EVOLUTIONARY BIOLOGY

Ancestral reconstruction of polyethylene terephthalate degrading cutinases reveals a rugged and unexplored sequence-fitness landscape

Vanessa Vongsouthi[1,2]*†, Rosemary Georgelin[1,2,3]†, Dana S. Matthews[1,2,3]†, Jake Saunders[1], Brendon M. Lee[1], Jennifer Ton[2], Adam M. Damry[1], Rebecca L. Frkic[1,3], Matthew A. Spence[1,2,3]*, Colin J. Jackson[1,3,4]*

The use of protein engineering to generate enzymes for the degradation of polyethylene terephthalate (PET) is a promising route for plastic recycling, yet traditional engineering approaches often fail to explore protein sequence space for optimal enzymes. In this work, we use multiplexed ancestral sequence reconstruction (mASR) to address this, exploring the evolutionary sequence space of PET-degrading cutinases. Using 20 statistically equivalent phylogenies of the bacterial cutinase family, we generated 48 ancestral sequences revealing a wide range of PETase activities, highlighting the value of mASR in uncovering functional variants. Our findings show PETase activity can evolve through multiple pathways involving mutations remote from the active site. Moreover, analyzing the PETase fitness landscape with local ancestral sequence embedding (LASE) revealed that LASE can capture sequence features linked to PETase activity. This work highlights mASR's potential in exploration of sequence space and underscores the use of LASE in readily mapping the protein fitness landscapes.

INTRODUCTION

More than 350 million tonnes of plastic are produced annually, with the majority being derived from petrochemicals (*1*). However, the current global recycling rate of plastics is estimated to be less than 10%, partially due to the technological constraints of mechanical (melt-extrusion) recycling (*2*). There is a pressing need to develop advanced and scalable recycling methods that will enable the transition toward a circular plastic economy. In the past decade, enzymatic depolymerization of plastics has emerged as one such advanced recycling method (*3*). The enzymatic degradation of polyethylene terephthalate (PET), a versatile thermoplastic commonly found in food and beverage packaging, and polyester textiles has received particular interest (*3*).

PET-degrading enzymes that have been characterized to date—including those classified as cutinases (*4*–*6*) (EC 3.1.1.74), lipases (*7*) (EC 3.1.1.3), and carboxylesterases (*8*) (EC 3.1.1.1)—belong to the esterase subclass (EC 3.1) and are characterized by a catalytic triad (Ser-His-Asp) typical of the α/β hydrolase fold superfamily. Cutinases have garnered substantial interest due to their ability to hydrolyse both aromatic and aliphatic polyesters (*9*). Cutinases of bacterial (*5*, *9*, *10*), fungal (*7*), and metagenomic (*11*–*13*) origins have been studied for their PETase activity, including Thc_Cut1 and Thc_Cut from *Thermobifida cellulosilytica* (*5*), TfCut2 from *Thermobifida fusca* (*5*), HiC from *Humicola insolens* (*14*), FsC from *Fusarium solani pisi* (*15*), and leaf-branch compost cutinase (LCC) from leaf-branch compost metagenome (*16*). However, extant cutinases are often unsuitable for direct use in industrial processes, necessitating optimization to improve properties such as catalytic efficiency (*17*), stability under harsh conditions (*18*, *19*), and to alleviate product inhibition (*20*). For example, active site optimization through rational design (*6*, *10*, *21*) and evolutionarily guided engineering (*20*) has yielded substantial improvements in PET hydrolysis across various cutinase backgrounds. While structure-guided protein engineering has been effective in improving the efficiency of extant PET hydrolases (PETases), it has provided little insight on the mechanisms by which PETase activity emerges or has been optimized by evolution.

Ancestral sequence reconstruction (ASR)—which uses a phylogenetic tree and a statistical model of evolution to infer the sequences of extinct, ancestral proteins—can provide critical insights into molecular evolution and functional diversification within protein families (*22*). ASR places sequences within the context of an evolutionary hypothesis and reconstructs residues according to epistatic constraints that are conserved within phylogenetic lineages (*23*). Sequences reconstructed with ASR are therefore an effective exploration of evolutionarily accessible sequence space. This has made ASR useful in studying sequence-function relationships over protein families (*24*, *25*); this insight can shed light on the topologies of fitness landscapes that dictate the adaptive potential of proteins (*26*). For example, ASR can give insight into the ruggedness (i.e., complexity) of a fitness landscape over large spans of evolutionarily accessible sequence space to reveal fitness peaks that are inaccessible through stepwise mutational approaches (*25*). In addition to this, ASR is also frequently used to engineer enzymes with enhanced industrial properties (*22*, *27*–*29*), making it a valuable tool in both protein engineering and understanding the mechanisms of molecular evolution.

More recently, protein representation learning has become a widely used method in protein engineering and evolutionary inference (*30*, *31*). Protein language models (PLMs), which are deep neural networks trained to predict the identities of masked residues in a corpus of protein sequences (*30*–*33*), can map information sparse and high dimensional protein sequences to fixed-length vector

[1]Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. [2]Samsara Eco, Sydney, NSW 2065, Australia. [3]ARC Centre of Excellence for Innovations in Peptide & Protein Science, Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. [4]ARC Centre of Excellence for Innovations in Synthetic Biology, Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia.
*Corresponding author. Email: vanessa.vongsouthi@samsaraeco.com (V.V.); colin.jackson@anu.edu.au (C.J.J.); matthew.spence@anu.edu.au (M.A.S.)
†These authors contributed equally to this work.

representations. These vector representations capture the evolutionary and biophysical features of protein sequences in the representation model's latent embedding space (*30*, *31*). PLMs have been used to learn the structure of fitness landscapes (*34*–*36*) and can learn evolutionary features when trained with ancestral reconstructed sequence data (*37*).

In this study, we applied ASR to explore the evolutionarily accessible sequence space of PET-degrading cutinases. Using a dataset of 397 extant cutinase sequences with significant homology to known cutinases with PETase activity, we adopted a multiplexed ASR (mASR) approach to generate a diverse library of ancestral cutinase sequences. Through experimental characterization of 48 ancestral cutinases, we identified a broad range of PETase activities, including between equivalent nodes on distinct yet statistically indifferent phylogenetic topologies. These findings highlight the importance of sampling diverse phylogenetic backgrounds to uncover functional ancestral variants. Furthermore, our study analyzed the topology of the PETase fitness landscape through two sequence embedding schemes: one-hot encoding (OHE) and the more recently described local ancestral sequence embedding (LASE) (*37*). We found that LASE was more effective in capturing cutinase sequence features pertinent to PETase activity, demonstrating a clear pattern of iterative improvement in PETase functionality throughout the sequence exploration phases of our study. This comprehensive approach not only highlights the utility of mASR in uncovering previously unknown PETase variants but also emphasizes the role of advanced embedding techniques in mapping PETase fitness landscapes.

RESULTS

mASR of PETases from diverse phylogenetic backgrounds

We used ASR to explore the PETase functional sequence space of the cutinase family. To maximize diversity in the local sequence space around known PETases, such as *Tf*Cut2 and LCC, we used the recently described mASR (*37*). In brief, mASR samples multiple statistically indifferent phylogenetic backgrounds from which to reconstruct ancestral sequences from. This produces diverse libraries of ancestral proteins that span functional sequence space over a distribution of realistic phylogenies. To achieve this, we performed 20 replicates of maximum likelihood (ML) phylogenetic inference and ASR on a single dataset of 397 extant cutinases with significant homology to *Tf*cut2 and LCC (E value $\leq 1 \times 10^{-5}$). Consistent with previous phylogenetic studies (*38*), PET hydrolytic cutinases were resolved as a polyphyletic group of two monophyletic lineages: the *Thermobifida* sp. PETases, which include *Tf*cut2, *Thermobifida cellulosytica* cutinase, and *Thermobifida alba* esterase 1 and the LCC-like PETases that include LCC and BhrPETase. This topology was resolved consistently over all 20 phylogenetic priors used for mASR (Fig. 1A). The placement of the PETase clades were supported by high ultra-fast bootstrap approximations (*39*) (≥0.95), and the ancestral nodes separating these groups were reconstructed in CodeML from the phylogenetic analysis by maximum likelihood (PAML) (*40*) suite with relatively high mean posterior probability [>0.9; based on ASR statistical benchmarking studies (*41*)]. The resulting ancestral sequence library comprised approximately 1600 unique sequences that encompassed the evolutionarily accessible sequence space of the bacterial cutinase family.

We selected 48 ancestral nodes from 20 different trees with homology to the most recent common ancestor of LCC and *Tf*Cut2 for experimental characterization and comparison to extant cutinases. Specifically, these nodes were chosen from the recent ancestors of the LCC and *Tf*Cut2 lineages (or the most recent common ancestor of both), with at least a single sequence sampled from each of the 20 distinct phylogenetic trees. Each ancestral sequence was selected as the maximum a posteriori (MAP) sequence that maximizes the posterior probability over the full length of the reconstructed protein. The variants were expressed, purified, and tested for PETase activity. The soluble expression level of each variant was measured using the Bradford assay (fig. S1). PETase activity against amorphous PET film was measured by ultraviolet (UV) absorbance at 260 nm ($A_{260\ nm}$) to detect the soluble products of PET hydrolysis after the removal of undigested film, including terephthalic acid (TPA), mono(2-hydroxyethyl) terephthalate (MHET), and bis(2-hydroxyethyl) terephthalate (BHET) (fig. S2). Among the extant cutinases, the highest PETase activities were observed for LCC ($A_{260\ nm} = 0.34 \pm 0.03$), BhrPETase from bacterium HR29 ($A_{260\ nm} = 0.49 \pm 0.01$), and a previously engineered variant of LCC with the mutations F243I/D238C/S283C/Y127G (LCC ICCG) ($A_{260\ nm} = 0.54 \pm 0.07$) (*6*), with other cutinases showing lower activity (Fig. 1E). Of the 48 ancestral variants, we observed PETase activity for a number of ancestral cutinases from a diverse range of trees (Fig. 1F). Several variants (1_533, 35_442, 55_547, 70_439, 70_440, and 85_456) exhibited similar activity to the most active extant cutinases, LCC and BhrPETase. From this group, we selected ancestors from tree 35, node 442 (35_442; $A_{260\ nm} = 0.37 \pm 0.06$) and tree 55, node 547 (55_547; $A_{260\ nm} = 0.59 \pm 0.09$), for further investigation. Notably, these two ancestors belonged to equivalent positions from two independent phylogenetic trees (Fig. 1, B to D), representing the most recent common ancestor of LCC and TfCut2 and sharing 98.5% sequence identity to one another.

Within the dataset of 48 ancestral sequences, 12 represent the same phylogenetic node (the most recent common ancestor of LCC and *Tf*Cut2) over 12 different phylogenetic backgrounds. Of these, 9 exhibited PETase activity, while 3 were inactive on PET. The PETase ancestor 55_547 and the inactive ancestor 10_543 both represent equivalent positions (the most recent common ancestor of *Tf*Cut2 and LCC) in their respective phylogenetic backgrounds and differ by only 9 of 261 positions, yet 55_547 has activity against PET and 10_543 does not. The functional differences between these sequences arise solely from the topology of the phylogenetic tree used to reconstruct them and markedly alter how the evolution of PETase activity could be inferred. For example, in the case of the 55_547 phylogeny, PETase activity appears to be a promiscuous ancestral trait that existed before the discovery of extant PETases (such as LCC and *Thermobifida* cutinases), whereas the 10_543 phylogeny supports the contradictory hypothesis that PETase activity emerged independently in LCC-like and *Thermobifida* PETase lineages from an ancestor without PETase activity. As both topologies failed rejection by the approximately unbiased (AU) (*42*) test and are statistically indifferent at representing the observed alignment data, it is not possible to reject one of these hypotheses purely from a phylogenetic perspective. We also observe no correlation between the mean posterior probability of an ancestor and its activity (fig. S5). This observation is somewhat counterintuitive, as the mean posterior probability of a sequence, which is the statistical confidence in the identity of a reconstructed ancestor, is often used to discriminate between poorly reconstructed sequences (and hence likely to be less fit) and those that are likely to be functional and fit (*41*). Together,

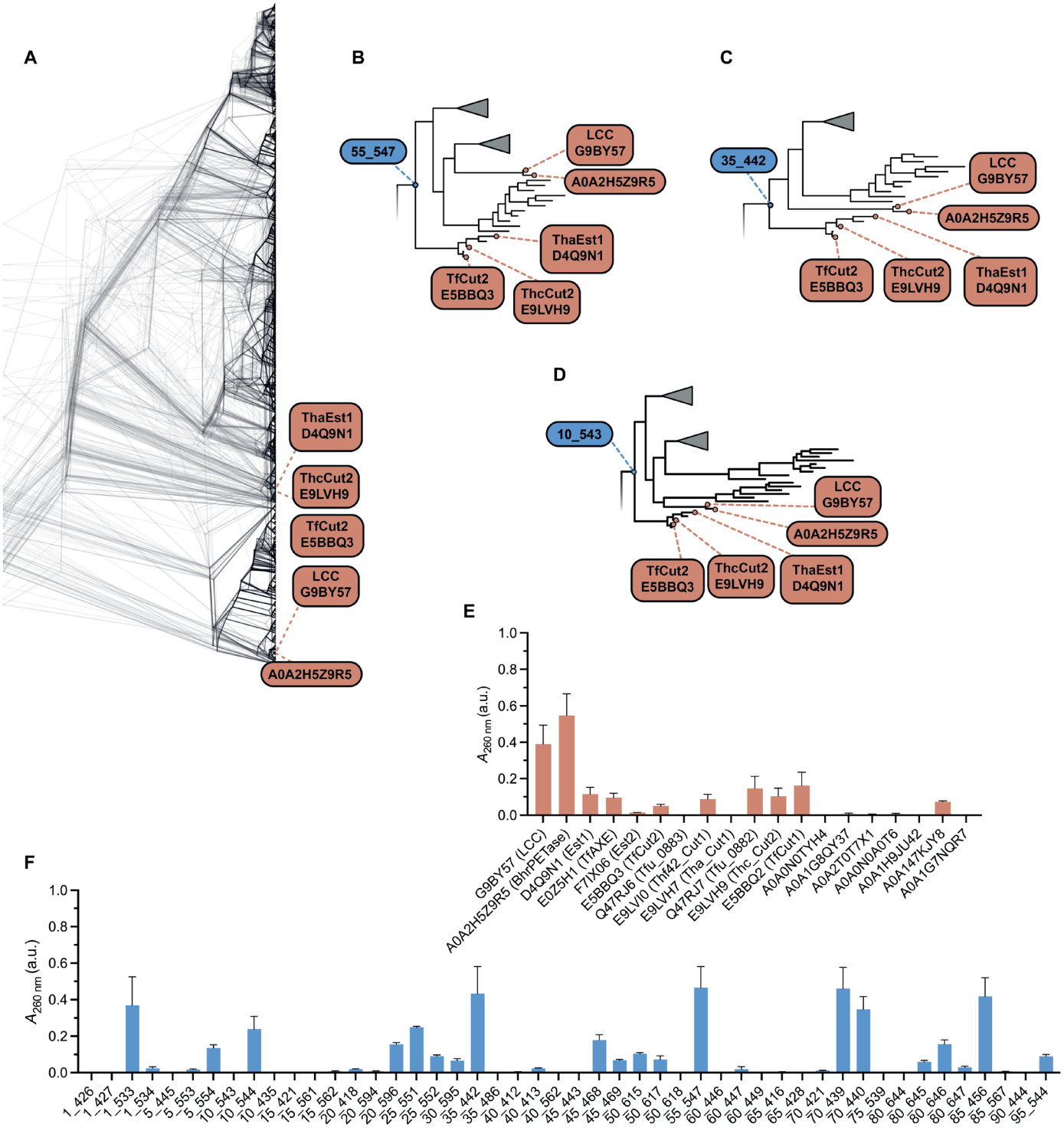

Fig. 1. ASR and characterization of ancestral cutinases. (**A**) Twenty replicates of phylogenetic reconstruction of the cutinase family with fixed extant node locations. All 20 presented topologies failed rejection by the AU test and are equally valid representations of the underlying sequence alignment. Extant tips of interest [*Thermobifida* cutinases, E5BBQ3, D4Q9N1, and E9LVH9 and metagenomic assembled cutinases A0A2H5ZR95 and G9B757 (LCC)] are labeled. (**B**) Phylogenetic tree 55, (**C**) 35, and (**D**) 10, with ancestors 35_442, 55_547, and 10_543 highlighted. Each ancestor belongs to equivalent nodes from independent phylogenetic topologies and is the most recent common ancestor of LCC and TfCut2, extant cutinases with known PETase activity. (**E**) Twenty extant and (**F**) 48 ancestral cutinases. The bulk soluble products of PET hydrolysis (TPA, MHET, and BHET) are measured by $A_{260\,nm}$ after 16 hours incubation with the enzyme at 60°C. Data are represented as the means ± SEM ($n = 3$).

these results highlight the importance of sampling diverse phylogenetic backgrounds during ASR when accessing the functional sequence space of a protein family as minor phylogenetic incongruencies can translate into substantial functional differences in (equivalent) reconstructed ancestral sequences.

Structural characterization and analysis of ancestral cutinases

A comparative structural and sequence analysis of inactive and active ancestors suggests that PET hydrolysis appears to be a trait that is evolutionarily accessible from within the cutinase background. Unlike the *Ideonella sakaiensis* PETase, which emerged through conformational optimization of the first and second shells via obvious selection within the active site (*43*–*45*), PETase activity in the TfCut2 and LCC PETase lineages appears to emerge through nonspecific and likely neutral mutations that are often distal to the active site (Fig. 2, A to C). Analysis of ancestral cutinase sequences reveals virtually no differences in the PET binding sites between PETase active and inactive variants (fig. S6). Furthermore, mutations associated with a gain of function vary between the different phylogenetic trees used to reconstruct the ancestral sequences. For example, ancestors 55_547 and 35_442 are each separated by nine unique mutations from their closest relatives without PETase activity (ancestors 10_543 with A15S, A36V, T49S, T92S, N109D, R114S, N145R, I180V, S226A and 75_539 with A36V, T49D, T92S, M105Q, R114S, S124N, R167T, P199S, and A226S, respectively). Of these loss-of-function mutations, three are shared between either background, five are unique, and one is a reversion (S226A for 55_547 and A226S for 35_442). Intuitively, a combination (with at least one) of these mutations is required to be fixed in either respective cutinase background to impart PETase activity. Similarly, ancestor 10_543 is separated by eight mutations from its closest relative with PETase activity (70_439 with E28Q, S49D, S124N, R145N, V180I, P199S, S215T, and A226S); of the eight mutations associated with gain of function in the background of 10_543, only three are specifically shared with gain-of-function mutations in 55_547. Nearly all functionally consequential

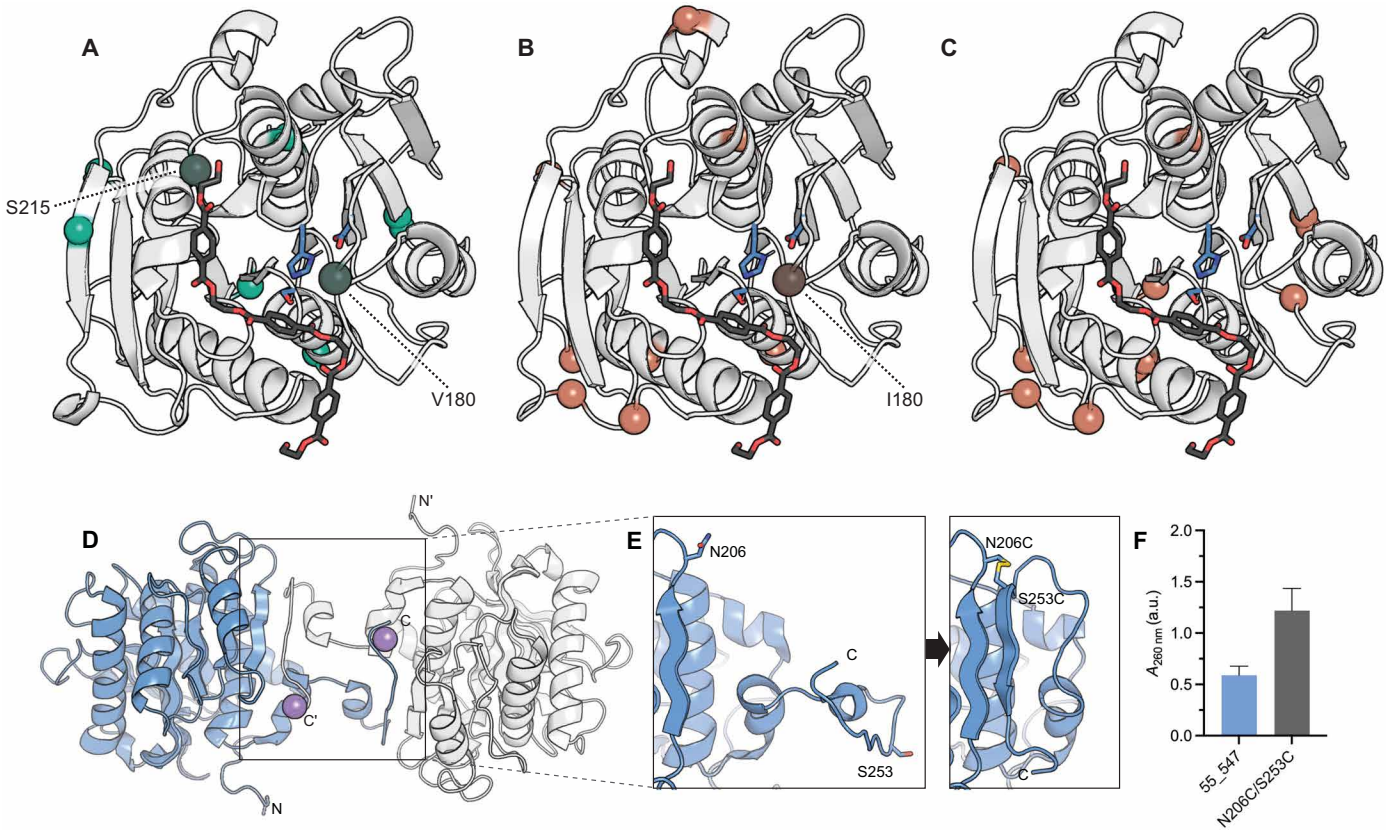

Fig. 2. Structural characterization of 55_547 and 35_442. (**A**) Positions of gain-of-function (GOF) mutations in the inactive ancestor 10_543, and (**B**) loss-of-function (LOF) mutations in the active ancestors 55_547 and (**C**) 35_442. GOF and LOF mutations are shown as teal and red spheres, respectively. AlphaFold (*67*, *68*) was used to generate models of each ancestor. Most GOF/LOF mutations are distal from the catalytic site residues (blue) and the putative PET binding site highlighted by the docked 2HE-(MHET)3 ligand (black). Mutations within 5 Å of the active site residues (highlighted in blue) are labeled (S215, V180, and I180) and shaded in a darker teal/red. Structural analysis of the active site residues H210, S132, and D178 in ancestors 10_543, 55_547, and 35_442 shows uniform alignment, indicating that PETase activity does not arise from modifications within the active sites (fig. S4) (**D**) Domain-swapped homodimer observed in the crystal structure of ancestor 55_547 with two C-terminal strands swapped (PDB 8ETX). The N and C termini of each subunit are indicated. The dimer is coordinated by sodium ions (represented as purple spheres) in the crystal lattice. (**E**) Positions on the flexible C terminus of ancestor 55_547 targeted for disulfide engineering, as shown in the crystal structure (left; PDB: 8ETX), resulting in the disulfide mutant N206C/S253C, as generated by AlphaFold (*67*, *68*) (right). The predicted formation of an additional β strand at the C terminus, constrained by the disulfide bond, is shown. (**F**) PETase activity of ancestor 55_547 and the disulfide mutant N206C/S253C. The bulk soluble products of PET hydrolysis (TPA, MHET, and BHET) are measured by $A_{260\ nm}$ after 16 hours incubation with the enzyme at 60°C. Data are represented as the means ± SEM ($n = 3$).

mutations across backgrounds are distal to the active site and PET binding pocket. This suggests that (i) there are numerous (and diverse) molecular mechanisms by which PETase activity can emerge from an ancestral cutinase without PETase activity, (ii) these mechanisms are not obviously associated with a restructuring of the enzyme active site, and (iii) PETase activity is readily evolutionarily accessible within the cutinase family, consistent with recent observations of PETases that have emerged from cutinase backgrounds (*11*, *12*).

We next determined the crystal structures of two active ancestors, 35_442 [Protein Data Bank (PDB) 8ETY] and 55_547 (PDB 8ETX), and identified an additional pathway to optimizing PETase activity beyond modification of the active site. Both ancestors crystallized in the C222 space group at resolutions of 1.5 to 1.8 Å (table S1). Structural analysis revealed the formation of domain-swapped dimers (Fig. 2D). The observation of domain swapping, likely induced by the high protein concentration in the crystallization conditions, suggests an intrinsic flexibility of the C terminus, a characteristic commonly observed in proteins previously characterized to form domain-swapped dimers (*46*). Given the flexible nature of the C terminus, it was hypothesized that introducing a disulfide bond to constrain this region into forming an intramolecular β sheet would enhance enzyme stability (Fig. 2E). Furthermore, the chosen site for the disulfide bond coincided with the predicted Ca^{2+}/Mg^{2+} binding site of the ancestors based on homology to extant cutinases, a region previously targeted for disulfide bond engineering in Tf-Cut2 and LCC for improved thermostability and PETase activity (*6*, *47*, *48*). We introduced the disulfide mutation N206C/S253C to ancestor 55_547 and experimental characterization of the resulting mutant demonstrated an approximately twofold improvement in whole-cell activity ($A_{260\ nm} = 1.22 \pm 0.22$) (Fig. 2F) and a 1.5-fold increase in soluble expression levels compared to the 55_547 background determined by Bradford assay (fig. S5). In combination with our comparative analysis of the selected inactive and active ancestors, the successful optimization of PETase activity via the introduction of a disulfide bond distal to the active site illustrates that PETase activity within the cutinase family can emerge and be enhanced through mechanisms that extend beyond the structural reconfiguration of the active site. It is therefore likely that activity optimization is, at least in part, being driven by thermodynamic and kinetic stabilization.

Alternate reconstructions of ancestral cutinases 35_442 and 55_547

We next experimentally characterized all alternate sequences of ancestors 55_547 (40 mutants) and 35_442 (29 mutants) that individually sampled each mutation that had been ambiguously reconstructed. Here, we define ambiguity as any site where at least two amino acids are reconstructed with a posterior probability of ≥0.2. All single mutations were introduced into the sequence of the node that was reconstructed ambiguously and typically represented conservative changes between physiochemically similar amino acids. This provided a high-resolution mutagenic map of the local sequence space around either of the ancestral PETase variants. Ambiguously reconstructed sites were spatially distributed over the protein and not localized to any specific functional area (fig. S7). Having already established that the mean posterior probability of an ancestral sequence is a poor indicator of PETase activity (fig. S4B) and that relatively minor changes in the cutinase sequence can notably alter PETase activity, we hypothesized that the neighborhood of evolutionarily possible (albeit less probable) sequences may contain mutations that benefit PETase activity. This hypothesis was guided by recent ASR studies on the *I. sakaiensis* PETase branch of the cutinase phylogeny where PETase activity appeared to emerge transiently within ancestral lineages (*38*). Experimental characterization revealed that all of the alternate reconstructions of ancestors 55_547 and 35_442 demonstrated PETase activity that was comparable (or reduced) to either respective ancestral background (Fig. 3, A and B).

A subset of randomly sampled combinations of single mutations from the alternate reconstructions were selected to explore potential epistatic interactions and their impact on PETase activity. The seven selected mutations—E14D, E28Q, A36V, S52A (in 55_547), F94L, S196A, and H247S—were recombined in various combinations from 2 to 5 point mutations in the background of 35_442 and 55_547. E28Q, A36V, S52A, S196A, and H247S are surface mutations, distal from the catalytic and putative PET binding site (Fig. 4A). In contrast, F94L is positioned within the PET binding site, as deduced from docking of 2HE-$(MHET)_3$ into both 35_442 and 55_547 (Fig. 4A), as well as structural homology to previously predicted PET binding sites in LCC. As single mutations, E14D, E28Q, A36V, F94L, and H247S were considered neutral based on PETase activity relative to the 35_442 background, while S52A and S196A exhibited decreased activity relative to 55_547. In our experimental characterization, we identified 24 recombined variants with increased PETase activity relative to their ancestral background (Fig. 4B). Specifically, 10 recombined variants in the background of 55_547 showed increased PETase activity, with the most active being E14D/E28Q/A36V ($A_{260\ nm} = 1.20 \pm 0.26$), E28Q/S196A ($A_{260\ nm} = 1.08 \pm 0.05$), and E28Q/S52A/S196A ($A_{260\ nm} = 1.03 \pm 0.09$). Similarly, 14 recombinations in the background of 35_442 demonstrated increased activity, with the most active being E28Q/S196A ($A_{260\ nm} = 1.42 \pm 0.29$), E28Q/F94L ($A_{260\ nm} = 0.99 \pm 0.17$), and E14D/H247S ($A_{260\ nm} = 0.77 \pm 0.08$). We identified 17 recombined variants that exhibited improved PETase activity relative to LCC ICCG, the most active extant variant in the study, with 35_442 E28Q/S196A, displaying ~2.5-fold higher activity.

The mutational analysis of the recombined variants highlighted E28Q as a key mutation, present in 14 of the 17 variants that demonstrated enhanced activity compared to the engineered LCC ICCG variant. This suggests a positive effect of E28Q on PETase activity; however, this was only observed in the presence of other mutations from the recombinations. The context dependence of E28Q is especially important when considering the double mutation E28Q/S196A, which was among the most active in both ancestral backgrounds. When assessed independently in the 55_547 background, the E28Q mutation adversely affected activity ($A_{260\ nm} = 0.00 \pm 0.00$), and the S196A mutation similarly led to a reduction in activity relative to the ancestor ($A_{260\ nm} = 0.16 \pm 0.02$). However, the combination of these mutations resulted in an increase in activity beyond the additive effects of the single mutations ($A_{260\ nm} = 1.08 \pm 0.05$), indicative of positive epistatic interactions.

Ruggedness analysis of PETase fitness landscape

We next analyzed the topology of the PETase fitness landscape over the 196 cutinase sequences characterized in this study. This was done both to build a holistic overview of PETase evolution and function in the cutinase family and to assess the effectiveness of this approach in exploring new-to-nature fitness peaks across functional sequence space. To achieve this, we embedded cutinase sequences as

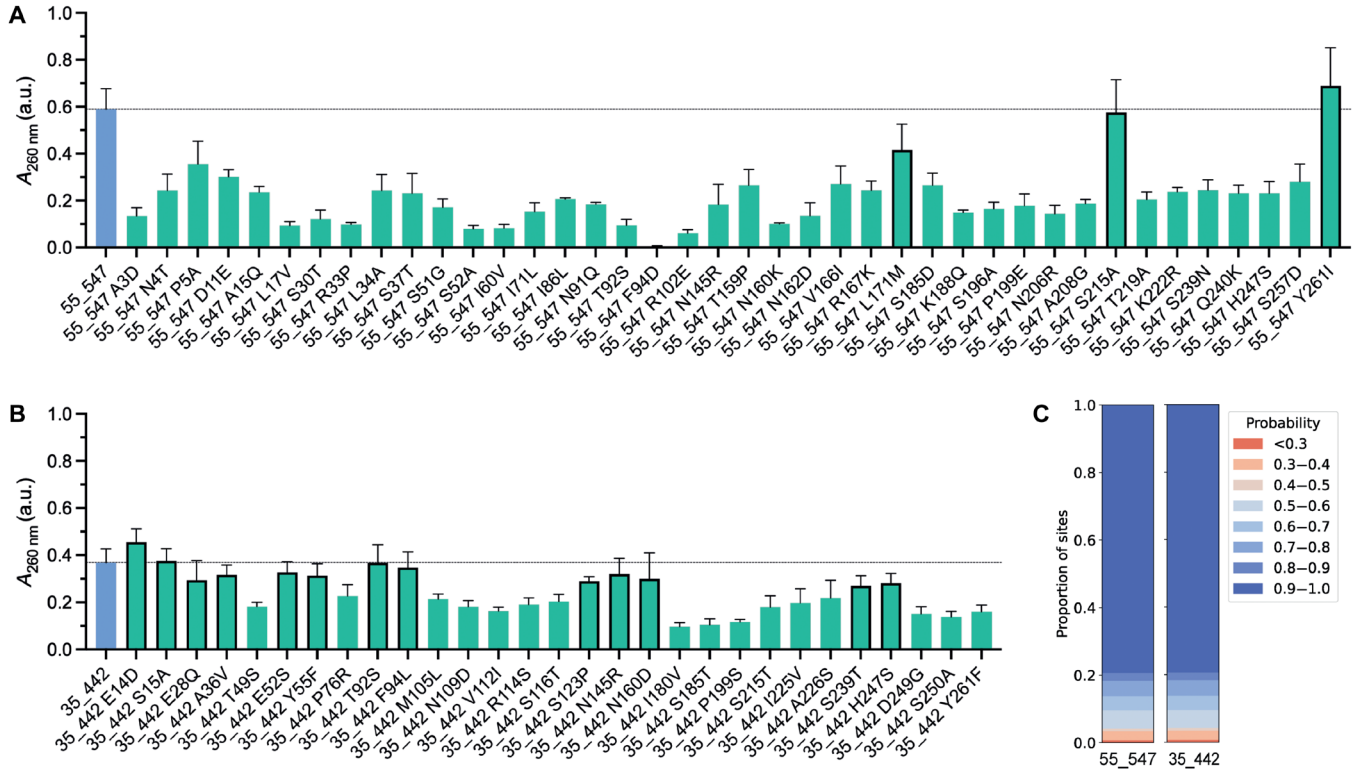

Fig. 3. Alternate reconstructions and random recombinations of ancestral cutinases 35_442 and 55_547. Activity of alternate reconstructions of (**A**) ancestor 55_547 and (**B**) 35_442 against amorphous PET film. The bulk soluble products of PET hydrolysis (TPA, MHET, and BHET) are measured by $A_{260\,nm}$ after 16 hours incubation with the enzyme at 60°C. The horizontal line in each panel represents the activity of the background ancestor for each single mutation. Single mutations considered neutral are highlighted by bold outlines. Data are represented as the means ± SEM ($n = 3$). (**C**) Posterior probability distributions for ancestors 55_547 and 35_442. a.u., arbitrary unit.

nodes in a network graph where edges connect each node to its *k*-nearest neighbors (*k*NNs) (Euclidean) neighbors. The scheme used to embed sequences therefore dictates the topology of the network graph. When this is the OHE, the Euclidean distance between nodes in the network graph is proportional to the number of mutations between the sequences they represent. The OHE graph network therefore captures PETase activity as a function of the mutations between sequences when the signal over the graph is the measured PETase activity (Fig. 5A).

Encoding protein sequences as the hidden states of a PLM can capture comparatively richer features than the OHE (*30*, *31*), albeit at the cost of interpretability over the network graph (*37*). Euclidean distances in the latent space of a representation model may share no interpretable relationship with equivalent Euclidean distances in the OHE space, therefore confounding the interpretation of PETase activity over the network graph to nonlinear distances instead of simple mutational distances.

We embedded all characterized cutinases in an OHE basis and visualized the resulting landscape in two dimensions with t-distributed stochastic neighbor embedding (tSNE) dimensionality reduction. Projection onto the two-dimensional tSNE basis did not meaningfully disrupt local structure in the full OHE basis according to the tSNE trustworthiness metric (0.92; lossless projection corresponds to a score of 1.0) (*49*, *50*), suggesting that clustering in the tSNE basis is not artifactual. In this space, ancestral and extant cutinases form homogenous, random clusters. 35_442 and 55_547 belong to distinct sequence clusters that are sparsely connected (Fig. 5 and fig. S9), reflecting their disparate evolutionary backgrounds. The alternate reconstructions of ancestors 55_547 and 35_442 and their recombinations are grouped as closely connected clusters around their respective ancestral backgrounds. Despite their comparable PETase activity, sequences in both backgrounds are resolved as independent components that share few direct edge connections, suggesting that the PETase fitness landscape is multimodal (i.e., there are multiple solution spaces to competent PETase activity) when considering only naive, residue-wise OHE sequence embeddings.

We then reconstructed the PETase fitness landscape in a learned representation space. To ensure that local features of the cutinase sequence space were captured by a PLM, we used a recently described method of LASE (*37*). In brief, LASE trains a small and family-specific deep learning model on ancestral reconstructed sequence datasets. The features learned by LASE capture the functional properties of proteins, such as catalytic efficiency, in a nonlinear space that is not interpretable with simple mutational distances. Embedding in LASE can therefore reveal topological features of the fitness map that are obfuscated in an OHE embedding.

Sequences projected by tSNE from the LASE embedding space cluster according to their relative fitness and how they were sampled (e.g., extant, ancestral, alternate, or recombined; Fig. 5B). For example, all point mutations from alternate reconstructions and recombinations group together in a single connected component in

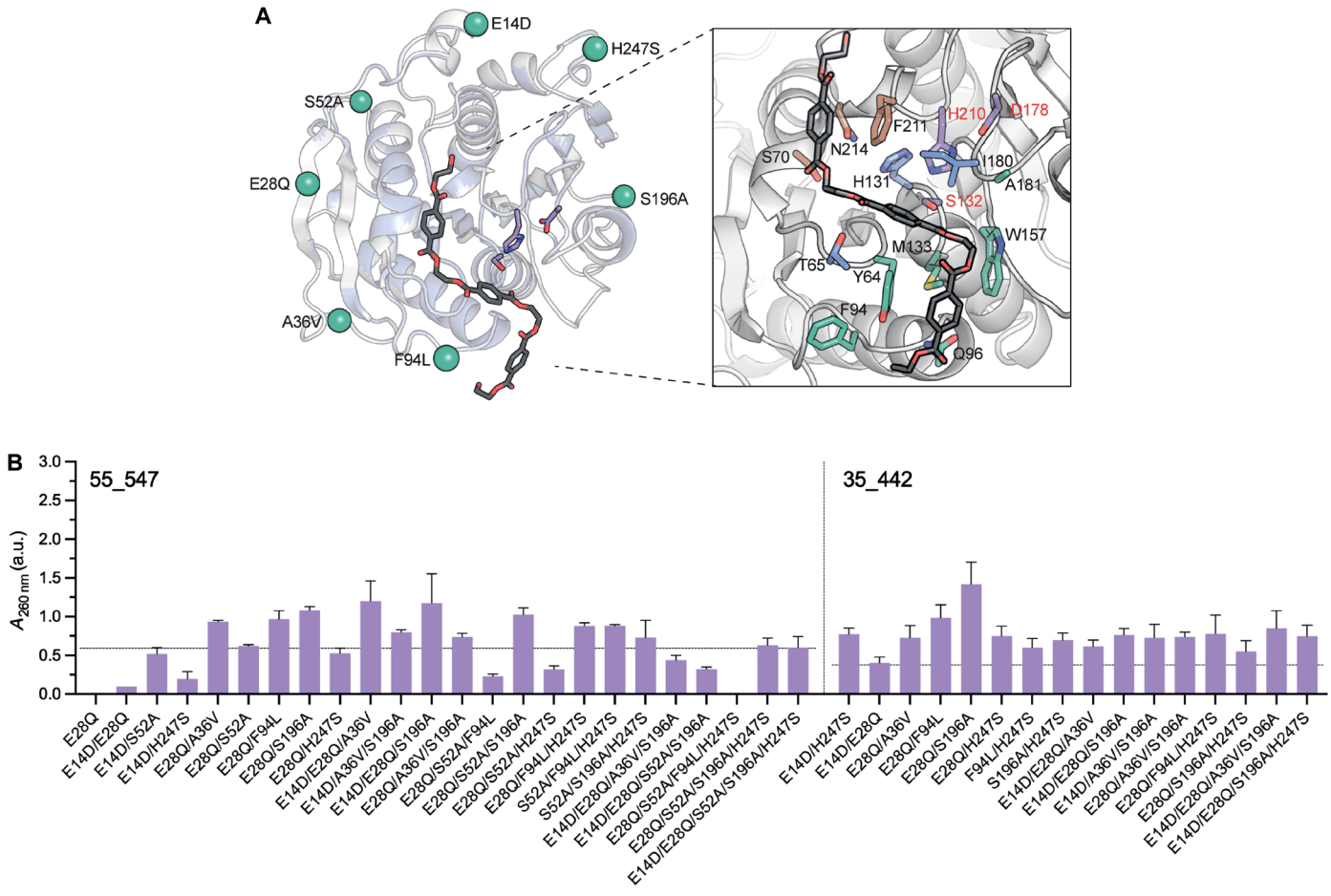

Fig. 4. PETase activities of recombined alternate reconstructions. (**A**) Structure of 35_442 and 55_547 aligned. Positions selected for recombination are highlighted as spheres. The docked pose of 2HE-$(MHET)_3$ is shown in a close-up of the binding and catalytic site, and residues that form the putative PET binding site based on homology to LCC are shown. Residues comprising subsites −2 (green), −1 (blue), and +1 (brown) are highlighted. The catalytic residues Ser^{132}-His^{210}-Asp^{178} are also shown (purple; red labels). (**B**) Activity of recombined mutations from alternate reconstructions of 35_442 and 55_547 against amorphous PET film. The bulk soluble products of PET hydrolysis (TPA, MHET, and BHET) are measured by $A_{260\,nm}$ after 16 hours incubation with the enzyme at 60°C. The horizontal lines represent the activity of the background ancestor for each of the recombined alternate reconstructions. Data are represented as the means ± SEM ($n = 3$) and grouped based on mutations recombined in the background of 55_547 (left) and 35_442 (right).

the LASE space, irrespective of the genetic background (ancestors 55_547 or 35_442) they were introduced into. Moreso, ancestral and extant sequences cluster into disconnected components despite their often high degree of site-wise sequence similarity. The LASE representation space is highly structured relative to the OHE space, where ancestral and extant sequences co-occur across distinct identity groups in the network graph. For example, 35_442 and 55_547 cluster together into the "ancestral cluster" in the LASE embedding space, whereas are resolved as independent components in the OHE domain. As with tSNE projection from the OHE domain, the local structure of the LASE embedding space is preserved nearly perfectly (trustworthiness = 0.98), suggesting that clustering in the tSNE basis is not artifactual. This indicates that our sequence sampling strategy is highly structured and systematic over an evolutionarily informative representation space while appearing somewhat random in the OHE space.

Last, we used a graph signal processing approach to quantify ruggedness in the cutinase PET fitness landscape. We define ruggedness as the nonlinearity between the fitness of a sequence and its neighbors in the network graph. We use the Dirichlet energy (DE) of the graph, which describes the nonlinearity of a signal over a graph, to measure this (*37*, *51*, *52*). To make the Dirichlet energy interpretable as a node-wise local quantity, we calculate it over each subgraph in the network defined by an edge-length of exactly 1, thus reducing its interpretation to the deviation from linearity a node demonstrates relative only to its immediate neighbors; the fitness signal over the network graph changes as a linear function (i.e., is smooth) over nodes of the graph that are characterized by low local DEs. The local Dirichlet energy is therefore a descriptor of how confounded by epistasis a sequence is. This analysis revealed that the fitness landscape is most rugged over the cutinase variants with the greatest PETase activity. The node with the single highest local Dirichlet energy is also the fitness peak (ancestor 55_547_E28Q/S196A). This was true for both OHE and LASE network graphs (Fig. 5C and fig. S9), indicating that the combinatorial mutations E28Q/S196A in the 55_547 background would be unlikely to be introduced through a stepwise mutagenesis approach due to the relatively low activity of each single mutation in isolation. We performed

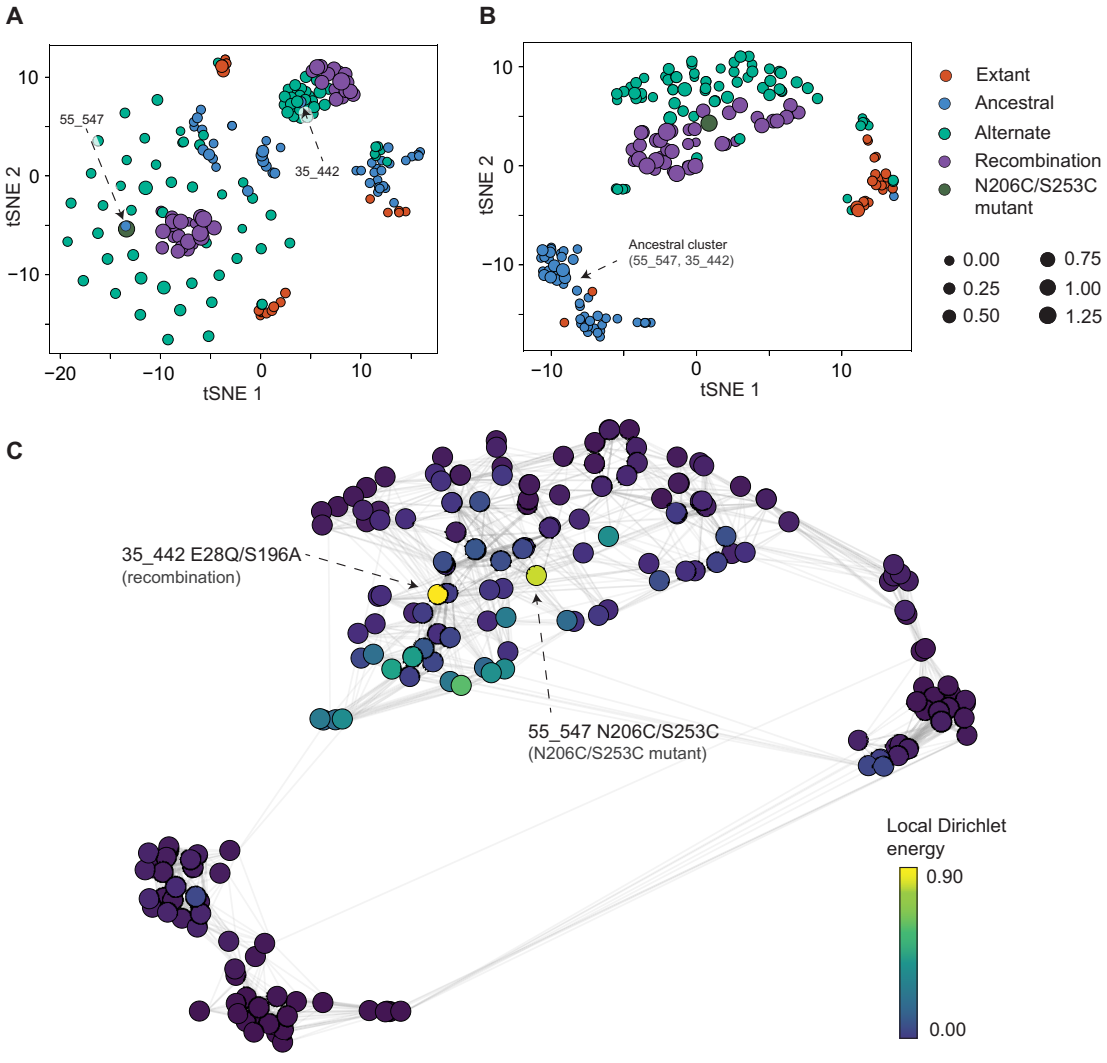

Fig. 5. Analysis and regression on PETase sequence space. PETase sequences were represented in (**A**) OHE and (**B**) LASE forms and projected into a two-dimensional space with tSNE. Color represents design stage, and size represents activity [$A_{260\,nm}$ (a.u.)]. The OHE and LASE sequence data were then used to train regression models. (**C**) Local Dirichlet energy for each variant was determined over subgraphs that include the variant's immediate neighbors as determined by *k*NN. Edges connect variants that were found to be neighbors, and color represents the local Dirichlet energy calculated.

analyses on how sensitive the local energies within the OHE and LASE landscapes were to both the underlying graph connectivity (the k parameter in the k-nearest neighbor (kNN) graph) and the sparsity of the data. We find that node-wise energies are significantly correlated (P value $\ll 0.0001$) for all k and sparsity values tested (figs. 10 to 13), indicating that results on landscape ruggedness are robust to the underlying graph structure. Together, these analyses demonstrate that mASR can effectively guide protein engineering by finding evolutionary features that are not immediately apparent and can help navigate rugged regions of sequence space to find fit enzyme variants that are not rationally obvious.

DISCUSSION

In our investigation into the evolutionary sequence space of bacterial cutinases with PETase activity, we used mASR (*37*) to enhance our exploration. This method allowed us to reconstruct and analyze ancestral cutinases across 20 diverse phylogenetic topologies, moving beyond the constraints of a single-tree perspective. Our experimental characterizations of 20 extant and 48 ancestral cutinases unveiled a wide spectrum of activities against amorphous PET film ($A_{260\,nm} = 0.00$ to 0.59). Of particular interest were two ancestors, 35_442 ($A_{260\,nm} = 0.37 \pm 0.06$) and 55_547 ($A_{260\,nm} = 0.59 \pm 0.09$), each representing the most recent common ancestor of LCC (*13*) and TfCut2 (*5*) on independent trees and exhibiting similar activity to the most active extant cutinases, LCC ($A_{260\,nm} = A_{260\,nm} = 0.34 \pm 0.03$) and BhrPETase ($A_{260\,nm} = 0.49 \pm 0.01$). Notably, the use of mASR was crucial in uncovering the variability of ancestral reconstructed sequences, as it revealed that equivalent nodes from independent phylogenetic topologies exhibited highly varied activity despite all topologies being equally valid representations of the underlying sequence alignment based on the AU test. While some

ancestors were inactive, others showed PETase activity comparable to characterized extant PETases. These observations highlight the utility of the mASR method in improving both the success and robustness of ASR applications in protein engineering. By embracing a wider array of evolutionary scenarios, mASR allowed for a more comprehensive and reliable identification of functional sequences relative to a single-tree approach. mASR may therefore become an important component of ASR- and evolutionary-guided enzyme design strategies in the future.

Our sequence and structural analysis of active and inactive ancestral cutinases provided insights on the evolutionary emergence of PETase activity in the cutinase family. In particular, our findings suggest that PETase activity in the TfCut2 and LCC PETase lineages has emerged through neutral mutations that are distal to the active site, rather than specific changes localized to the active site itself, as observed in the evolution of *I. sakaiensis* PETase (*38*). Our detailed examination of ancestors 35_442 and 55_547, along with their closest inactive counterparts, highlighted unique gain-of-function mutations leading to PETase activity from these distinct phylogenetic backgrounds. These findings imply the existence of multiple, distinct evolutionary pathways for acquiring PETase functionality and indicate that PETase activity is readily accessible within the cutinase family.

Our structural analysis also revealed the formation of domain-swapped dimers between adjacent symmetry mates in the crystal structures of the ancestral variants 35_442 and 55_547. While possibly an artifact of the crystallization conditions, this observation suggests a certain degree of structural flexibility in the C terminus of these ancestors even in solution. Based on this observation, we introduced a disulfide bond in the C terminus of ancestor 55_547 that coincided with the predicted Ca^{2+}/Mg^{2+} binding site, a strategy homologous to similar successful modifications in TfCut2 and LCC that were observed to improve stability and activity (*6*, *47*, *48*). The introduction of the disulfide N206C/S253C resulted in a variant of 55_547 with an approximately twofold improvement in whole-cell activity and 1.5-fold improvement in soluble expression relative to the background ancestor.

We deepened our exploration of the evolutionary sequence space of the cutinase family by addressing ambiguously reconstructed positions in the initial mASR. Ambiguous positions were identified based on a posterior probability threshold of ≥0.2 for the second most probable residue. Introducing these 69 alternate reconstructions as single mutations into 35_442 and 55_547, we observed that most mutations were either neutral or decreased PETase activity relative to the ancestral background. However, when we recombined a random subset of the alternate reconstructions, we observed several recombinations in the background of both 35_442 (E28Q/S196A) and 55_547 (E14D/E28Q/A36V) that displayed increased PETase activity, relative not only to the initial ancestors but also to an engineered variant of LCC with enhanced activity, $LCC^{ICCG\ 6}$. Notably, the most active recombinations exhibited activity improvements that exceeded the additive effects predicted from their individual mutations, highlighting positive epistatic interactions that contribute to their PETase activity.

To complement our experimental investigations, we modeled the sequence-fitness landscape of all 196 cutinase sequences characterized in our study using a graph signal processing approach that elucidates the complex topology of the PETase fitness landscape. This methodology involved embedding cutinase sequences as nodes in a network graph, where each node was connected to its *k*NNs (Euclidean), thereby enabling a direct comparison between two embedding schemes, OHE and LASE. While OHE provides a straightforward mutational distance metric between sequences, it often oversimplifies the nuanced relationship between sequence variation and function. In contrast, LASE, by training a family-specific deep learning model on our dataset of ancestral reconstructed sequences, captures the functional properties of proteins in an abstract manner, revealing nonlinear sequence-activity relationships that are obscured in OHE representations.

Using LASE to map the sequence-fitness landscape, we highlighted iterative improvements in PETase activity throughout the different phases of our sequence exploration of the cutinase family. This embedding not only facilitated a deeper understanding of the functional implications of sequence variation but also allowed us to identify clusters of sequences with similar functional profiles, regardless of their evolutionary background. Moreover, our analysis of the fitness landscape's ruggedness via LASE provided insights into the role of epistasis in PETase activity. The local Dirichlet energy calculations revealed that sequences with the highest PETase activity were associated with the greatest ruggedness, suggesting that the most functionally optimized variants emerge from complex interplays of multiple mutations rather than from linear accumulations of beneficial single mutations. This observation highlights the significance of epistatic interactions in driving the evolutionary innovation of PETase activity, illustrating that successful enzyme variants often lie in regions of the fitness landscape that are not readily accessed through single mutational steps but can be accessed using evolutionarily-guided approaches, such as mASR.

MATERIALS AND METHODS

Ancestral sequence reconstruction

Protein sequences (1000) were collected from the National Center for Biotechnology Information nonredundant (nr) database with pblast using LCC (UniProt: G9BY57) as seed and an E-value cutoff of 1×10^{-5}. Sequence redundancy was removed to 90% sequence identity (ID) in CD-HIT (*53*). Signal peptides were deleted using SignalP4.0 (*54*). Alignment was performed using the GINSI protocol of MAFFT (*55*). Replicates (100) of independent model parameterization and tree search (default parameters) were performed using IQTREE2 (*56*) on the National Computational Infrastructure (NCI) Gadi supercomputer. The sequence evolution model was parameterized using ModelFinder (*57*), as implemented in IQTREE2. Branch supports were determined as the ultrafast bootstrap approximation (*39*) calculated to 1000 replicates, as implemented in IQTREE2. The AU test (*42*) was conducted to 10,000 replicates for all ML topologies. Empirical Bayesian ASR was performed on 20 of the 100 trees that failed rejection by the AU test in CodeML (*40*) using the ML replacement matrix (LG) (*58*) with rates modeled as a discrete gamma distribution parameterized with four rate categories.

Small-scale protein expression and purification

Plasmids were transformed by heat shock into chemically competent BL21(DE3) *Escherichia coli* cells and plated onto lysogeny broth (LB) agar supplemented with kanamycin (100 μg/ml) and incubated at 37°C overnight. A single colony was used to inoculate 1.5-ml autoinduction media supplemented with kanamycin (100 μg/ml) in a 2.2 ml of 96-well deep well block and grown at 1050 rpm at 37°C for 5 hours, followed by room temperature (RT; 25°C) for 16 hours.

Cells were harvested by centrifugation at 2000*g* for 15 min at RT and resuspended in lysis buffer [1× BugBuster Protein Extraction Reagent (Merck-Millipore), 20 mM tris, 300 mM NaCl, turbonuclease (1 U/ml; pH 8; Sigma-Adrich)]. The cell suspension was left to incubate at RT for 20 min with gentle shaking. The lysate was separated from the insoluble cell debris by centrifugation at 2250*g* for 1 hour at RT.

The clarified lysate was then diluted with 100 μl of equilibration buffer [20 mM tris and 300 mM NaCl (pH 8)] and purified by nickel-charged immobilized metal affinity chromatography (IMAC) using a 96-well HisPurTM Ni-NTA Spin Plate (Thermo Fisher Scientific) equilibrated in equilibration buffer, washing the sample three times with 250 μl of wash buffer [20 mM tris, 300 mM NaCl, and 10 mM imidazole (pH 8)], and eluting with 250 μl of elution buffer [20 mM tris, 300 mM NaCl, and 150 mM imidazole (pH 8)]. All centrifugation steps following addition of wash or elution buffer were at 1000*g* for 1 min at RT. The eluate was stored at 4°C. Bradford assay was used to quantify the soluble expression levels post-purification (fig. S1).

UV absorbance assay for PET-degrading activity

For purified protein from 96-well expression and purification, 15 μl of the eluate from the 96-well Ni-NTA purification and 285 μl of reaction buffer [50 mM bicine (pH 9)] was added to a clear 96-well plate. For purified protein from large-scale expression and purification, 300 μl of 100 nM enzyme in reaction buffer was added to a clear 96-well plate. A single disk of amorphous PET (Goodfellow, ES301445) with 4 mm in diameter and 0.25 mm in thickness was added to each well. The plate was incubated at 60°C for 16 hours. Following incubation, 100 μl of the reaction solution was transferred to a clear UV-transparent 96-well plate, and the absorbance was measured between 240 and 300 nm in 10-nm steps using the Epoch Microplate Spectrophotometer (BioTek) (fig. S2). For comparison of activity of all variants, the $A_{260\ nm}$ was used. Assays were repeated in technical triplicate for each variant.

Assay data processing

To correct for possible systematic error between replicate data points, the mean $A_{260\ nm}$ for each replicate was determined. Using the mean absorbance, a scaling coefficient was assigned to the replicates with the lowest and highest mean values such that the mean of the scaled absorbance values equaled the mean of the replicate with the mid-range mean value. To ensure scaling improved consistency between replicates, correlograms of the data before (fig. S3) and after (fig. S4) scaling were produced to confirm that the monotonic (rank) correlation between replicates was preserved.

Cloning of TEV-PETase variants for crystallography studies

Primer pairs containing the DNA sequence for the tobacco etch virus (TEV) cleavage site (5'-GAAAACCTGTATTTTCAAAGC-3′) were constructed, specific to each PETase variant. Polymerase chain reaction was performed using these primers and PETase variant genes to create mutant fragments. These fragments were reassembled using Gibson Assembly (*59*) and checked through Sanger sequencing to ensure that the TEV cleavage site was correctly introduced.

Large scale protein expression and purification

The TEV-PETase ancestral variants plasmids were transformed using electroporation into electrocompetent E. cloni cells (Lucigen) and plated onto LB agar supplemented with kanamycin (100 μg/ml). The plates were incubated overnight at 37°C. A single colony was inoculated into a 10-ml solution of LB media with kanamycin (100 μg/ml) and incubated overnight at 37°C and 180 rotations per minute (rpm). This liquid starter culture was then added to 1 liter of autoinduction media (*60*) [6 g of Na_2HPO_4, 3 g of KH_2PO_4, 20 g of tryptone, 5 g of yeast extract, 5 g of NaCl, 10 ml of 60% (v/v) glycerol, 5 ml of 10% (w/v) glucose, and 25 ml of 8% (w/v) lactose] with kanamycin (100 μg/ml) and incubated for 24 hours at room temperature and 180 rpm. The cells were separated from the media by centrifugation at 5000*g* for 15 min at 4°C and resuspended in lysis buffer (400 mM NaCl, 25 mM imidazole, turbonuclease (1 U/mL; Sigma-Aldrich), and 50 mM tris-HCl (pH 8.0)]. The resuspended cell solution was lysed using two rounds of sonication at 50% power and pulse time for 5 min, with 5 min on ice between sonication steps. Next, the sample was centrifuged at 32,000*g* for 60 min at 4°C, and the soluble cell solution was separated from the insoluble cell material and filtered through a 0.45-μm pore size filter. The filtered soluble cell solution was passed through an equilibrated nickel-charged IMAC using a 5-ml HisTrap HP (GE Healthcare Life Sciences) in lysis buffer. The protein bound to the column was eluted using elution buffer [400 mM NaCl, 500 mM imidazole, and 50 mM tris-HCl (pH 8.0)]. The protein sample was buffer exchanged to TEV reaction buffer [100 mM NaCl, 0.5 mM EDTA, 1 mM dithiothreitol, 1% (v/v) glycerol, and 50 mM tris-HCl (pH 8.0)] using a PD-10 desalting column and diluted to 50 ml in this buffer. A 1 ml of solution containing purified TEV protease (1 mg/ml) was added and incubated at room temperature overnight. The cleaved sample was passed through an equilibrated Nickel-charged IMAC using a 5-ml HisTrap HP (GE Healthcare Life Sciences), and the flowthrough was collected. This flowthrough was concentrated using the 3-kDa Amicon ultra 15-ml centrifugal filters and filtered through a 0.22-μm filter. Last, the cleaved protein was purified to homogeneity using size exclusion chromatography, and the HiLoad 26/600 Superdex 200 column (GE Healthcare Life Sciences) was equilibrated in size exclusion buffer [150 mM NaCl and 25 mM Hepes (pH 7.5)].

Protein crystallization and structure determination

Proteins were concentrated using the 3-kDa Amicon ultra 15-ml centrifugal filters to 15 to 36 mg/ml and crystallized in 20% (w/v) polyethylene glycol (PEG) 3350 alongside 0.2 M salt and bis-tris buffer solution, specifically ancestor 55_547 in 0.2 M sodium/potassium tartrate, 0.1 M bis-tris propane (pH 7.5), and 20% (w/v) PEG 3350 and ancestor 35_442 in 0.2 M sodium malonate, 0.1 M bis-tris propane (pH 6.5), and 20% (w/v) PEG 3350. The x-ray diffraction data were collected on the MX2 beamline at the Australian Synchrotron (*61*). The data were processed using XDS (*62*), and the phase problem was resolved with molecular replacement using the PETase wild-type structure (PDB: 6EQE) as the search model. The ligands and solvent molecules were removed and then used as the search model part of Phaser (CCP4) (*63*). The structure was refined using phenix.refine (*64*) through multiple iterative steps and rebuilt each time with Coot (*65*). The structures of ancestors 55_547 and 35_442 were deposited in the PDB under the PDB IDs 8ETX and 8ETY, respectively.

Protein sequence representations

To analyze PETase sequence space, OHE and LASE were made. To produce OHE, aligned PETase sequences were converted into a (20by267) matrix, where gaps were represented as a zero vector of

length 20. The LASE embedding model was implemented as a Transformer in PyTorch 2.0.1 as previously described (*37*), with three encoder blocks with two-headed multihead attention (64 dimensions) and a feed-forward fully connected layer (128 dimensions). The LASE embedding model was trained with a masking percent of 15%, more than 100 epochs with a batch size of 32 using the Adam optimizer. Loss was determined as the categorical cross entropy loss. Performance over training was assessed with perplexity and categorical accuracy (fig. S8).

Local Dirichlet energy calculations

To estimate the local ruggedness of each PETase variant, the *k*NNs of each node were used to produce a *k*NN subgraph for each variant with scikit-learn 1.2.1. The kNN subgraphs were made symmetric by considering (bi)directional edges as undirectional. The dirichlet energy was calculated as previously described (*52*, *66*)

$$\lambda_m = 1/N\, \boldsymbol{y}^T \boldsymbol{L}\boldsymbol{y}$$

where, λ_m is the normalized dirichlet energy, N is the number of variants in the subgraph, $\boldsymbol{y}$ is the activity (absorbance) of each variant, $\boldsymbol{L}$ is the graph Laplacian operator of the adjacency matrix of each *k*NN subgraph, and ^{T} denotes a transpose operation.

Supplementary Materials

This PDF file includes:

Figs. S1 to S13

Table S1

REFERENCES AND NOTES

1. OECD, *Plastic pollution is growing relentlessly as waste management and recycling fall short, says OECD* (accessed 16 December 2023); www.oecd.org/environment/plastic-pollution-is-growing-relentlessly-as-waste-management-and-recycling-fall-short.htm.
2. OECD, *Global Plastics Outlook* (2022); https://www.oecd.org/en/publications/2022/06/global-plastics-outlook_f065ef59.html.
3. B. Zhu, D. Wang, N. Wei, Enzyme discovery and engineering for sustainable plastic recycling. *Trends Biotechnol.* **40**, 22–37 (2022).
4. M. R. Egmond, J. de Vlieg, Fusarium solani pisi cutinase. *Biochimie* **82**, 1015–1021 (2000).
5. E. Herrero Acero, D. Ribitsch, G. Steinkellner, K. Gruber, K. Greimel, I. Eiteljoerg, E. Trotscha, R. Wei, W. Zimmermann, M. Zinn, A. Cavaco-Paulo, G. Freddi, H. Schwab, G. Guebitz, Enzymatic surface hydrolysis of PET: Effect of structural diversity on kinetic properties of cutinases from thermobifida. *Macromolecules* **44**, 4632–4640 (2011).
6. V. Tournier, C. M. Topham, A. Gilles, B. David, C. Folgoas, E. Moya-Leclair, E. Kamionka, M. L. Desrousseaux, H. Texier, S. Gavalda, M. Cot, E. Guémard, M. Dalibey, J. Nomme, G. Cioci, S. Barbe, M. Chateau, I. André, S. Duquesne, A. Marty, An engineered PET depolymerase to break down and recycle plastic bottles. *Nature* **580**, 216–219 (2020).
7. A. Carniel, É. Valoni, J. Nicomedes Junior, A. C. Gomes, A. M. Castro, Lipase from *Candida Antarctica* (CALB) and cutinase from *Humicola insolens* act synergistically for PET hydrolysis to terephthalic acid. *Process Biochem.* **59**, 84–90 (2017).
8. G. von Haugwitz, X. Han, L. Pfaff, Q. Li, H. Wei, J. Gao, K. Methling, Y. Ao, Y. Brack, J. Mican, C. G. Feiler, M. S. Weiss, D. Bednar, G. J. Palm, M. Lalk, M. Lammers, J. Damborsky, G. Weber, W. Liu, U. T. Bornscheuer, R. Wei, Structural insights into (Tere)phthalate-Ester hydrolysis by a carboxylesterase and its role in promoting PET depolymerization. *ACS Catal.* **12**, 15259–15270 (2022).
9. C. Gamerith, M. Vastano, S. M. Ghorbanpour, S. Zitzenbacher, D. Ribitsch, M. T. Zumstein, M. Sander, E. Herrero Acero, A. Pellis, G. M. Guebitz, Enzymatic degradation of aromatic and aliphatic polyesters by *P. pastoris* expressed cutinase 1 from *Thermobifida cellulosilytica*. *Front. Microbiol.* **8**, 938 (2017).
10. C. Silva, S. Da, N. Silva, T. Matamá, R. Araújo, M. Martins, S. Chen, J. Chen, J. Wu, M. Casal, A. Cavaco-Paulo, Engineered *Thermobifida fusca* cutinase with increased activity on polyester substrates. *Biotechnol. J.* **6**, 1230–1239 (2011).
11. L. A. Amaral-Zettler, E. R. Zettler, T. J. Mincer, Ecology of the plastisphere. *Nat. Rev. Microbiol.* **18**, 139–151 (2020).
12. X. Qi, M. Ji, C. F. Yin, N. Y. Zhou, Y. Liu, Glacier as a source of novel polyethylene terephthalate hydrolases. *Environ. Microbiol.* **25**, 2822–2833 (2023).
13. S. Sulaiman, S. Yamato, E. Kanaya, J. J. Kim, Y. Koga, K. Takano, S. Kanaya, Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Appl. Environ. Microbiol.* **78**, 1556–1562 (2012).
14. A. M. Ronkvist, W. Xie, W. Lu, R. A. Gross, Cutinase-catalyzed hydrolysis of poly(ethylene terephthalate). *Macromolecules* **42**, 5128–5138 (2009).
15. M. A. M. E. Vertommen, V. A. Nierstrasz, M. Veer, M. M. C. G. Warmoeskerken, Enzymatic surface modification of poly(ethylene terephthalate). *J. Biotechnol.* **120**, 376–386 (2005).
16. Y. Akutsu-Shigeno, Y. Adachi, C. Yamada, K. Toyoshima, N. Nomura, H. Uchiyama, T. Nakajima-Kambe, Isolation of a bacterium that degrades urethane compounds and characterization of its urethane hydrolase. *Appl. Microbiol. Biotechnol.* **70**, 422–429 (2006).
17. Y. Ma, M. Yao, B. Li, M. Ding, B. He, S. Chen, X. Zhou, Y. Yuan, Enhanced poly(ethylene terephthalate) hydrolase activity by protein engineering. *Engineering* **4**, 888–893 (2018).
18. Y. Cui, Y. Chen, X. Liu, S. Dong, Y. Tian, Y. Qiao, R. Mitra, J. Han, C. Li, X. Han, W. Liu, Q. Chen, W. Wei, X. Wang, W. du, S. Tang, H. Xiang, H. Liu, Y. Liang, K. N. Houk, B. Wu, Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy. *ACS Catal.* **11**, 1340–1350 (2021).
19. H. F. Son, I. J. Cho, S. Joo, H. Seo, H. Y. Sagong, S. Y. Choi, S. Y. Lee, K. J. Kim, Rational protein engineering of thermo-stable PETase from ideonella sakaiensis for highly efficient PET degradation. *ACS Catal.* **9**, 3519–3526 (2019).
20. R. Wei, T. Oeser, J. Schmidt, R. Meier, M. Barth, J. Then, W. Zimmermann, Engineered bacterial polyester hydrolases efficiently degrade polyethylene terephthalate due to relieved product inhibition. *Biotechnol. Bioeng.* **113**, 1658–1665 (2016).
21. K. N. Hellesnes, S. Vijayaraj, P. Fojan, E. Petersen, G. Courtade, Biochemical characterization and NMR study of a PET-hydrolyzing cutinase from *Fusarium solani pisi*. *Biochemistry* **62**, 1369–1375 (2023).
22. M. A. Spence, J. A. Kaczmarski, J. W. Saunders, C. J. Jackson, Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).
23. G. K. A. Hochberg, J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).
24. T. N. Starr, S. K. Zepeda, A. C. Walls, A. J. Greaney, S. Alkhovsky, D. Veesler, J. D. Bloom, ACE2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature* **603**, 913–918 (2022).
25. A. T. Meger, M. A. Spence, M. Sandhu, D. Matthews, J. Chen, C. J. Jackson, S. Raman, Rugged fitness landscapes minimize promiscuity in the evolution of transcriptional repressors. *Cell Syst.* **15**, 374–387.e6 (2024).
26. E. C. Hartman, D. Tullman-Ercek, Learning from protein fitness landscapes: A review of mutability, epistasis, and evolution. *Curr. Opin. Syst. Biol.* **14**, 25–31 (2019).
27. J. Livada, A. M. Vargas, C. A. Martinez, R. D. Lewis, Ancestral sequence reconstruction enhances gene mining efforts for industrial Ene reductases by expanding enzyme panels with thermostable catalysts. *ACS Catal.* **13**, 2576–2585 (2023).
28. R. E. S. Thomson, S. E. Carrera-Pacheco, E. M. J. Gillam, Engineering functional thermostable proteins using ancestral sequence reconstruction. *J. Biol. Chem.* **298**, 102435 (2022).
29. Y. Gumulya, J. M. Baek, S. J. Wun, R. E. S. Thomson, K. L. Harris, D. J. B. Hunter, J. B. Y. H. Behrendorff, J. Kulig, S. Zheng, X. Wu, B. Wu, J. E. Stok, J. J. de Voss, G. Schenk, U. Jurva, S. Andersson, E. M. Isin, M. Bodén, L. Guddat, E. M. J. Gillam, Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat. Catal.* **1**, 878–888 (2018).
30. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
31. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
32. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
33. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
34. L. Chen, Z. Zhang, Z. Li, R. Li, R. Huo, L. Chen, D. Wang, X. Luo, K. Chen, C. Liao, M. Zheng, Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Syst.* **14**, 706–721.e5 (2023).
35. S. D'Costa, E. C. Hinds, C. R. Freschlin, H. Song, P. A. Romero, Inferring protein fitness landscapes from laboratory evolution experiments. *PLOS Comput. Biol.* **19**, e1010956 (2023).
36. B. L. Hie, K. K. Yang, P. S. Kim, Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**, 274–285.e6 (2022).

37. D. S. Matthews, M. A. Spence, A. C. Mater, J. Nichols, S. B. Pulsford, M. Sandhu, J. A. Kaczmarski, C. M. Miton, N. Tokuriki, C. J. Jackson, Leveraging ancestral sequence reconstruction for protein representation learning. *Nat. Mach. Intell.* **6**, 1542–1555 (2024).
38. Y. Joho, V. Vongsouthi, M. A. Spence, J. Ton, C. Gomez, L. L. Tan, J. A. Kaczmarski, A. T. Caputo, S. Royan, C. J. Jackson, A. Ardevol, Ancestral sequence reconstruction identifies structural changes underlying the evolution of ideonella sakaiensis PETase and variants with improved stability and activity. *Biochemistry* **62**, 437–450 (2023).
39. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
40. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
41. G. N. Eick, J. T. Bridgham, D. P. Anderson, M. J. Harms, J. W. Thornton, Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Mol. Biol. Evol.* **34**, 247–261 (2017).
42. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
43. X. Han, W. Liu, J. W. Huang, J. Ma, Y. Zheng, T. P. Ko, L. Xu, Y. S. Cheng, C. C. Chen, R. T. Guo, Structural insight into catalytic mechanism of PET hydrolase. *Nat. Commun.* **8**, 2106 (2017).
44. B. Guo, S. R. Vanga, X. Lopez-Lorenzo, P. Saenz-Mendez, S. R. Ericsson, Y. Fang, X. Ye, K. Schriever, E. Bäckström, A. Biundo, R. A. Zubarev, I. Furó, M. Hakkarainen, P. O. Syrén, Conformational selection in biocatalytic plastic degradation by PETase. *ACS Catal.* **12**, 3397–3409 (2022).
45. A. Crnjar, A. Griñen, S. C. L. Kamerlin, C. A. Ramírez-Sarmiento, Conformational selection of a tryptophan side chain drives the generalized increase in activity of PET hydrolases through a Ser/Ile double mutation. *ACS Org. Inorg. Au* **3**, 109–119 (2023).
46. Y. Liu, D. Eisenberg, 3D domain swapping: As domains continue to swap. *Protein Sci. Publ. Protein Soc.* **11**, 1285–1299 (2002).
47. J. Then, R. Wei, T. Oeser, M. Barth, M. R. Belisário-Ferrari, J. Schmidt, W. Zimmermann, Ca^{2+} and Mg^{2+} binding site engineering increases the degradation of polyethylene terephthalate films by polyester hydrolases from *Thermobifida Fusca*. *Biotechnol. J.* **10**, 592–598 (2015).
48. J. Then, R. Wei, T. Oeser, A. Gerdts, J. Schmidt, M. Barth, W. Zimmermann, A disulfide bridge in the calcium binding site of a polyester hydrolase increases its thermal stability and activity against polyethylene terephthalate. *FEBS Open Bio* **6**, 425–432 (2016).
49. L. van der Maaten, Learning a parametric embedding by preserving local structure, in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (PMLR, 2009), pp. 384–391.
50. J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: An experimental study, in *Artificial Neural Networks — ICANN 2001*, G. Dorffner, H. Bischof, K. Hornik, Eds. (Springer, 2001), pp. 485–491.
51. M. Daković, L. Stanković, E. Sejdić, Local smoothness of graph signals. *Math. Probl. Eng.* **2019**, e3208569 (2019).
52. E. Castro, A. Godavarthi, J. Rubinfien, K. Givechian, D. Bhaskar, S. Krishnaswamy, Transformer-based protein generation with regularized latent space optimization. *Nat. Mach. Intell.* **4**, 840–851 (2022).
53. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
54. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
55. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
56. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
57. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
58. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
59. D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison III, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
60. F. W. Studier, Stable expression clones and auto-induction for protein production in *E. Coli*. *Methods Mol. Biol. Clifton NJ* **1091**, 17–32 (2014).
61. D. Aragão, J. Aishima, H. Cherukuvada, R. Clarken, M. Clift, N. P. Cowieson, D. J. Ericsson, C. L. Gee, S. Macedo, N. Mudie, S. Panjikar, J. R. Price, A. Riboldi-Tunnicliffe, R. Rostan, R. Williamson, T. T. Caradoc-Davies, MX2: A high-flux undulator microfocus beamline serving both the chemical and macromolecular crystallography communities at the australian synchrotron. *J. Synchrotron Radiat.* **25**, 885–891 (2018).
62. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
63. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
64. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
65. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
66. E. Castro, A. Benz, A. Tong, G. Wolf, S. Krishnaswamy, Uncovering the folding landscape of RNA secondary structure with deep graph embeddings. arXiv:06885 [cs.LG] (2022).
67. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
68. 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).

Acknowledgments

Funding: We acknowledge the ARC Centre of Excellence for Innovations in Peptide and Protein Science (CE200100012) (to C.J.J.), the ARC Centre of Excellence in Synthetic Biology (CE200100029) (to C.J.J.), and a Cooperative Research Centre Project Grant from the Australian Government (to C.J.J.). This research was undertaken in part using the MX2 beamline at the Australian Synchrotron, part of ANSTO, and made use of the Australian Cancer Research Foundation (ACRF) detector (to C.J.J.). This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government (to C.J.J.). **Author contributions:** V.V.: Investigation, Data curation, writing—original draft, writing—review & editing, methodology, data curation, supervision, formal analysis, visualization, conceptualization, project administration. R.G.: Investigation, validation. D.S.M.: Methodology, formal analysis, validation, investigation, writing—original draft, visualization, conceptualization, data curation, and software. J.S.: Investigation, methodology, data curation, validation, formal analysis, and software. B.M.L.: Investigation, writing—review and editing. J.T.: Investigation. A.M.D.: Conceptualization, supervision, writing—review and editing, and validation. R.L.F.: Investigation, writing—original draft, writing—review and editing, validation, supervision, formal analysis, and visualization. M.A.S.: Investigation, conceptualization, writing—original draft, writing—review and editing, visualization, conceptualization, methodology, data curation, formal analysis, and supervision. C.J.J.: Conceptualization, writing—original draft, writing—review and editing, supervision, funding acquisition, investigation, methodology, resources, validation, formal analysis, project administration, and visualization. **Competing interests:** The authors declare the following competing interest(s): C.J.J., V.V., M.A.S., R.G., D.S.M., J.S., A.M.D., and J.T. hold equity in the plastic recycling company, Samsara Eco. C.J.J., M.A.S., and V.V. are inventors of a patent (AU2022309300A1) assigned to Samsara Eco claiming the described enzymes. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Code for model training and analysis, trained LASE weights, phylogenetic trees, and sequence files are available on Zenodo (10.5281/zenodo.14708816) and GitHub (https://github.com/RSCJacksonLab/cutinase_lase. Deposited crystal structures are available on the PDB under accession IDs 8ETY and 8ETX.

Submitted 1 September 2024
Accepted 9 April 2025
Published 14 May 2025
10.1126/sciadv.ads8318