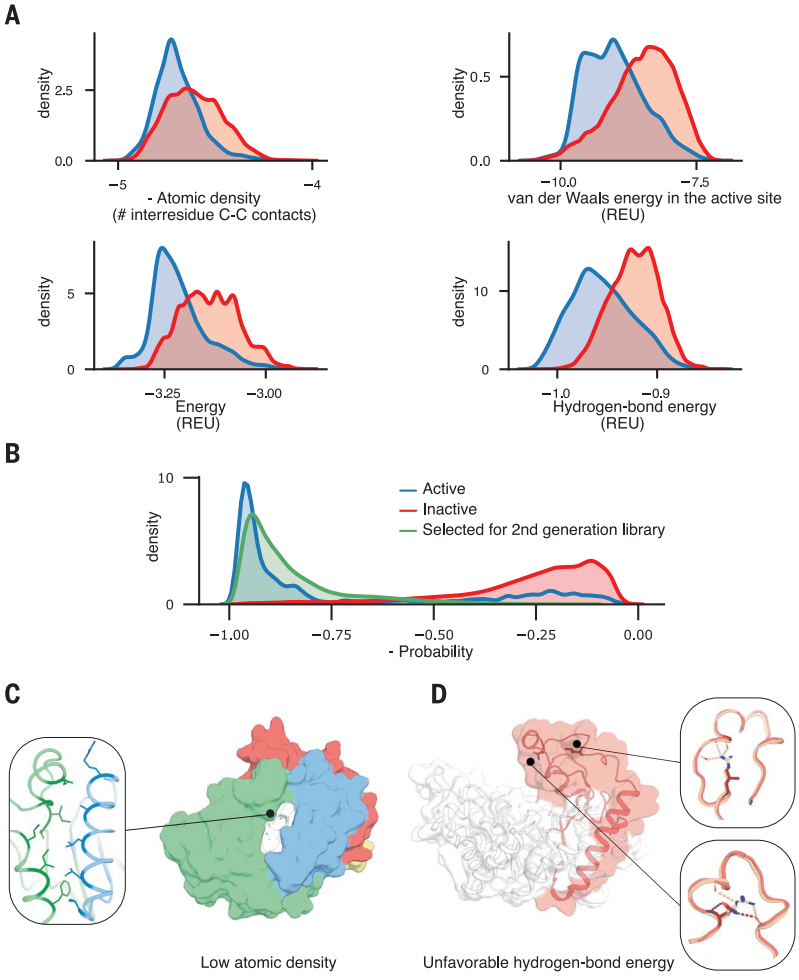

Fig. 4. Energy and structure features discriminate between active and inactive designs. (**A**) Representative features included in the activity predictor. The features show a statistically significant difference between presumed active and inactive designs with an independent two-sample *t* test, but no feature is individually an effective discriminator. All features are normalized by protein length. Low values are favorable. (**B**) Separation of designs recovered by ABPP versus other designs on the basis of a logistic regression model. In green, probability distribution for designs assembled by the fragments selected for the second-generation library. (**C** and **D**) Examples of backbone fragments eliminated by the activity predictor in the second-generation library. Fragment color scheme as in Fig. 1. (C) 2WYS in segment 2 (green) was selected for the first-generation library but discarded in the second because of 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 because of unfavorable hydrogen bond energy. Close inspection revealed two mutations introduced during sequence design, Arg^{282}→Asn and Arg^{289}→Leu [residue numbering refers to PDB entry: 1UQZ (*56*)], eliminating hydrogen bonds that are crucial for β/α loop backbone stabilization. Mutations in red.

Recovered designs are compact and preorganized for activity

The deep sequencing analysis provides a valuable 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 presumed active and inactive sets using an independent two-sample *t* test, finding that 63 metrics exhibited *P* values less than 10^{-10}. To select the most meaningful metrics, we visually inspected their distributions and focused on 10 (Fig. 4A and fig. S4) that were not significantly correlated (see materials and methods). We then trained a logistic regression model based on these 10 metrics to predict whether an enzyme is active (Fig. 4B and tables S1 to S3).

The 10 dominant predictive metrics 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 (*41*). 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 prevalence of

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

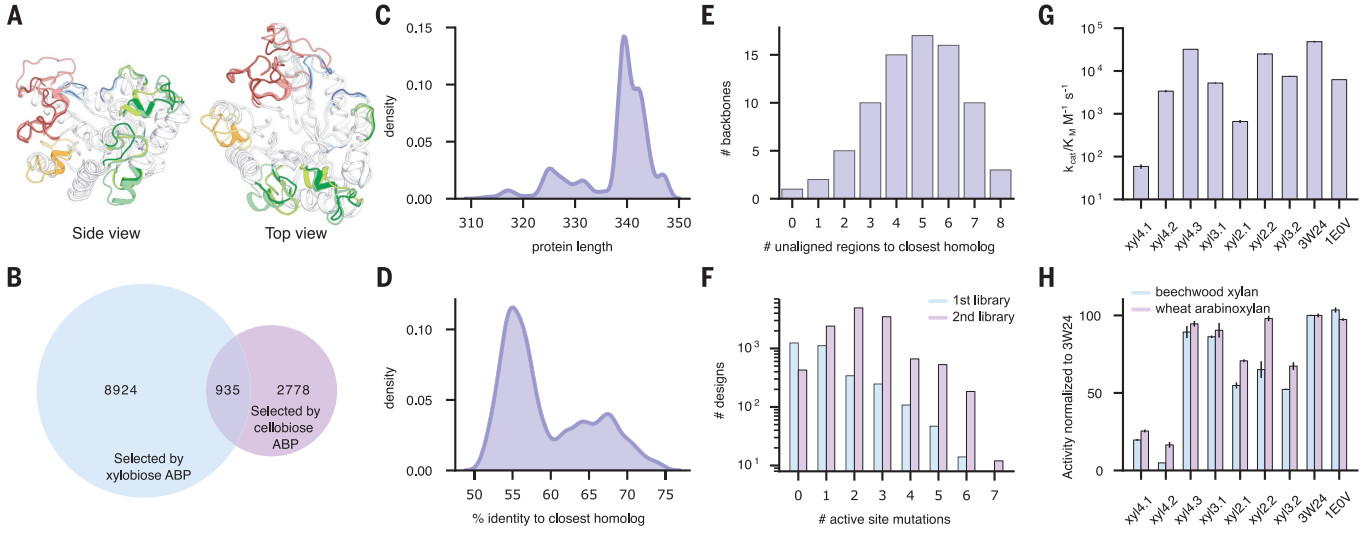

Fig. 5. Activity predictor significantly increases design success rate. (**A**) Backbone fragments selected for the second-generation library (colors as in Fig. 1, low diversity regions in white). (**B**) Number of sequences in the population selected by the xylobiose and cellobiose ABPs (blue and purple, respectively). In the overlap region, designs selected by both ABPs. (**C**) Protein sequence length of recovered designs in the second-generation library. (**D**) Distribution of sequence identity to nearest natural homolog of recovered designs in the second-generation library. (**E**) Number of unaligned regions to nearest natural homolog of recovered designs in the second-generation library. (**F**) Designs recovered 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 (*27*) 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.

buried long-range hydrogen bond networks in large proteins of a complex fold such as TIM barrels (fig. S5) (*42*, *43*). However, Rosetta system energy makes a small contribution to predicting activity, presumably because 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 in promoting active site preorganization. Focusing on the four catalytic residues only, the model includes a feature that penalizes high repulsive energy, further emphasizing 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 (*44*, *45*). Additionally, the design objective function is substantially different within the active site versus the remainder of the protein.

Recently, the AlphaFold2 ab initio structure prediction method (*46*) has been shown to discriminate correctly from incorrectly folded de novo–designed binders (*47*). However, when AlphaFold2 was applied to our set, no discernable difference was found between presumed 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 the designs, CADENZ generates sequences with native-like characteristics.

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 instead of ranking them on the basis of Rosetta energies, we ranked them according to the activity predictor (Fig. 4B and fig. S2B). We then 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, holding sidechain conformations of the core catalytic residues fixed in all design calculations. This second-generation library included three backbone fragments for each of the four segments and up to 11 sequence variants per fragment (for a total of 100 designed fragments), resulting in 334,125 designed full-length xylanases with 81 different backbones. To gain insight into the molecular features that are disfavored by the activity predictor, we analyzed which backbone fragments were chosen in the first library but discarded in the second. We found, for example, that atomic density (Fig. 4C) and hydrogen-bond energy (Fig. 4D) were unfavorable in many discarded fragments.

We synthesized and screened the second-generation library as we did with the first-generation library (fig. S7). Notably, sequencing confirmed 9859 active designs, an order-of-magnitude increase in the rate of positive hits compared with that of the first library (Fig. 5A). In addition to the xylobiose screen, we also screened the library with an ABP that is based on cellobiose (Fig. 5B), the disaccharide repeat moiety in cellulose rather than in xylan (*48*). We found 2778 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 design population). To verify that the ABP-selected designs exhibited the full catalytic cycle, we used plate-based validation with O-PNPX_2 and cellPNP (to detect cellulase activity) confirming 85 and 60% of active clones in the xylobiose- and cellobiose-labeled populations, respectively (fig. S7, C and D).

Ranking designs on the basis of the activity predictor resulted in a more focused library that included 79 of the 81 designed backbones and sequence lengths ranging from 312 to 347 amino acids (Fig. 5C). Although the activity predictor was blind to the identities of the designed fragments and mutations, we were

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

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% of these designs incorporated 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 they were in the first library, exhibiting 50 to 73% sequence identity to the most similar sequences in the nr database (*40*) (Fig. 5D) as well as 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 the second-generation library. The discrimination that we observed, however, was lower than in the first-generation 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 (*18*). On the basis of the AbDesign strategy (*26*), we used Rosetta atomistic modeling to enumerate all fragment combinations and design their sequences as full-length enzymes, followed by selection with the activity predictor. This strategy encodes more stabilizing intersegment interactions than when the fragments are designed independently, and the designs are therefore more likely to be stable and foldable. Thus, in this implementation, the design and selection process does not favor modularity but rather optimal structure and energy properties. The activity predictor selected 27 designs for experimental characterization with up to 143 mutations and 51 to 74% sequence identity to their nearest natural homologs. Although these designs were generated by a different process than the one used to train the activity predictor, notably, 25 (93%) of the designs were active in hydrolyzing O-PNPX_2 (table S4), compared with less than 50% in a previous application of AbDesign to GH10 enzymes (*18*). 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 those of natural GH10 xylanases from thermophiles, including against natural wood and wheat xylan (Fig. 5, G and H), despite incorporating over 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, for either high- or low-throughput screening.

Discussion

Modularity is a prerequisite for innovation in numerous engineering disciplines, but protein domains exhibit high epistasis, severely hampering the ability to combine fragments into stable and active structures. 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. We have also begun to investigate how EpiNNet can be implemented to design large repertoires of active site sequence variants (*49*). Our approach increases the number and diversity of functional enzymes that can be interrogated relative to the natural diversity, providing an alternative to metagenomic libraries (*12*). Current methods for optimizing and diversifying proteins rely on sequence statistics (*50*, *51*) or cycles of mutation, recombination, and screening (*52*). Because of high epistasis, these methods explore a small fraction of sequence and structure space, whereas we show in this system that CADENZ can generate 10^6 structurally diverse designs of which >10,000 are recovered on the basis of activity.

Our results also show that ABPP is an effective strategy for high-throughput isolation of successful CADENZ designs and could be extended to other substrates (*53*), either natural or engineered. The combined strategy of CADENZ and ABPP enabled us to implement effective design-test-learn cycles on many enzyme designs that have previously led to deeper understanding of the design principles for de novo–designed miniproteins (*44*, *45*). The rules we learned increased the design success rate by an order of magnitude and were directly transferable to automated small-scale design. Such functional data from many homologous yet structurally diverse enzymes may guide future improvements in macromolecular energy functions and advance efforts to develop AI-based enzyme design methods.

REFERENCES AND NOTES

1. M. E. Csete, J. C. Doyle, *Science* **295**, 1664–1669 (2002).
2. Y. Lazebnik, *Cancer Cell* **2**, 179–182 (2002).
3. M. Parter, N. Kashtan, U. Alon, *PLOS Comput. Biol.* **4**, e1000206 (2008).
4. D. M. Weinreich, R. A. Watson, L. Chao, *Evolution* **59**, 1165–1174 (2005).
5. D. A. Kondrashov, F. A. Kondrashov, *Trends Genet.* **31**, 24–33 (2015).
6. C. A. Voigt, C. Martinez, Z.-G. Wang, S. L. Mayo, F. H. Arnold, *Nat. Struct. Biol.* **9**, 553–558 (2002).
7. S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11691–11696 (2014).
8. N. Ferruz *et al.*, *J. Mol. Biol.* **432**, 3898–3914 (2020).
9. T. M. Jacobs *et al.*, *Science* **352**, 687–690 (2016).
10. K. Murphy, C. Weaver, *Janeway's Immunobiology* (Garland Science, 2016).
11. E. Dellus-Gur, A. Toth-Petroczy, M. Elias, D. S. Tawfik, *J. Mol. Biol.* **425**, 2609–2621 (2013).
12. P.-Y. Colin *et al.*, *Nat. Commun.* **6**, 10008 (2015).
13. P. Kast, M. Asif-Ullah, N. Jiang, D. Hilvert, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5043–5048 (1996).
14. S. G. Withers *et al.*, *Biochem. Biophys. Res. Commun.* **139**, 487–494 (1986).
15. E. Drula *et al.*, *Nucleic Acids Res.* **50** (D1), D571–D577 (2022).
16. CAZypedia Consortium, *Glycobiology* **28**, 3–8 (2018).
17. A. Tóth-Petróczy, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **26**, 131–138 (2014).
18. G. Lapidoth *et al.*, *Nat. Commun.* **9**, 2780 (2018).
19. R. Netzer *et al.*, *Nat. Commun.* **9**, 5286 (2018).
20. N. Nagano, C. A. Orengo, J. M. Thornton, *J. Mol. Biol.* **321**, 741–765 (2002).
21. R. Sterner, B. Höcker, *Chem. Rev.* **105**, 4038–4055 (2005).
22. B. Höcker, J. Claren, R. Sterner, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16448–16453 (2004).
23. B. Höcker, S. Beismann-Driemeyer, S. Hettwer, A. Lustig, R. Sterner, *Nat. Struct. Biol.* **8**, 32–36 (2001).
24. M. Richter *et al.*, *J. Mol. Biol.* **398**, 763–773 (2010).
25. D. L. Trudeau, M. A. Smith, F. H. Arnold, *Curr. Opin. Chem. Biol.* **17**, 902–909 (2013).
26. R. Lipsh-Sokolik, D. Listov, S. J. Fleishman, *Protein Sci.* **30**, 151–159 (2021).
27. X. Han *et al.*, *Proteins* **81**, 1256–1265 (2013).
28. A. Goldenzweig *et al.*, *Mol. Cell* **63**, 337–346 (2016).
29. S. R. Andrews *et al.*, *J. Biol. Chem.* **275**, 23027–23033 (2000).
30. B. Henrissat *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7090–7094 (1995).
31. O. Khersonsky *et al.*, *Mol. Cell* **72**, 178–186.e5 (2018).
32. C. Engler, R. Kandzia, S. Marillonnet, *PLOS ONE* **3**, e3647 (2008).
33. G. Chao *et al.*, *Nat. Protoc.* **1**, 755–768 (2006).
34. S. P. Schröder *et al.*, *ACS Cent. Sci.* **5**, 1067–1078 (2019).
35. D. Tull *et al.*, *J. Biol. Chem.* **266**, 15621–15625 (1991).
36. D. Tull, S. G. Withers, *Biochemistry* **33**, 6363–6370 (1994).
37. H. Taguchi, T. Hamasaki, T. Akamatsu, H. Okada, *Biosci. Biotechnol. Biochem.* **60**, 983–985 (1996).
38. A. Rhoads, K. F. Au, *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
39. A. Bateman *et al.*, *Nucleic Acids Res.* **49**, D480–D489 (2021).
40. E. W. Sayers *et al.*, *Nucleic Acids Res.* **50** (D1), D20–D26 (2022).
41. A. Goldenzweig, S. J. Fleishman, *Annu. Rev. Biochem.* **87**, 105–129 (2018).
42. O. Khersonsky, S. J. Fleishman, *BioDesign Research* **2022**, 1–11 (2022).
43. D. Baran *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10900–10905 (2017).
44. G. J. Rocklin *et al.*, *Science* **357**, 168–175 (2017).
45. L. Cao *et al.*, *Nature* **605**, 551–560 (2022).
46. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).
47. N. Bennett *et al.*, *bioRxiv* (2022), p. 2022.06.15.495993.
48. C. de Boer *et al.*, *RSC Chem. Biol.* **1**, 148–155 (2020).
49. J. Y. Weinstein *et al.*, *bioRxiv* 2022.10.11.511732 [Preprint] (2022). https://doi.org/10.1101/2022.10.11.511732.
50. W. P. Russ *et al.*, *Science* **369**, 440–445 (2020).
51. N. Ferruz, S. Schmidt, B. Höcker, *Nat. Commun.* **13**, 4348 (2022).
52. S. L. Lovelock *et al.*, *Nature* **606**, 49–58 (2022).
53. L. Wu *et al.*, *Curr. Opin. Chem. Biol.* **53**, 25–36 (2019).
54. R. Lipsh-Sokolik *et al.*, Data for: Combinatorial assembly and design of enzymes (2022); http://dx.doi.org/10.5281/zenodo.7382421.
55. T. Ihsanawati *et al.*, *Proteins* **61**, 999–1009 (2005).

ACKNOWLEDGMENTS

We thank N. London, B. Höcker, S. Barber-Zucker, and D. Listov for discussions and S. Warzsawski and K. Goldin for technical help. R.L.-S. is supported by a fellowship from the Arianne de Rothschild Women Doctoral Program. **Funding:** This work was funded by the Volkswagen Foundation grant 94747 (S.J.F.), the Israel Science Foundation grant 1844 (S.J.F.), the European Research Council through a Consolidator Award grant 815379 (S.J.F.), the Dr. Barry Sherman Institute for Medicinal Chemistry (S.J.F.), a donation in memory of Sam Switzer (S.J.F.), the Royal Society for the Ken Murray Research Professorship

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025

Corrected 8 August 2023. See full text.

(G.J.D.), the European Research Council grant ERC-2011-AdG-290836 'Chembiosphing' (H.S.O.), ERC-2020-SyG-951231 'Carbocentre' (H.S.O. and G.J.D.), and the Netherlands Organization for Scientific Research through the NWO TOP grant 2018-714.018.002 "Endoglycoprobe" (H.S.O.). **Author contributions:** Conceptualization: R.L.S., O.K., S.J.F.; Methodology: R.L.S., O.K., S.J.F.; Software: R.L.S.; Validation: R.L.S., O.K.; Formal analysis: R.L.S., S.Y.H.; Investigation: R.L.S., O.K., S.P.S., C.d.B.; Resources: R.L.S., O.K., S.P.S., C.d.B., H.S.O., S.J.F.; Data Curation: R.L.S.; Writing: R.L.S., S.J.F., H.S.O., G.J.D.; Visualization: R.L.S., O.K.; Supervision: S.J.F., H.S.O.; Project administration: S.J.F.; Funding acquisition: S.J.F., H.S.O., G.J.D. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Custom Python scripts, RosettaScripts, commandlines, Jupyter notebooks, and datasets are available at Zenodo (*54*). License information: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ade9434
Materials and Methods
Figs. S1 to S7
Tables S1 to S5
References (*56*–*70*)

View/request a protocol for this paper from *Bio-protocol*.

Submitted 19 September 2022; accepted 12 December 2022
10.1126/science.ade9434

Downloaded from https://www.science.org at University Van Amsterdam on April 13, 2025