Toward *De Novo* Protein Design from Natural Language

Fengyuan Dai[1†], Shiyang You[2†], Yudian Zhu[1], Yuan Gao[4], Lihao Fu[4], Xibin Zhou[1], Jin Su[1], Chentong Wang[5], Yuliang Fan[1], Xiaoxiao Ma[2], Xianjun Deng[2], Letong Yu[2], Hui Qian[1], Yan He[5], Yitao Ke[5], Chenchen Han[1], Xing Chang[5], Liangzhen Zheng[6], Sheng Wang[6], Yajie Wang[1], Anping Zeng[1], Shunzhi Wang[3], Tong Si[4*], Jianming Liu[1*], Hongyuan Lu[2*], Fajie Yuan[1*]

[1]School of Engineering, Westlake University, Hangzhou, 310024, China.
[2]Bioscience and Biomedical Engineering Thrust, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511400, China.
[3]Department of Biochemistry, University of Washington Seattle, WA 98195, USA.
[4]State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences Shenzhen, 518055, China.
[5]School of Life Sciences, Westlake University, Hangzhou, 310024, China.
[6]Shanghai Zelixir Biotech Co., Ltd., Shanghai, 201203, China.

*Corresponding author(s). E-mail(s): tong.si@siat.ac.cn; liujianming@westlake.edu.cn; hongylu@hkust-gz.edu.cn; yuanfajie@westlake.edu.cn;
†These authors contributed equally to this work.

Abstract

Programming biological function—designing bespoke proteins that execute specific tasks on demand—is a foundational goal of molecular engineering. Yet, current protein design paradigms remain fundamentally limited, typically requiring either an existing protein to evolve from, or deep, family-specific expertise to guide the design process. Here we introduce Pinal, a generative model that overcomes this barrier by translating functional descriptions in natural language directly into diverse and active proteins. This capability is built upon a 16-billion-parameter foundation model trained on an unprecedented synthetic corpus of 1.7 billion protein-text pairs, enabling it to ground functional language in the biophysical principles of protein structure. To provide definitive experimental validation, we tasked Pinal with designing four proteins from distinct functional classes: a fluorescent protein, a polyethylene terephthalate hydrolase, an alcohol dehydrogenase, and a metabolic H-protein. Remarkably, all four designs were functionally active and the two Pinal-designed enzymes achieved catalytic turnover for their respective reactions. Notably, the Pinal-designed H-protein even surpassed its natural counterpart, exhibiting 1.7-fold higher performance. Our results establish that natural language can serve as a programmable instruction set for biology, democratizing protein design and shifting the paradigm from the incremental modification of existing molecules to the direct creation of function from a conceptual description.

The ambition to program biological function—creating bespoke proteins to unlock solutions in medicine, sustainability, and industry—is rapidly moving from the realm of science fiction to a tangible frontier of molecular engineering. For decades, this quest has been dominated by two complementary philosophies: mimicking natural selection through directed evolution[1–7], or applying first principles of physics and chemistry to design proteins *de novo* using computational tools like Rosetta[8–15]. The recent advent of artificial intelligence (AI), particularly deep generative models[16], promised a powerful third paradigm, learning the intricate grammar of protein sequence[17–21], structure[22–31], and function[32–34] from vast biological datasets to redefine the boundaries of what is possible.

Yet, despite this remarkable progress, a fundamental bottleneck persists. The dialogue between the human scientist and the protein design algorithm remains tethered to a low-level language of structural constraints[35–37], mutational lists[38,39], and family-specific parameters[40,41]. This paradigm places the burden of translation on the user, demanding they possess deep, specialized expertise to convert a high-level conceptual goal—such as a plastic-degrading enzyme or a fluorescent imaging reporter—into a precise, machine-readable blueprint. This expertise barrier fundamentally limits the scope of creative exploration and restricts the power of protein engineering to a small cadre of specialists, leaving a vast universe of potential protein solutions unexplored.

We address this limitation by teaching machines to comprehend the natural language of scientific intent. Here, we introduce Pinal (http://www.denovo-pinal.com/, Fig. 1), a generative foundation model that redefines the interface for protein design. The model's central innovation is a principled architecture that decomposes the intractable language-to-sequence problem by using protein structure as a semantic and geometric bridge (t $\rightarrow$ c $\rightarrow$ s). This decomposition first translates natural language (t) into a structural blueprint (c), which in turn guides the generation of the final amino acid sequence (s). This framework is brought to life by the scale of a 16-billion-parameter model trained on an unprecedented synthetic corpus of 1.7 billion protein-text pairs, allowing it to master the complex grammar linking human language to the protein design landscape. We observe strong scaling trends, where both data and model size are critical drivers of performance. Our results show that Pinal faithfully translates complex language prompts into viable protein candidates, often discovering creative solutions with significant sequence novelty and diverse structures that expand beyond the catalog of experimentally verified proteins.

To demonstrate Pinal's practical capabilities, we guided it with natural language to generate four proteins from distinct functional families: a green fluorescent protein (GFP), a polyethylene terephthalate hydrolase (PETase), an alcohol dehydrogenase (ADH), and a metabolic H-protein. Experimental validation confirmed activity across all four designs. Notably, this included the first enzymes—the Pinal-designed PETase and ADH—generated with catalytic function through language-based instructions alone. In addition, the Pinal-designed H-protein exhibited a 1.7-fold improvement in performance over its natural counterpart, demonstrating that language-guided design can access solutions beyond those found in nature. Collectively, these results establish that human language can serve as a source code for biology, shifting the paradigm from modifying what exists to the direct authorship of what is imagined.

1 The Pinal network

Pinal's architecture is founded on a central hypothesis: that the exponential vastness of amino acid sequence space makes direct mapping from natural language intractable, but this problem becomes both tractable and controllable when decomposed through a structural intermediate, since protein structures are far more conserved than their corresponding sequences[42]. Our preliminary experiments confirmed this hypothesis. To validate it, we compared two models with identical encoder-decoder architectures trained on the same protein-text dataset. The direct language-to-sequence model failed to achieve meaningful accuracy on the validation set. In stark contrast, the language-to-structure model, which generates Foldseek 3Di structure token sequences[43], achieved substantially superior performance and robust convergence (Fig. 1a,b). Grounded in this principle, Pinal is realized as a generative framework that models the joint probability of a sequence (s) and its structure (c), conditioned on the text (t). This is formalized through a simple Bayesian graphical model which decomposes the probability as: $p(s, c|t) = p(c|t)p(s|c, t)$. This probabilistic factorization naturally defines Pinal's two-stage process (Fig. 1c, Extended Data Fig. 1): first, generating a structural scaffold (c) from the text prior $p(c|t)$, and second, designing a sequence (s) conditioned on both the scaffold and the text, according to $p(s|c, t)$.

The first stage is handled by T2struct, a text-to-structure module that translates a natural language prompt (t) into a structural blueprint (c). Critically, T2struct represents protein backbones not as 3D coordinates, but as a 1D sequence of discrete 3Di structural tokens from Foldseek[43]. This operation transforms the formidable challenge of 3D geometry generation into a tractable sequence-to-sequence learning problem. A key feature of this token-wise autoregressive model is its ability to assign a computable log-probability $\log p(c|t)$ to each generated sequence of structures—a capability that proves essential for the optimal sampling strategy described below. The discrete structural representation strategy also benefits from the superior scalability demonstrated by recent protein language models[44–46]. T2struct employs an encoder-decoder Transformer architecture: its encoder, a pre-trained PubMedBERT model[47], is specialized for interpreting biological text, while its decoder, a GPT-2 style model trained from scratch, autoregressively predicts the 3Di token sequence (Fig. 1D), connected via a text-3Di cross-attention module.

With a structural blueprint in hand, the second stage is performed by SaProt-T, our text-conditioned sequence designer. This module receives the generated structural tokens and designs a viable amino acid sequence (s) that is conditioned on both the structure and the original textual prompt (Fig. 1d, right). SaProt-T is a modified adaptation of the powerful SaProt[44] model, engineered with a novel mechanism to ensure the final sequence respects the fine-grained functional constraints of the text. This is achieved by projecting the text embedding from the T2struct encoder and prepending it to the input 3Di token sequence (Methods 7.3.1). This combined representation is then processed by SaProt-T, whose cross-attention layers can directly fuse semantic cues from the language with the geometric constraints of the scaffold. The model is trained using a masked language modeling objective, learning to predict the correct amino acid for each structural token position (Extended Data Fig. 5).

The power of Pinal's architecture lies not merely in the sequential execution of the two generation modules, but in their probabilistic unification to elegantly solve the global optimization problem defined in Eq. 1. This is because a naive, greedy approach—first finding the single most probable structure and then designing the best sequence for it—is mathematically suboptimal, as highlighted in Eq. 2. Such a myopic strategy risks overlooking a slightly less probable structure that, crucially, can accommodate a far superior functional sequence. Pinal circumvents this pitfall by employing a global sampling and re-ranking strategy, a process made possible by the fact that both of its two generative stages yield tractable log-probabilities ($\log \prod_{n=1}^{N} p(c_n \mid c_{<n}, t)$ and $\log \prod_{n=1}^{N} p(s_n \mid c, t)$). In practice, we first sample K structural scaffolds from T2struct's distribution, $p(c|t)$. For each of these K scaffolds, we then use SaProt-T to design its optimal corresponding sequence. Finally, these K complete sequence-structure pairs are re-ranked based on their full joint log-probabilities (Eq. 1). This allows Pinal to identify globally superior candidates that a purely sequential, greedy decision process would almost certainly miss (Methods 7.4, Extended Data Fig. 3).

$$\begin{aligned}\arg\max_{s} p(s \mid t) = \arg\max_{s,c} p(s, c \mid t) &= \arg\max_{s,c} \log p(s, c \mid t) = \arg\max_{s,c} (\log p(c \mid t) + \log p(s \mid c, t)) \\ &= \arg\max_{s,c} \Bigg[\underbrace{\sum_{n=1}^{N} \log p(c_n \mid c_{<n}, t)}_{\text{T2struct}} + \underbrace{\sum_{n=1}^{N} \log p(s_n \mid c, t)}_{\text{SaProt-T}} \Bigg]. \end{aligned} \tag{1}$$

$$\textbf{Note:} \quad \arg\max_{s,c} (\log p(c \mid t) + \log p(s \mid c, t)) \neq \left(\arg\max_{c} \log p(c \mid t) + \arg\max_{s} \log p(s \mid c, t) \right). \tag{2}$$

2 Scaling Pinal with massive synthetic data

While scaling laws have revolutionized general AI [46,48,49], their potential in natural language-guided protein generation is untapped due to critical data limitations. The premier human-annotated database, Swiss-Prot, contains merely 0.5 million proteins, far smaller than the billion-scale datasets powering modern language models. Moreover, its template-driven annotations do not reflect the complexity and diversity of queries common in biological research.

To overcome the constraints of scale and diversity, we developed an unprecedented synthetic training dataset containing 1.7 billion text-protein pairs. This comprehensive dataset encompasses over 100 billion word tokens across 135 million unique protein sequences and structures (Methods 7.2, Supplementary Table 2 and Table 4). The dataset construction involved a systematic, multi-faceted approach to generate high-quality natural language descriptions of proteins. For UniProt/TrEMBL entries, we employed the multimodal protein language model ProTrek[50] to predict diverse functional descriptions from multiple perspectives, creating the TrEMBL50-ProTrek subset. These multi-perspective descriptions were then synthesized using large language models (LLMs) to produce coherent, integrated functional annotations, resulting in the TrEMBL50-ProLLM dataset. For Swiss-Prot proteins, we applied a parallel strategy, utilizing LLMs to consolidate multiple functional descriptions into comprehensive annotations (SwissProt-LLM). Additionally, we preserved the original structured Swiss-Prot entries (Swiss-Annot) while generating LLM-paraphrased versions to enhance linguistic diversity (Swiss-Aug). To ensure robust coverage across varying

description lengths and annotation types, we incorporated nearly 800 million keyword-protein pairs from the InterPro database. All protein structural information was sourced from the AlphaFold database (AFDB), providing a unified structural foundation for the entire dataset.

We present the first empirical demonstration of data scaling effects in natural language-guided protein design. Our scaling experiments revealed two key findings that illuminate the relationship between training data size and model performance. First, on challenging test sets with less than 30% sequence identity, model performance—as measured by ProTrek score (Methods 7.5.3, Extended Data Fig. 6) and GT-TMscore—exhibited substantial and consistent improvement with increasing training data size (Fig. 2b, right) (detailed evaluation settings and measures are described in the next section). Second, on standard test sets with less than 50% sequence identity, the ProTrek score demonstrates a steady upward trend throughout the scaling range, while GT-TMscore values plateau rapidly between 100 million and 1 billion word tokens before showing modest gains with further scaling (Fig. 2b, left). These findings demonstrated that synthetic data benefits manifested most prominently under demanding generalization scenarios, where the complexity of the task necessitated larger-scale training data to achieve meaningful performance improvements.

We further investigated Pinal's model scaling trends. Recent studies on protein language models (PLMs) have observed that scaling model parameters may not yield significant benefits for corresponding protein prediction tasks[51,52]. Similarly, on the ProteinGym benchmark[53], ESM2's 15 billion parameter version performs even worse than the 650 million parameter variant. We demonstrated Pinal's model scaling capabilities by training T2struct at three distinct scales: 200 million, 1.2 billion, and 15.5 billion parameters (Supplementary Table 1). As model capacity increased, we observed substantial reductions in validation loss (Fig. 2c, upper). More critically, this improvement translated into superior practical performance: larger models generate proteins with markedly higher predicted foldability (pLDDT scores) and enhanced alignment with language instructions (ProTrek scores) across both standard and stringent test sets (Fig. 2c). Together with our data scaling findings, these experiments established that both model scaling and data scaling serve as fundamental drivers of Pinal's performance achievements.

3 Pinal designs recapitulate native protein signatures

Here, we systematically evaluated proteins designed by Pinal, assessing both their structural correctness and functional relevance. Structural correctness was quantified using folding confidence (pLDDT), Predicted Aligned Error (PAE), and similarity to native structures (TM-score). To address the challenge of evaluating functional relevance—a critical gap in the nascent field of language-guided protein design—we employed the ProTrek protein-to-text matching score (Methods 7.5.3, Extended Data Fig. 6), which measured the semantic distance between protein sequence/structure and their functions.

We benchmarked Pinal against the few existing attempts at language-guided protein design (Methods 7.3.3, Extended Data Fig. 4a). Among them, ProteinDT[54] was the sole preceding model explicitly built for this task, yet it represented a preliminary exploration with a model and dataset orders of magnitude smaller than Pinal's. The other models demonstrated similarly clear limitations: Chroma[55], by the

authors' own account, showed only occasional alignment between designed proteins and intended prompts, while ESM3[46], a concurrent work (by preprint date), was restricted to simple keyword inputs and lacks the ability to interpret complex language. Most critically, no prior work has achieved the goal of generating experimentally validated functional proteins guided solely by natural language. While ESM3 did produce a validated protein, this was not a true language-guided design, as it was conditioned on predefined sequential and structural fragments as input, rather than language.

We designed a multi-faceted evaluation framework comprising three distinct settings: (1) long, descriptive sentences that mimic the complex queries of biologists; (2) short, function-specific phrases to test core generative accuracy; and (3) specialized keywords to enable a direct, rigorous comparison with ESM3. The detailed construction of each test set was outlined in Methods 7.5.1. In tasks involving both long and short sentences, Pinal markedly outperformed ProteinDT and Chroma. Designs from ProteinDT and Chroma failed to align with their textual prompts, yielding ProTrek and TM-scores statistically indistinguishable from a random baseline (Fig. 2d). Furthermore, a significant fraction of their generated sequences failed to fold into plausible structures, as indicated by low pLDDT and high PAE values. In stark contrast, proteins designed by Pinal not only exhibited high textual & functional alignments, with ProTrek scores and TM-scores close to those of real natural proteins, but also possessed structural features (pLDDT, PAE) indicative of stable folding, comparable to those of well-characterized proteins in the Swiss-Prot database (Fig. 2e). Pinal's superiority stemmed from a combination of three key advantages: an advanced architecture, a vastly expanded parameter space (tens of times larger), and a massive training dataset (thousands to tens of thousands times larger than ProteinDT and Chroma). This superiority is consistently observed for both complex, multi-aspect descriptions and concise, function-specific sentences (Fig. 2f, Extended Data Fig. 2).

To further assess its capabilities, we compared Pinal against ESM3 on keyword-conditioned generation. To ensure the most stringent and equitable comparison, we benchmarked Pinal on ESM3's "home ground": generating proteins from InterPro database keywords—the same dataset used to train ESM3—and measuring performance with their established "keyword recovery" metric. Even under these conditions, which inherently favored ESM3, Pinal demonstrated marked superiority (Fig. 2g). Protein designs from Pinal exhibited consistently higher predicted foldability, whereas ESM3's outputs showed considerable variability. Most notably, Pinal achieved substantially better functional alignment, with a keyword recovery rate of 50%—a five-fold improvement over ESM3's approximate 10% rate.

In addition to these findings, A recent independent benchmark study[56] has confirmed Pinal's effectiveness, showing it achieved the best performance on 11 out of 14 metrics focused on sequential plausibility, structural foldability, and language alignment. Compared to prominent models like ProteinDT, Chroma, ESM3, and PAAG[57], Pinal demonstrated performance gains ranging from several-fold to tenfold on multiple metrics (such as ProTrek score, pLDDT, and PAE).

4 Generative design beyond natural evolution

Having established Pinal's accuracy in recapitulating known functions, we next evaluated its capacity for novel protein design. We first investigated whether the model could generate sequence diversity beyond that observed in nature for known protein families. To this end, we challenged Pinal to generate sequence diversity for four distinct families: CoA Ligase, Glucokinase, Aldehyde Dehydrogenase, and Carbonic Anhydrase. These families span a diverse range of catalytic mechanisms, structural folds, and metabolic pathways, which helped test Pinal's generalization capabilities. Phylogenetic analysis revealed that while natural sequences formed tight, evolutionarily conserved clusters, Pinal's designs expanded the known sequence space, populating a vast and diverse phylogenetic landscape (Fig. 3a).

We then assessed Pinal's capacity for structural innovation by designing folds for proteins that are currently experimentally uncharacterized. To investigate this, we prompted the model with functional descriptions of proteins that lack any known structural homologue in the PDB (Methods 7.5.2). The results in Fig. 3b demonstrate a clear capacity for structural innovation: a substantial population of Pinal's designs clustered in the high-confidence foldability region (high pLDDT), while simultaneously showing low sequence identity (<30%) and low structural similarity (TM-score < 0.5) to any known protein or fold. Pinal was effectively designing in a truly uncharted territory of the protein fold space.

To confirm that this computational novelty translated to physical reality, we selected 20 novel proteins for experimental synthesis.We found that 17 of these 20 proteins (85%) were successfully expressed in soluble forms, a strong indicator of stable folding (Supplementary Fig. 1 and Supplementary Table 6). These results provide definitive experimental evidence that Pinal has internalized fundamental principles of protein stability. It can move beyond diversifying existing superfamilies to authoring structurally new, physically viable protein architectures without relying on structural templates, using nothing more than a language prompt (Fig. 3c).

5 Experimental realization of diverse, active proteins

Having demonstrated that Pinal can generate novel, stable protein architectures, we next sought to determine if this generative capability extends to creating biochemical function. To this end, we designed a validation framework targeting four proteins of significant scientific and industrial importance (Fig. 4). Each target was chosen to probe a distinct class of biochemical complexity: the intrinsic photophysics of a GFP; the enzymatic catalysis of a synthetic polymer by a PETase; the cofactor-dependent mechanism of an ADH; and the molecular coordination of an H-protein within a multi-protein system. For each, the generative process of Pinal was guided solely by a high-level text prompt describing the desired function (Supplementary Table 7), with no structural or sequential templates provided as input. The outcomes of these four distinct challenges, presented below, provide a rigorous test of Pinal's capabilities as a universal, programmable protein design engine.

Our first challenge for Pinal was the computational design of a novel GFP, a classic benchmark of protein folding and function[58]. Guided solely by a text prompt describing autofluorescence (Supplementary

Table 7), Pinal generated an initial set of 1,000 sequences, which a multi-tiered computational pipeline filtered to 15 promising candidates (Extended Data Fig. 7a). These sequences shared $< 30\%$ identity with known fluorescent proteins but were all confidently predicted to adopt the characteristic β-barrel architecture (Fig. 5b). To confer function, we computationally grafted the essential "TYG" chromophore-forming motif into these novel scaffolds and synthesized all 15 designs[58,59]. Remarkably, 11 of the 15 proteins (73%) expressed in a soluble form. Screening this pool of soluble proteins for function revealed one highly active variant, pinalGFP, which produced green fluorescence (460 nm, 510 nm) visible to the naked eye (Fig. 5a and Extended Data Fig. 7c). Quantitatively, pinalGFP achieved 50% of the fluorescence intensity of the highly-optimized esmGFP[46] and 20% of the wildtype GFP (avGFP) within 24 hours. Notably, this performance under minimal constraints surpassed the equivalent design-esmGFP-b8 from ESM3, which was non-fluorescent at the same 24-hour time point due to slow maturation (Fig. 5c). Furthermore, analytical gel filtration showed that while avGFP and esmGFP are primarily dimeric, pinalGFP exists in a more complex oligomeric state, with fluorescence confined to the dimeric fraction exhibiting a quantum yield of 0.44 (Extended Data Fig. 7c-d). This observation highlights a clear avenue for future optimization.

Next, we escalated the challenge from designing a functional protein to creating a novel enzyme for catalysis. To this end, we targeted a PETase to degrade the recalcitrant plastic polyethylene terephthalate (PET). Pinal generated 1,000 sequences, from which we selected 30 candidates that contained both the canonical Ser-His-Asp catalytic triad and a key substrate-binding motif ("SMGGG-D-H") required to form a dynamic active site (Extended Data Fig. 8a and (Supplementary Table 10))[60]. Upon expression in *E. coli*, these designs proved highly viable: 25 of the 30 candidates (83%) were soluble, a majority exhibiting high thermostability (Supplementary Table 10)—a critical attribute for industrial applications. This high rate of structural integrity was matched by functional success: 15 of the 25 soluble proteins (60%) displayed clear esterase activity against a model substrate-4-nitrophenyl butyrate, confirming the viability of the designed active sites within the novel scaffolds (Fig. 5d). We then advanced the four most active variants (pinalPETase01/02/03/07) to assess their performance on amorphous PET films[61]. While all four were active, one variant, pinalPETase01, stood out for its superior and sustained performance. Over a seven-day period, it continuously degraded the polymer, yielding 5.4 mM of degradation products (Fig. 5e). In contrast, the activity of the wild-type benchmark, *Is*PETase, plateaued after the first day, indicating a loss of function. This sustained catalysis by pinalPETase01 resulted in visible, macroscopic degradation of the PET film, underscoring its efficacy and stability (Fig. 5f and Extended Data Fig. 8b).

Having demonstrated the design of a novel PETase, we next tested Pinal's ability to navigate the complexity of a large enzyme superfamily and generate diverse solutions from its distinct sub-classes. For this, we targeted ADH, a versatile class of biocatalysts for producing chiral alcohols that are essential building blocks for pharmaceuticals[62]. Pinal was provided with a single, general prompt for an NAD^+-dependent ADH (Supplementary Table 7), imposing no constraints on metal dependence or chain length. From 1,000 generated sequences, we computationally filtered for candidates with high predicted structural confidence and correctly formed NAD^+ binding pockets, selecting eight for experimental validation (Extended Data Fig. 9a and Supplementary Table 9). Of these, five were successfully expressed and purified. Notably, these

five functional designs spanned three distinct ADH subfamilies: Fe^{2+}-dependent long-chain (pinalADH02, pinalADH05), medium-chain (pinalADH04), and short-chain variants (pinalADH01, pinalADH03) (Fig. 5g). In silico analysis confirmed that the predicted folds and catalytic metal coordination were consistent with their respective subfamilies (Extended Data Fig. 9b). This structural diversity also resulted in varied biophysical properties. For instance, the enzymes exhibited different optimal reaction temperatures, with the short-chain variant performing best at 37 °C (Fig. 5h and Extended Data Fig. 9c). This result demonstrates Pinal's ability to capture the core functional principles of distinct enzyme subfamilies, allowing it to generate active catalysts across different protein architectures.

As the culminating test of Pinal's capabilities, we challenged it to move beyond single-enzyme function and design a protein requiring precise coordination within a complex multi-protein system. We targeted the H-protein, a key shuttle protein within the glycine cleavage system[63,64], whose function requires both correct post-translational lipoylation and dynamic interactions with its partner enzymes to become active (Fig. 5i). From a single functional prompt, Pinal generated 1,000 sequences, from which we selected 20 for gene synthesis and experimental validation (Extended Data Fig. 10a and Supplementary Table 11). The designs yielded a high success rate, with 14 of the 20 candidates (70%) showing soluble expressions. When tested for CO_2 fixation in a reconstituted assay, five of the designed proteins were active (Extended Data Fig. 10b). The top-performing variant, pinalH02, drove glycine synthesis to a final titer of 11.1 mM, a 1.7-fold increase over the native *E. coli* H-protein (ecH), highlighting its potential for robust biomanufacturing applications (Fig. 5j)[65]. Molecular dynamics simulations revealed a clear mechanistic basis for this superior performance. While the native ecH locks its reactive lipoamide arm in a protective cavity for regulatory purposes, the arm of pinalH02 is largely liberated. This key architectural difference bypasses the native protein's rate-limiting step, facilitating more efficient intermediate shuttling and leading to enhanced carbon fixation (Fig. 5k and Extended Data Fig. 10c-f). Successfully designing a protein that not only functions within a complex enzymatic network but also surpasses its natural counterpart demonstrates Pinal's capacity to generate bespoke, high-performance components for advanced cell-free biomanufacturing[66].

6 Discussion

Pinal's significance lies not in its performance alone but, more fundamentally, in its realization of a new paradigm for molecular engineering. For decades, protein design has been the domain of specialists, requiring a deep, intuitive understanding of biophysics to translate a desired function into the low-level language of structural coordinates and residue mutations. Our work demonstrates that this paradigm can be shifted. By leveraging a generative framework that comprehends high-level natural language, we effectively lower the barrier to entry, transforming protein design from a specialized craft into a programmable discipline. The dialogue is no longer about atoms and angles, but about function and intent, democratizing the ability to create novel functional matter.

Our results indicate that two core principles are key to achieving this vision. The first is architectural: the problem of directly mapping the vast, combinatorial space of language to the even vaster space of protein

sequences is intractable. Our decoupled, two-stage framework, which uses protein structure as a compressed, semantic, and geometric bridge (t $\rightarrow$ c $\rightarrow$ s), is a crucial innovation that renders the problem tractable. The second principle is scale. Our work provides compelling evidence that the scaling effects observed in general AI are applicable to the complex task of language-guided protein design. The performance of our 16-billion-parameter model, enabled by an unprecedented 1.7-billion-pair synthetic dataset, confirms that massive-scale training is essential for mastering the intricate grammar that links human language to the language of life.

Perhaps the most profound implication of our work is the bridge it builds between abstract concepts and functional, physical reality. The successful experimental realization of four distinct classes of proteins provides definitive validation for the generality of this paradigm. Notably, this set includes the first catalytically active enzymes, designed from nothing more than textual descriptions. In addition, the fact that our Pinal-designed H-protein functionally outperforms its natural counterpart reveals a new principle of biological design: that the optimal solution to a biological problem may not have been discovered by natural evolution, but can be authored directly from human language. We are no longer limited to rediscovering or incrementally improving what nature has provided; we can now compose entirely new biological solutions.

While this work provides a definitive proof-of-concept, we acknowledge the frontiers that remain. Pinal's current architecture, trained primarily on single polypeptide chains, is not yet optimized for designing multi-protein complexes or their intricate interfaces. Furthermore, like all data-driven models, its capabilities are constrained by the current landscape of biological annotation. This limits fine-grained quantitative control—for instance, specifying a precise emission wavelength for a fluorescent protein remains a challenge—and highlights a fundamental need for richer, more quantitative community-wide datasets. The experimental validation presented here was also necessarily conservative. A crucial next step will be to couple Pinal's generative power with high-throughput functional screening to fully map the boundaries of this new design space and uncover the full diversity of solutions it contains.

Looking forward, the ability to program proteins with language does not merely accelerate existing research; it fundamentally changes its nature. The paradigm shifts from the slow, manual discovery of molecular solutions to their direct, conceptual design. Future iterations of this technology could move beyond interpolating known functions towards true creative reasoning, tackling prompts that require an understanding of first-principles chemistry and physics to invent entirely new catalytic mechanisms. We envision Pinal as a foundational step towards a "biological compiler" where biological function itself can be programmed directly from high-level intent. When integrated with AI systems capable of autonomously proposing novel scientific hypotheses, this approach could enable fully automated, closed-loop discovery cycles. This is the ultimate promise of our work: to establish that human language can serve as the source code for biology, unlocking a universe of bespoke therapeutics, sustainable biocatalysts, and smart materials, authored not by evolution, but by imagination.

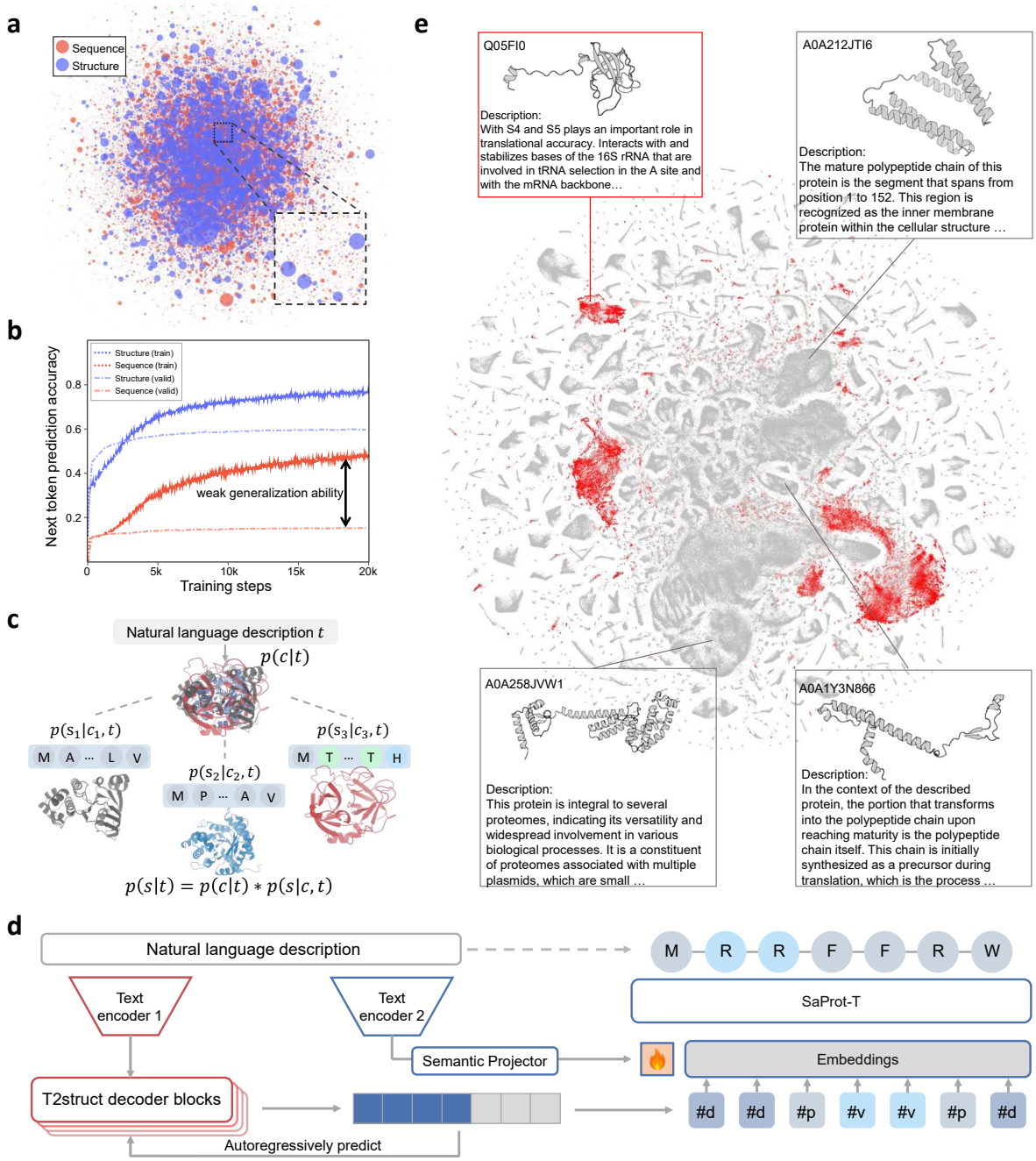


Fig. 1: **Pinal is a generative foundation model that leverages a two-stage design process and massive training data to master the text-to-protein link. a**: Protein sequence versus structure space in the Swiss-Prot database. The 65,733 sequence clusters (at 30% identity) map to only 21,368 structure clusters, demonstrating greater structural conservation. Each circle is a cluster; its size is proportional to the protein count. **b**: Next token prediction accuracy curves trained on the Swiss-Prot dataset. The blue results represent text-to-structure generation, while the red color results represent text-to-sequence generation. The training and validation sets are split by 30% sequence identity. "train" and "valid" are accuracy on the training set and validation set respectively. **c**: The chain rule of probability enables the decomposition of text-to-protein design modeling into two sequential stages, mediated by an intermediate structural representation. **d**: Pinal model architecture. The diagram illustrates the data flow (arrows) and module attribution. Modules belonging to the T2struct component are outlined in red, while those belonging to the SaProt-T component are outlined in blue. **e**: Visualization of the distribution of human-reviewed Swiss-Prot data (545K proteins, 27M protein-text pairs, red) and our synthetic dataset (135M proteins, 1.7B protein-text pairs, gray) in a low-dimensional space, obtained by t-SNE reduction of the ProTrek embeddings. This visualization demonstrates how the synthetic dataset vastly expands the coverage of the protein design space, encompassing and extending beyond regions covered by the scarce human-annotated data.

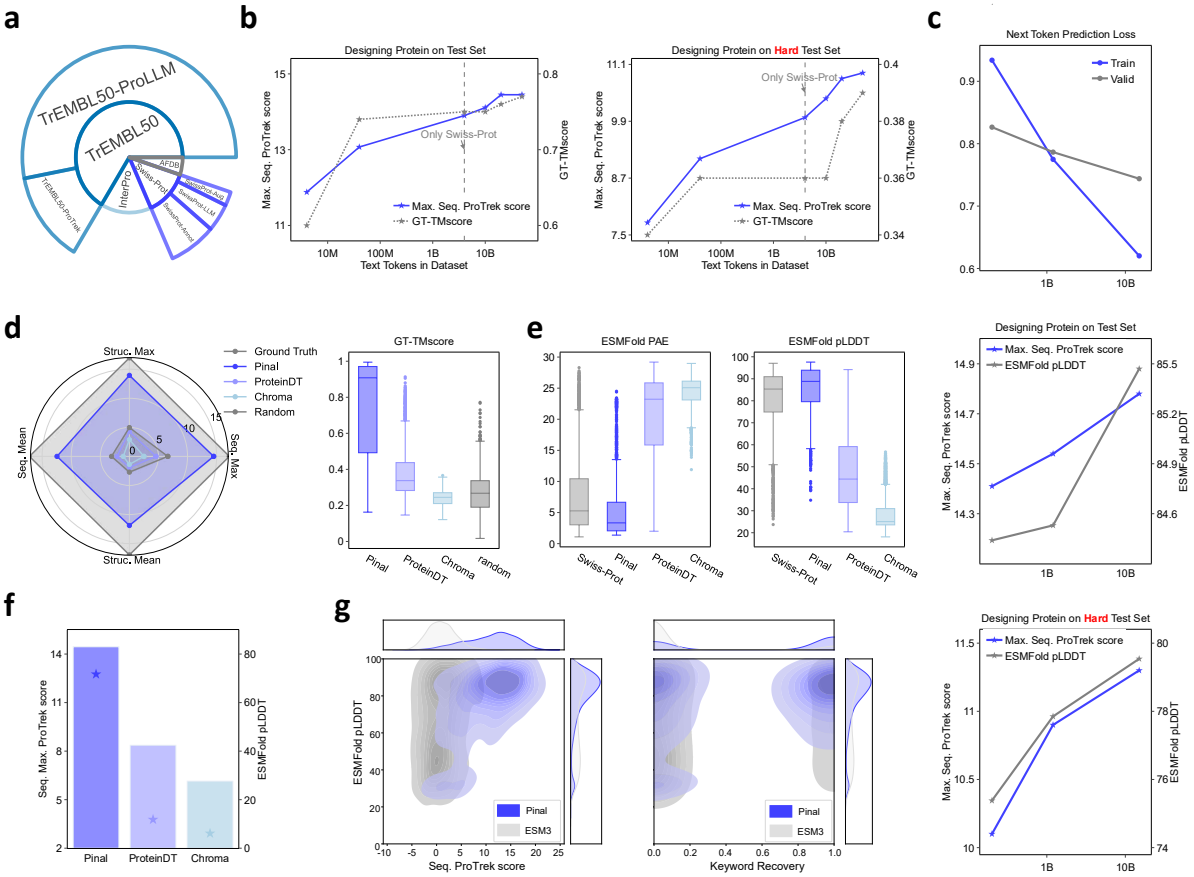


Fig. 2: **Scaling T2struct and benchmarking Pinal's design performance.** **a**: Source of Pinal training datasets. **b**: Protein design performance improves as training data size increases, with the x-axis plotted on a log scale. **c**: Scaling behavior of T2struct from 200 million to 15 billion parameters, with the x-axis plotted on a log scale; both test loss and the quality of designed proteins improve with larger parameter sizes. The three plots are arranged vertically. **d**: Alignment of designed proteins with given language descriptions, assessed by ProTrek score (sequence-to-text and structure-to-text) and GT-TMscore. The the maximum (Max) and mean (Mean) ProTrek score values are reported. **e**: ESMFold evaluates the foldability of designed proteins. **f**: The foldability and textual alignment of designed proteins conditioned on short sentences. Foldability, measured by ESMFold pLDDT, is represented by the bars (right axis), while textual alignment, measured by the max. sequence ProTrek score, is indicated by the stars (left axis). **g**: Comparing with ESM3. In both panels, deeper colors in the upper right corner indicate stronger model performance.

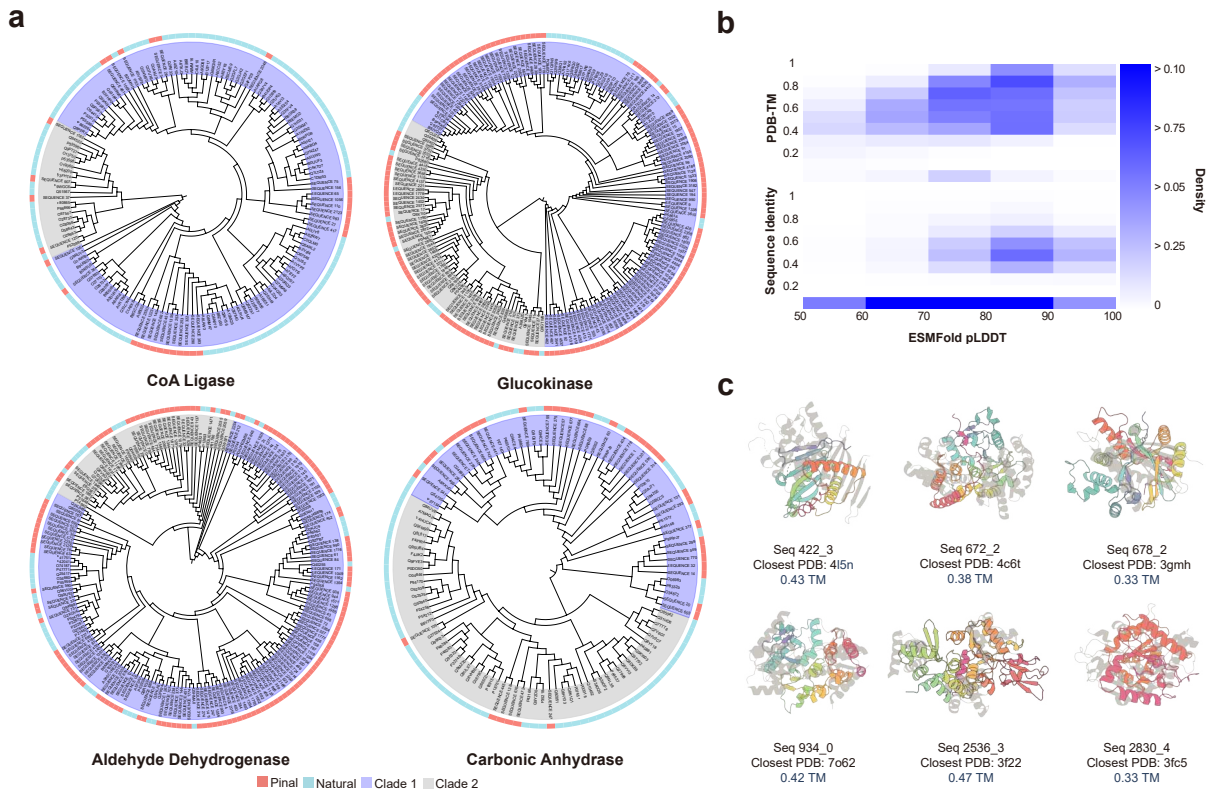


Fig. 3: **Pinal generates a diverse landscape of novel proteins. a,** Phylogenetic analysis comparing Pinal-designed sequences with their natural counterparts for four distinct enzyme families. For each family, the natural sequences form conserved clusters, whereas Pinal's designs are broadly distributed across the phylogenetic space, demonstrating its ability to generate high sequence diversity. **b,** Heatmaps illustrate the relationship between predicted structural confidence (ESMFold pLDDT, X-axis) and protein novelty. The top panel shows structural novelty (TM-score to known proteins in PDB, Y-axis), and the bottom panel shows sequence novelty (sequence identity to known proteins in UniRef100, Y-axis). Darker shades indicate higher density of designed proteins. **c,** Representative examples of novel protein structures designed by Pinal. For each design, the closest structural homologue in the PDB is shown, along with a low TM-score, providing visual confirmation of Pinal's ability to generate previously unknown protein folds.

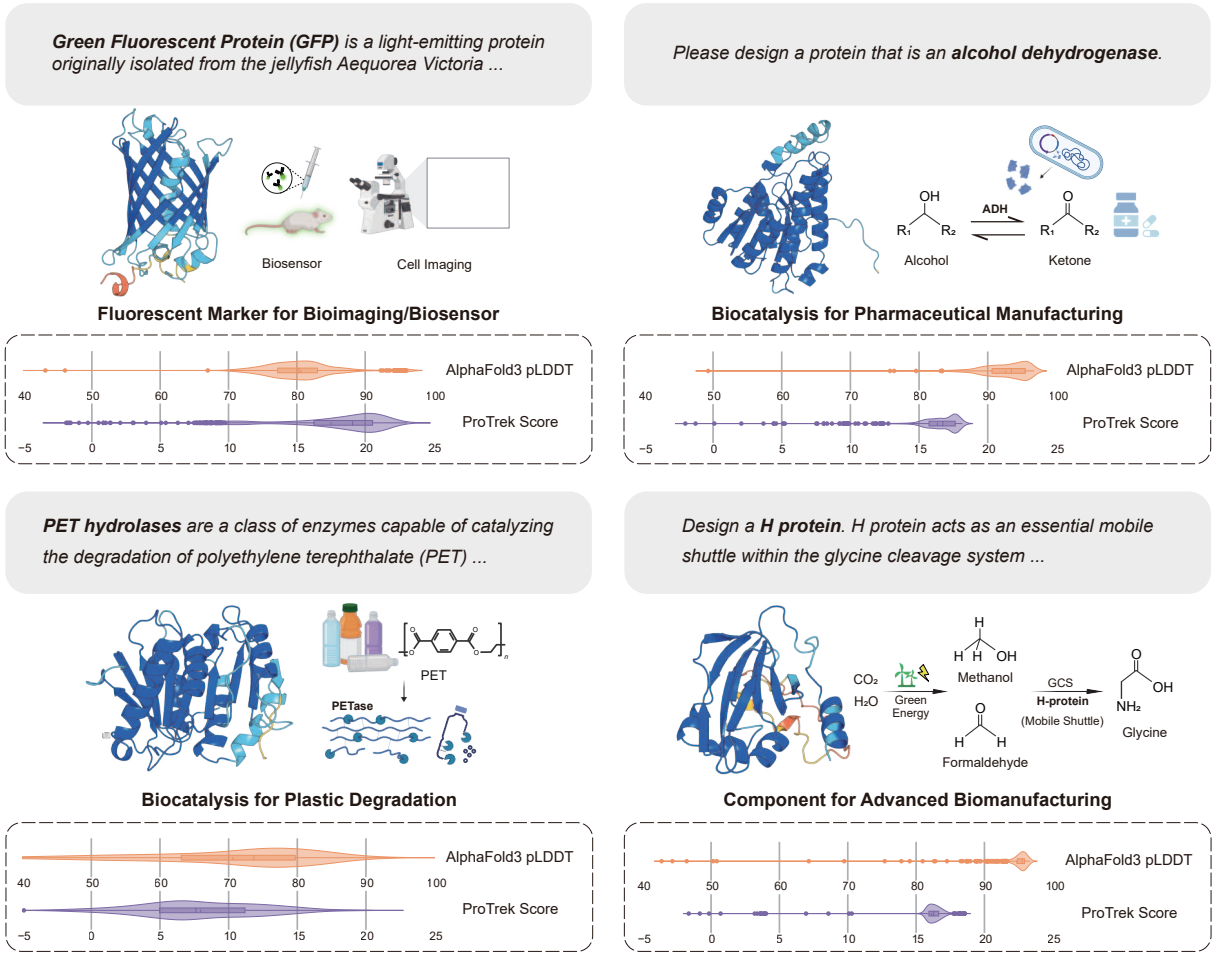


Fig. 4: **Pinal translates diverse functional prompts into high-quality protein candidates without filtering.** The conceptual application, natural language prompt, and final design are presented for GFP, ADH, PETase, and H-protein. Violin plots show the high quality of the candidates, with high AlphaFold3 pLDDT scores indicating structural confidence and high ProTrek scores demonstrating close functional alignment with the prompt.

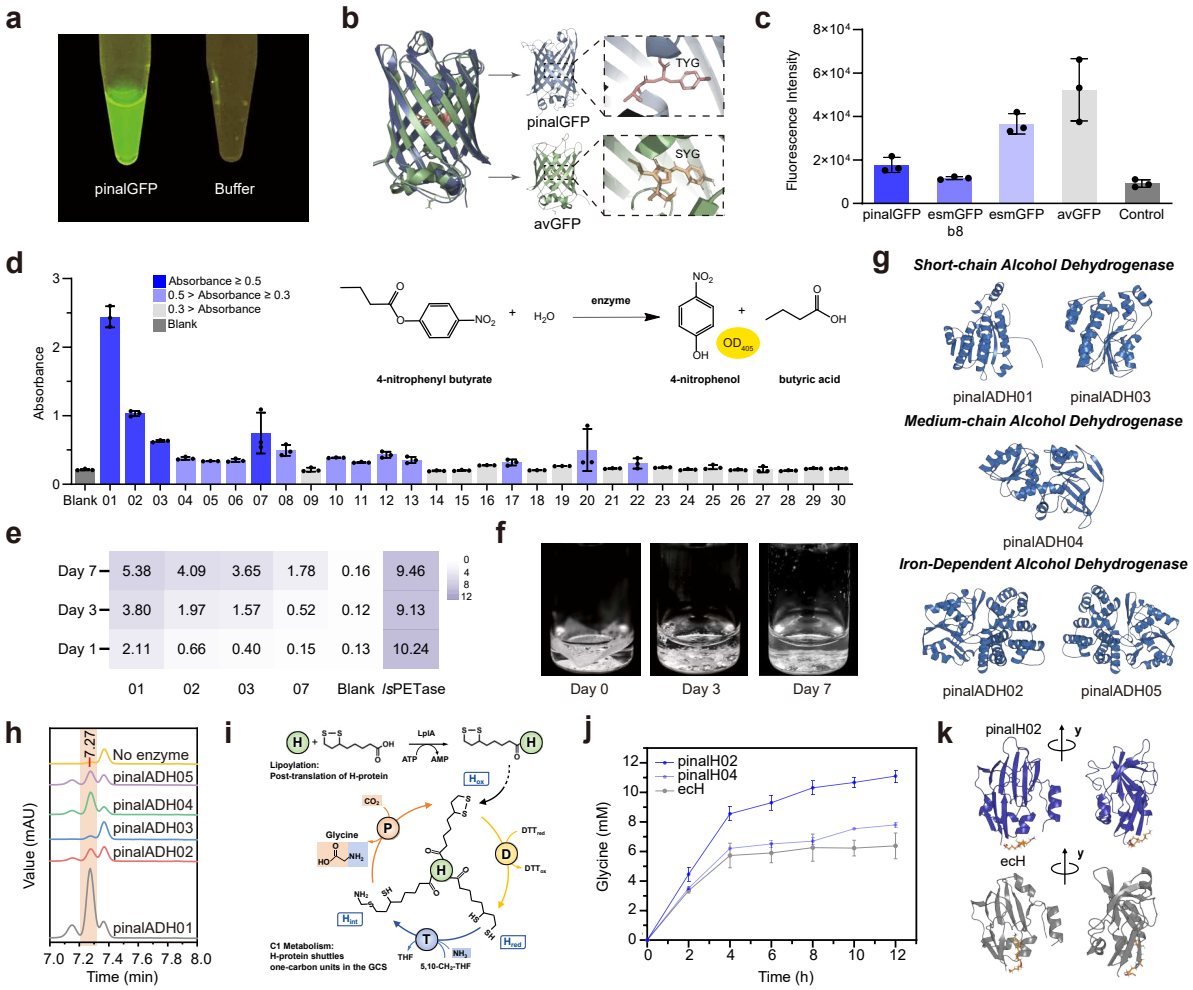


Fig. 5: **Experimental validation of Pinal-designed proteins across four distinct functional classes. a,** Visual evidence of purified pinalGFP (5 mg mL^{-1}) under 365 nm UV illumination demonstrating the characteristic green fluorescence of functional variants. **b,** Structural alignment of pinalGFP (blue) with avGFP (green, PDB: 2WUR) showing conservation of the β-barrel architecture. The engineered chromophore residues (T65-Y66-G67) are highlighted in the pinalGFP structure. **c,** Quantitative fluorescence analysis of GFP variants expressed in *E. coli* BL21(DE3) cells 24 hours post-induction. Fluorescence intensity was normalized to cell density (OD_{600}). Data are shown as mean ± s.d. (n = 3). **d,** High-throughput screening of 30 Pinal-designed PETase candidates for esterase activity. Activity was measured by monitoring the hydrolysis of the model substrate 4-nitrophenyl butyrate to 4-nitrophenol, quantified by absorbance at 405 nm. The reaction scheme is shown above. Data are shown as mean ± s.d. (n = 3). **e,** Time-course analysis of PET degradation, showing the concentration (μM) of total soluble products (MHET and TPA) released from amorphous PET films by the top four PETase candidates and wild-type IsPETase over seven days. **f,** Visual evidence of macroscopic degradation of an amorphous PET film by the most active variant, pinalPETase01, over a seven-day incubation period. **g,** Predicted structural models of the five experimentally validated ADH designs, illustrating the structural diversity across short-chain, medium-chain, and iron-dependent long-chain subfamilies. **h,** Confirmation of catalytic activity for the five purified ADH variants. The chromatogram overlay shows the conversion of the substrate by each enzyme compared to a no-enzyme control. **i,** H lipoylation to generate H_{ox}, which carries a lipoyl arm on the surface of H protein. The glycine cleavage system (GCS) runs in reverse to synthesize glycine from CO_2. H is the central shuttle protein that accepts the aminomethyl group from T protein and delivers it to the P protein, which then catalyzes the fixation of CO_2 to synthesize glycine. LplA, lipoate-protein ligase A; P, glycine decarboxylase, EC 1.4.4.2; T, aminomethyltransferase, EC 2.1.2.10; DTT_{ox} and DTT_{red}, oxidized and reduced forms of dithiothreitol; 5,10-CH_2-THF, N^5, N^{10}-methylene-tetrahydrofolate; H_{int}, aminomethylated form of the lipoylated H protein; H_{ox} and H_{red}, oxidized and reduced form of the lipoylated H protein, respectively. **j,** Glycine production based on pinalH02, pinalH04 and ecH from *E. coli*. Data are mean ± s.d. (n = 3). **k,** The confirmation of pinalH02 and ecH at the time point of 200 ns in MD simulations. For natural H protein, the aminomethyl lipoate arms are locked in the hydrophobic cavity.

References

[1] Frances H Arnold. Directed evolution: bringing new chemistry to life. *Angewandte Chemie (International Ed. in English)*, 57(16):4143, 2017.

[2] Yajie Wang, Pu Xue, Mingfeng Cao, Tianhao Yu, Stephan T Lane, and Huimin Zhao. Directed evolution: methodologies and applications. *Chemical reviews*, 121(20):12384–12444, 2021.

[3] Jason Yang, Ravi G Lal, James C Bowden, Raul Astudillo, Mikhail A Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, 2025.

[4] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[5] Huimin Zhao, Lori Giver, Zhixin Shao, Joseph A Affholter, and Frances H Arnold. Molecular evolution by staggered extension process (step) in vitro recombination. *Nature biotechnology*, 16(3):258–261, 1998.

[6] Willem PC Stemmer. Rapid evolution of a protein in vitro by dna shuffling. *Nature*, 370(6488):389–391, 1994.

[7] George P Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.

[8] Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *science*, 302(5649):1364–1368, 2003.

[9] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

[10] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B Acton, Gaetano T Montelione, and David Baker. Principles for designing ideal protein structures. *Nature*, 491(7423):222–227, 2012.

[11] Sarel J Fleishman, Timothy A Whitehead, Damian C Ekiert, Cyrille Dreyfus, Jacob E Corn, Eva-Maria Strauch, Ian A Wilson, and David Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.

[12] Neil P King, William Sheffler, Michael R Sawaya, Breanna S Vollmar, John P Sumida, Ingemar André, Tamir Gonen, Todd O Yeates, and David Baker. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science*, 336(6085):1171–1174, 2012.

[13] Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17(7):665–680, 2020.

[14] Kristian W Kaufmann, Gordon H Lemmon, Samuel L DeLuca, Jonathan H Sheehan, and Jens Meiler. Practically useful: what the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010.

[15] Greg Winter, Andrew D Griffiths, Robert E Hawkins, and Hennie R Hoogenboom. Making antibodies by phage display technology. *Annual review of immunology*, 12(1):433–455, 1994.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[17] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[18] Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

[19] Yufeng Liu, Lu Zhang, Weilun Wang, Min Zhu, Chenchen Wang, Fudong Li, Jiahai Zhang, Houqiang Li, Quan Chen, and Haiyan Liu. Rotamer-free protein sequence design based on deep learning and self-consistency. *Nature computational science*, 2(7):451–462, 2022.

[20] Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nature biotechnology*, 42(2):216–228, 2024.

[21] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[22] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[23] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Altae-Tran, et al. Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv*, pages 2025–04, 2025.

[24] Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, pages 1–10, 2025.

[25] Susana Vázquez Torres, Melisa Benard Valle, Stephen P Mackessy, Stefanie K Menzies, Nicholas R Casewell, Shirin Ahmadi, Nick J Burlet, Edin Muratspahić, Isaac Sappington, Max D Overath, et al. De novo designed proteins neutralize lethal snake venom toxins. *Nature*, pages 1–7, 2025.

[26] Yufeng Liu, Sheng Wang, Jixin Dong, Linghui Chen, Xinyu Wang, Lei Wang, Fudong Li, Chenchen Wang, Jiahai Zhang, Yuzhu Wang, et al. De novo protein design with a denoising diffusion network independent of pretrained structure prediction models. *Nature Methods*, 21(11):2107–2116, 2024.

[27] Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, 2023.

[28] Isaac D Lutz, Shunzhi Wang, Christoffer Norn, Alexis Courbet, Andrew J Borst, Yan Ting Zhao, Annie Dosey, Longxing Cao, Jinwei Xu, Elizabeth M Leaf, et al. Top-down design of protein architectures with reinforcement learning. *Science*, 380(6642):266–273, 2023.

[29] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[30] Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023.

[31] Bin Huang, Yang Xu, Xiuhong Hu, Yongrui Liu, Shanhui Liao, Jiahai Zhang, Chengdong Huang, Jingjun Hong, Quan Chen, and Haiyan Liu. A backbone-centred energy function of neural networks for protein design. *Nature*, 602(7897):523–528, 2022.

[32] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106, 2023.

[33] Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.

[34] Geraldene Munsamy, Ramiro Illanes-Vicioso, Silvia Funcillo, Ioanna T Nakou, Sebastian Lindner, Gavin Ayres, Lesley S Sheehan, Steven Moss, Ulrich Eckhard, Philipp Lorenz, et al. Conditional language models enable the efficient design of proficient enzymes. *bioRxiv*, pages 2024–05, 2024.

[35] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[36] Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

[37] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

[38] Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876, 2009.

[39] Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3):e0282689, 2023.

[40] Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.

[41] Anna Lauko, Samuel J Pellock, Kiera H Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alex Shida, Andrew Hunt, Indrek Kalvet, et al. Computational design of serine hydrolases. *Science*, page eadu2454, 2025.

[42] Kristoffer Illergård, David H Ardell, and Arne Elofsson. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

[43] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022.

[44] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

[45] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[46] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, page eads0018, 2025.

[47] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[48] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[49] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

[50] Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. Protrek: Navigating the protein universe through tri-modal contrastive learning. *Nature Biotechnology (accepted for publication*, pages 2024–05, 2024.

[51] Francesca-Zhoufan Li, Ava P Amini, Yisong Yue, Kevin K Yang, and Alex X Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pages 2024–02, 2024.

[52] Quentin Fournier, Robert M Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: is scaling necessary? *bioRxiv*, pages 2024–09, 2024.

[53] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

[54] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *Nature Machine Intelligence*, pages 1–12, 2025.

[55] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

[56] Jiahao Kuang, Nuowei Liu, Changzhi Sun, Tao Ji, and Yuanbin Wu. Pdfbench: A benchmark for de novo protein design from function. *arXiv preprint arXiv:2505.20346*, 2025.

[57] Chaohao Yuan, Songyou Li, Geyan Ye, Yikun Zhang, Long-Kai Huang, Wenbing Huang, Wei Liu, Jianhua Yao, and Yu Rong. Annotation-guided protein design with multi-level domain alignment. *arXiv preprint arXiv:2404.16866*, 2024.

[58] Karen S. Sarkisyan et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[59] M. Ormo, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien, and S. J. Remington. Crystal structure of the aequorea victoria green fluorescent protein. *Science*, 273:1392–1395, 1996.

[60] Wei Liu, Chuang Li, Bin Li, Liying Zhu, Dengming Ming, and Ling Jiang. Structure-guided discovery and rational design of a new poly(ethylene terephthalate) hydrolase from alphafold protein structure database. *Journal of Hazardous Materials*, 480:136389, 2024.

[61] J. Chen, Y. Jia, Y. Sun, et al. Global marine microbial diversity and its potential in bioprospecting. *Nature*, 633:371–379, 2024.

[62] Hongyuan Lu, Shiqin Yu, Fengyu Qin, Wenbo Ning, Xiaoqiang Ma, Kaiyuan Tian, Zhi Li, and Kang Zhou. A secretion-based dual fluorescence assay for high-throughput screening of alcohol dehydrogenases. *Biotechnology and Bioengineering*, 118(4):1605–1616, 2021.

[63] C. Cohen-Addad, S. Pares, L. Sieker, M. Neuburger, and R. Douce. The lipoamide arm in the glycine decarboxylase complex is not freely swinging. *Nature Structural Biology*, 2:63–68, 1995.

[64] G. Kikuchi, Y. Motokawa, T. Yoshida, and K. Hiraga. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.*, 84:246–263, 2008.

[65] Z. Liu, K. Wang, Y. Chen, T. Tan, and J. Nielsen. Third-generation biorefineries as the means to produce fuels and chemicals from co2. *Nature Catalysis*, 3:274–288, 2020.

[66] J. Liu, H. Zhang, Y. Xu, H. Meng, and A.-P. Zeng. Turn air-captured co2 with methanol into amino acid and pyruvate in an atp/nad(p)h-free chemoenzymatic system. *Nature Communications*, 14:2772, 2023.

7 Methods

7.1 Detailed model architecture

Pinal's generative capability is realized through a two-stage network composed of two main modules: T2struct and SaProt-T.

T2struct is an encoder-decoder Transformer that translates natural language descriptions into structural (3Di) token sequences (Extended Data Fig. 1a). The text encoder, initialized from a pretrained T5 model[67], first processes the input prompt to generate a high-level text representation. This representation then guides the T2struct decoder, a classical GPT-2-style[68] autoregressive model composed of a stack of N identical blocks. As shown in the detailed view, each block sequentially refines the structural representation using three core sub-modules. First, the structural context module, implemented as a masked multi-head self-attention layer, models the internal dependencies of the generated backbone. Next, the text-conditioning module, a multi-head cross-attention layer, integrates semantic guidance by querying the text representation to ensure alignment with the prompt's intent. Finally, the feature refinement network, a pointwise feed-forward block, applies further non-linear transformations before the output is passed to the next decoder block.

The second stage of the Pinal network, SaProt-T, is an architectural adaptation of the SaProt model designed to generate a protein sequence from a structural scaffold and a text prompt (Extended Data Fig. 1b). To condition the generation on language, a separate text encoder (a pretrained PubMedBERT[47] model) produces a text representation, which is then passed to a semantic projector—a trainable linear layer—to map the text embedding into the protein embedding space. This projected text vector is then prepended to the sequence of input structural token embeddings to form a text-protein joint representation. This joint representation is subsequently processed by a stack of N SaProt-T blocks, each composed of self-attention and feed-forward modules, to predict the final amino acid sequence in a single forward pass.

7.2 Dataset construction

Training a large-scale, text-guided protein design model requires extensive and diverse datasets. However, a fundamental challenge exists: high-quality, manually curated databases such as Swiss-Prot are limited in scale (containing 0.5 million proteins), making them insufficient for training billion-parameter models. Conversely, large sequence databases like UniProt/TrEMBL[70], while containing hundreds of millions of proteins, lack reliable text annotations.

To overcome this data gap, we devised a comprehensive, multi-pronged strategy to construct a hybrid dataset. Our approach encompasses two primary components: (1) augmenting the high-quality Swiss-Prot annotations using large language models (LLMs) to improve their diversity and comprehensiveness, and (2) generating massive-scale synthetic text-protein pairs from extensive, less-annotated protein database such as UniProtKB/TrEMBL and AlphaFold database (AFDB). This strategy yields a final dataset of over 1.7 billion text-protein pairs, the composition of which is detailed in the following subsections and summarized in Supplementary Table 2.

7.2.1 SwissProt-Annot

Following the procedure in ProTrek[50], we construct the SwissProt-Annot dataset by extracting protein function information from 56 distinct subsections, including protein-level (*e.g.* Function, Miscellaneous, Caution) and residue-level (*e.g.* Active site, Binding site, DNA binding), from Swiss-Prot database (Release 2023_03). We employ GPT-4 to enhance the linguistic diversity of our training data, using distinct strategies tailored to the format of the Swiss-Prot annotations. For each protein-level description, we generate 10 paraphrases. For the residue-level annotations, we prompt GPT-4 to generate a diverse pool of sentence templates, which are then populated with specific annotations to form a variety of complete sentences (see Supplementary Table 5 for the generation prompt and Supplementary Table 4 for an illustrative example).

To provide the model with natural language descriptions of enzymatic function, rather than just symbolic chemical equations from Swiss-Prot, we incorporated the comprehensive textual annotations for each EC number from the ENZYME database[72]. Specifically, we download data from the official website[1] and extract the descriptions and comments (*i.e.*, the "DE" and "CC" lines). This process results in an additional 253,279 protein-text pair data for SwissProt-Annot.

7.2.2 SwissProt-Aug

While the SwissProt-Annot dataset is highly structured, it typically presents each piece of functional information as isolated sentences. To create more comprehensive and coherent functional descriptions, we integrate these individual sentences into unified narratives. We first establish 56 canonical questions, each corresponding to one of the 56 subsections of a Swiss-Prot entry. To ensure linguistic diversity, we then use GPT-4 to generate 100 distinct paraphrases for each canonical question. Subsequently, we prompt the *glm-4-flash* model[73] with a prompt constructed from two components: a domain-specific question (selected from the paraphrases) and the collection of existing descriptions for a given protein from SwissProt-Annot. The language model synthesizes the provided descriptions to generate a comprehensive answer, which serves as the new functional description for the protein. The complete prompts used in this process are detailed in Supplementary Table 5.

7.2.3 TrEMBL50-ProTrek

To address the scale limitations of the high-quality Swiss-Prot database, we generate a massive-scale dataset by pairing proteins in TrEMBL with descriptions from SwissProt-Annot. Specifically, we first cluster the TrEMBL database at 50% sequence identity, yielding a set of representative proteins (hereafter referred to as TrEMBL50). Next, we construct a comprehensive text pool by aggregating all functional descriptions from SwissProt-Annot. Finally, for each protein in TrEMBL50, we employed the ProTrek model to calculate the protein-text matching scores. By selecting the top 10 descriptions for each protein, this process yielded the TrEMBL50-ProTrek dataset, containing approximately 412 million synthetic protein-text pairs. Additionally, we made targeted additions for proteins with limited functional annotations in the training set. We curated 46 key protein classes (Supplementary Table 1) and used their names as textual queries for

[1] https://ftp.expasy.org/databases/enzyme/enzyme.dat

ProTrek to search against the TrEMBL50 database, generating approximately 22,000 additional synthetic protein-text pairs.

7.2.4 SwissProt-LLM and TrEMBL50-ProLLM

We incorporated two datasets introduced by Evolla[74], which contain LLM-generated question-answer pairs for proteins from Swiss-Prot and TrEMBL50, respectively. We extracted only the answers from each pair to serve as functional descriptions, creating two datasets: SwissProt-LLM and TrEMBL50-ProLLM.

7.2.5 Keyword-protein pairs from InterPro database

To further expand the diversity of textual prompts, we incorporate keyword-protein pairs from the InterPro database (version 101.0)[2]. We collect these pairs from InterProScan annotations, using the functional keywords directly as a concise form of textual input. This dataset ensures the model is exposed to a wide range of short, function-centric descriptions.

7.2.6 Unannotated structures from the AlphaFold database

To enhance the structural diversity of our training set, we include a large corpus of unannotated proteins from the AlphaFold Database (AFDB). As these proteins lack descriptive text, we pair them with a randomly selected prompt from a predefined set of 50 non-informative sentences (*e.g.* "design a protein randomly"). This strategy of using a various set of generic prompts encourages the model to explore the vastness of the protein fold space independently of any specific functional constraints.

7.3 Training regimen

We implement the training using PyTorch. To optimize GPU memory usage, we employ activation checkpointing[75] and apply DeepSpeed ZeRO-2[76] for gradient partitioning and optimizer partitioning. All models are trained using BFloat16 mixed-precision to enhance efficiency.

The models were trained with the standard AdamW optimizer, using hyperparameters of $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon =$ 1e-8. We used a global batch size of 512 and applied a weight decay of 0.1 to prevent overfitting. The learning rate was warmed up over the first 10% of training steps, followed by a cosine decay schedule. Additional details on other hyperparameters are provided in Supplementary Table 3.

To ensure a balanced representation of protein space and mitigate distributional biases, our training regimen leverages the AFDB cluster[77] for sampling proteins from the InterPro and AFDB datasets. This is implemented through a hierarchical, two-stage sampling strategy: a cluster is first selected uniformly at random, and a protein is then sampled uniformly from within the chosen cluster. For the remaining datasets, proteins are sampled directly, without this cluster-based approach. Furthermore, each dataset is assigned a specific sampling weight, as detailed in Supplementary Table 2.

[2] https://www.ebi.ac.uk/interpro/download/InterPro/

7.3.1 SaProt-T/O training details and rationale

To generate amino acid sequences in the second stage of the Pinal pipeline, we develop two specialized models, SaProt-T and SaProt-O, based on the SaProt architecture.

SaProt-T(ext) is specifically designed for the task of protein reverse folding: predicting amino acid sequences given both protein backbone structures (as 3Di tokens) and textual descriptions.

SaProt-O(mni) is trained for maximum flexibility to better align with diverse, real-world protein (re)design workflows. It is trained on various combination of inputs, which may include or exclude textual prompts, structural information, or portions of the amino acid sequence (see Extended Data Fig. 5b for data combination details). This "omni-modal" training strategy enables SaProt-O to effectively handle a wide range of design scenarios, such as infilling, redesign, or sequence generation from structure and text alone.

To validate our core hypothesis—that conditioning on text is crucial for generating high-quality, functionally relevant sequences—we conducted an ablation study comparing our models against the original SaProt. While existing models like baseline SaProt[44] excel at predicting sequences from structural tokens alone, we posit that text provides essential fine-grained cues not present in the coarse-grained 3Di tokens. To test this hypothesis, we compared our text-conditioned models SaProt-T and SaProt-O (760M) with the original SaProt (650M).

The findings (Extended Data Fig. 5a) demonstrate that sequences generated by SaProt-T and SaProt-O consistently exhibit superior textual alignment and predicted foldability. This result underscores the significance of our text-conditioning strategy and justifies the development of SaProt-T/O for high-fidelity, language-guided sequence design.

In this paper, we use SaProt-T for all Pinal experiments. By contrast, SaProt-O is more suited for protein editing tasks where partial sequences or structures are provided. For example, after proteins are designed by Pinal, one can use SaProt-O to re-engineer the protein by mutating multiple amino acids. We leave the systematic exploration of SaProt-O applications as future work.

7.3.2 The stability issue of large-scale model training

Training the 15.5B T2struct model presented a challenge due to its size. We observed that both training and validation losses exhibited spikes and failed to converge after the learning rate warmup phase. Inspired by prior work[46,80] suggesting that learning rates should be adapted based on model parameters, we reduced the learning rate to 1e-5, which is 10 times smaller than the initial value. This adjustment effectively resolved the training stability issues, enabling the model to converge reliably and maintain a consistent loss trajectory throughout training.

7.3.3 Training phase for T2struct

The training protocol for the 15.5B T2struct model was executed in three sequential phases. The initial phase consisted of 320,000 training steps on the complete data corpus, utilizing an experimental sampling distribution specified in Supplementary Table 2. A key feature of this stage was the stratified sampling of the TrEMBL-ProTrek dataset, wherein pairs with a ProTrek score of 12 or higher were sampled with 80%

probability, versus 20% for pairs with scores below 12. This was followed by a 305,000-step training phase on a curated corpus. For this second phase, the TrEMBL-ProTrek data component was wholly substituted with a high-quality subset, which was derived by filtering for pairs with a ProTrek score exceeding 12, thereby reducing its volume from 412 million to approximately 60 million pairs. The sampling weights for all datasets held constant from the initial phase. While the two phases training have already established Pinal's superior performance over baselines like ESM3, Chroma, and ProteinDT, we identified a tendency in all models to generate repetitive amino acid sequences (e.g., "AAAAAAA")[81]. Our investigation revealed that the root cause was not the sequence design module (SaProt-T), but rather the generation of repetitive Foldseek structural tokens—a tendency correlated with training data derived from lower-quality predicted structures in the AFDB (Extended Data Fig. 4b).

Based on these findings, we introduce a third training phase to specifically address this repetition issue. This phase involves fine-tuning the model for an additional 115,000 steps on a further refined dataset, from which we exclude protein structures with a pLDDT score below 70. The effectiveness of this refined training strategy is confirmed by evaluations on the hard test set (Extended Data Fig. 4a). Specifically, this third phase significantly improves the model's performance on foldability metrics (pLDDT and PAE). Furthermore, as shown by a comparative analysis of the model before and after this phase (Phase 2 *vs.* Phase 3), training with this higher-quality structural data substantially decreases the percentage of designed proteins containing excessive amino acid repeats (Extended Data Fig. 4b), validating our approach.

The total training across all phases cost approximately 50,000 H800 Nvidia GPU hours.

7.4 Inference for programmable protein design

Protein generation with Pinal involves a two-stage sampling process. First, the T2struct model generates structural backbones from input text using multinomial sampling (($\tau = 0.8$) to produce diverse structures. Second, SaProt-T generates sequences using a flexible sampling strategy tailored to the specific goal.

For large-scale in silico evaluations, we aim to obtain the single most probable sequence that SaProt-T can design for a given structure. We employ argmax sampling (greedy decoding), where the model selects the amino acid with the highest predicted probability at each position. This deterministic approach yields a single, reproducible "best guess" sequence, which is ideal for systematic benchmarking.

For generating a large number of candidates for wet-lab experimental validation, exploring a broader range of viable sequences is crucial. We employ the multinomial sampling for the SaProt-T to explore a broader sequence space. This approach increases the diversity of the final generation, enhancing the probability of discovering novel proteins with high functional activity.

To determine the optimal number of candidates (K) (see section 1), we conducted an empirical study on the hard test set. We varied K from 5 to 100 and found that performance, measured by the ProTrek score and pLDDT of the top-ranked designs, improves notably as K increases to 50, after which the gains become marginal (Extended Data Fig. 3). We therefore set $K = 50$ as the standard for all experiments, offering a strong balance between performance and computational cost.

Inference is performed on NVIDIA H800 GPUs, which require at least 40GB of memory to run the 15.5B T2struct. On a single GPU, Pinal generates sequences at a rate of approximately 300 residues per second. The model's maximum generation length is 1022 residues, with inference time increasing with sequence length.

7.5 Benchmarking and evaluation protocol

7.5.1 Construction of evaluation sets

Our evaluation protocol assesses two distinct capabilities of Pinal: generalization from complex functional descriptions and translation of specific functional prompts into proteins.

Design protein from complex text descriptions (long sentences evaluation). To evaluate the generalization capability of Pinal on unseen text-protein pairs, we constructed two distinct evaluation sets from a held-out portion of the Swiss-Prot dataset. First, we established a standard test set comprising 3,304 proteins. This set was achieved by clustering the Swiss-Prot database at a 50% sequence identity threshold and selecting protein clusters that were disjoint from those used in training. For each protein in this set, we generated comprehensive descriptions by simply concatenating all available functional annotations. These descriptions were subsequently paraphrased using GPT-4 to ensure the resulting prompts were textually novel to our model. From this pool of unique prompts, we randomly sampled 500 for the long-sentence-based design evaluation. Second, we created a more stringent 'hard' test set containing 465 proteins. There proteins were derived by filtering the standard test set, retaining only proteins that share less than 30% sequence identity with any protein in the training clusters. We applied the prompt generation procedures used for both the standard test set and SwissProt-Aug, and randomly sampled another 500 prompts to constitute the final 'hard' evaluation set. For each description, we then generated five protein candidates, resulting in a total of 2,500 designed proteins per setting for our analysis.

Translating specific functional prompts (short sentences & keyword-based evaluation). To assess the model's ability to accurately translate functional concepts, we created two evaluation sets by sampling prompts directly from the entire datasets. The short sentence set was constructed to cover a broad range of functional commands while simultaneously testing the model's robustness to textual variation. To achieve this, we first systematically sampled 50 descriptions from each of the 17 distinct protein-level sub-sections within SwissProt-Annot. These 850 descriptions were then processed using the same paraphrasing protocol as the long sentences to ensure their phrasing was novel to the model. The keyword-based set was designed for a direct, head-to-head comparison with ESM3. This set was constructed by randomly sampling 500 entry names from the InterPro database to serve as textual inputs for both models, ensuring a fair benchmark of their keyword-conditioned generative capabilities. Similar to the long-sentence evaluation, we then generated five protein candidates for each prompt in these sets.

7.5.2 Prompt generation for structurally uncharacterized protein

The construction of these prompts follows a multi-step pipeline. First, we identify a pool of structurally uncharacterized proteins by comparing AlphaFold-predicted structures against the entire PDB database

using Foldseek. We retain 6,000 proteins whose predicted structures show no significant similarity (TM-score < 0.5) to any known experimental structure. For this set of novel proteins, we first use ProTrek to predict functional text for them, and then we use a large language model[73] to paraphrase this retrieved text to ensure linguistic diversity. These rewritten descriptions serve as the final input prompts for the novelty generation task.

7.5.3 Metrics

Evaluation metrics are grouped into three primary categories: textual alignment, foldability, and novelty.

Textual alignment. We use three metrics to assess alignment with the input prompt. For designs with a known native structure, we report the GT-TMscore, defined as the TM-score[3] between the design's predicted structure and the ground-truth structure. For keyword-based prompts, we additionally report keyword recovery, calculated with InterProScan[4] to measure functional consistency, strictly following the protocol of ESM3[46]. For a more general assessment, we use the ProTrek score[50], which measures the cosine similarity between the prompt's text embedding and the embeddings of the designed sequence (Seq. ProTrek) or structure (Struc. ProTrek). For both sequence and structure modalities, the ProTrek scores of native protein-text pairs show high Spearman correlation with scores against semantically equivalent, paraphrased descriptions, confirming the metric's robustness. Furthermore, scores for correctly matched pairs are significantly higher than for randomly shuffled, mismatched pairs, demonstrating strong discriminative ability (see Extended Data Fig. 6 for full analysis).

Foldability. The structural plausibility of each designed sequence is predicted with ESMFold[82] or AlphaFold3[83]. We report two standard confidence metrics: the predicted Local Distance Difference Test (pLDDT), a per-residue confidence score, and the Predicted Aligned Error (PAE), which measures the confidence in relative domain positions. Higher pLDDT and lower PAE values indicate a more confident, well-folded structure.

Novelty. The novelty of a design is determined by structurally comparing it against all entries in the Protein Data Bank (PDB). We report the highest query TM-score against any PDB entry as the PDB-TM.

7.6 Phylogenetic analysis of Pinal-generated sequence diversity

To evaluate the sequence diversity of Pinal-designed proteins, we retrieved natural homologous sequences for target protein families from the Swiss-Prot database. Concurrently, 5,000 sequences were generated per family using Pinal and filtered for high quality based on a ProTrek score > 15, pLDDT > 85, and PAE < 7. To reduce redundancy, both the natural and the filtered generated datasets were independently subjected to a two-step clustering process with CD-HIT, first at 90% identity (c = 0.9, aS = 0.8) and subsequently at 70% identity (c = 0.7, aS = 0.8). The final representative sequences from both sets were then combined, aligned using Clustal Omega, and used to construct a maximum likelihood phylogenetic tree with IQ-TREE 2 under an automatically selected best-fit model. The resulting tree was visualized with a midpoint root to assess the diversity of Pinal's designs by analyzing their distribution relative to their natural counterparts.

[3] https://zhanggroup.org/TM-score/
[4] https://www.ebi.ac.uk/interpro/download/InterProScan/

7.7 Materials, strains, and cultivation

Glycine, nicotinamide adenine dinucleotide (NAD^+, NADH), and tetrahydrofolic acid (THF) were obtained from Sigma-Aldrich (St. Louis, MO, USA). Dithiothreitol (DTT), pyridoxal 5'-phosphate monohydrate (PLP), Ethanol, Acetaldehyde, 2,4-Dinitrophenylhydrazine (DNPH), Acetonitrile (ACN), Trifluoroacetic acid (TFA), Imidazole, NaOH, HCl, KH_2PO_4, 1,1,1,3,3,3-Hexafluoro-2-propanol (HFIP), 4-nitrophenyl butyrate (pNPB), standards of bis-hydroxyethyl terephthalate (BHET), terephthalic acid (TPA) and mono(2-hydroxyethyl) terephthalate (MHET) were obtained from Aladdin (Shanghai, China). *E. coli* BL21(DE3) (Tsingke Biotechnology, Beijing, China) was used as the primary host for recombinant protein expression. Cultures were routinely grown in Luria-Bertani (LB) medium at 37 °C with shaking at 200–225 rpm.

7.8 Recombinant protein expression and purification

7.8.1 Expression and purification of pinalGFP

The codon-optimized gene sequence for pinalGFP, along with benchmark proteins, was synthesized and cloned into a pET28a(+) vector containing an N-terminal 6×His-tags (GENEWIZ, Suzhou, China). The construct was transformed into *E. coli* BL21(DE3) cells (Tsingke, Beijing, China). For large-scale protein production, transformed BL21(DE3) cells were cultured in 4 L of LB medium at 37 °C until the OD_{600} reached 0.6–0.8. Protein expression was induced with 0.5 mM IPTG, followed by incubation at 18 °C for 16–20 hours. Cells were harvested by centrifugation, resuspended in lysis buffer (50 mM Tris-HCl, 500 mM NaCl, 10 mM imidazole, pH 8.0), and lysed by sonication on ice. The cleared lysate was obtained by centrifugation (4,000 × g, 30 min, 4 °C). The supernatant was first purified using Ni-NTA agarose affinity chromatography (Qiagen). The bound protein was eluted with a buffer containing 250 mM imidazole. Following concentration using Amicon Ultra-15 centrifugal filters (30 kDa MWCO, Millipore), the protein was further purified by size exclusion chromatography on a Superdex 200 pg 16/600 column using an ÄKTA prime PLUS system (GE Healthcare), equilibrated with 25 mM Tris-HCl, 150 mM NaCl, pH 8.0.

7.8.2 Expression and purification of pinalADH and pinalPETase

Candidate gene sequences for pinalADH and pinalPETase, codon-optimized for *Escherichia coli*, were commercially synthesized (Guangzhou Aiji Biotechnology Co., Ltd.). The synthesized genes were cloned into the pET28a(+) expression vector containing an C-terminal 6×His-tags and transformed into *E. coli* BL21(DE3) competent cells. For protein expression, transformed cells were cultured in 100 mL of LB medium at 37 °C with shaking at 225 rpm until the OD_{600} reached 0.6. Expression was induced with a final concentration of 0.1 mM IPTG, and the cultures were subsequently incubated at 16 °C and 120 rpm for 24 hours. Cells were harvested by centrifugation (4,500 × g, 30 min), and the pellet was washed twice with 10 mM phosphate buffer. Cell disruption was performed using a high-pressure homogenizer (KS-2000M, Kaibai Nano Technology). The lysate was cleared by centrifugation at 12,000 × g for 30 min at 4 °C. The clarified supernatant was loaded onto a 5 mL HISTRAP FF CRUDE column (Cytiva) using an ÄKTA pure 25 system (Cytiva).

The column was washed with 50 mM phosphate buffer (pH 7.8) containing 60 mM imidazole and 500 mM NaCl. The protein was then eluted with 50 mM phosphate buffer (pH 7.8) containing 250 mM imidazole and 500 mM NaCl. Protein concentration was determined using a BCA Protein Assay Kit (Biogene), and purity was assessed by SDS-PAGE.

7.8.3 Expression and purification of pinalH

The selected H-protein genes, codon-optimized for *E. coli*, were synthesized with C-terminal 6×His-tags and cloned into the pET28a vector (Tsingke Biotechnology). The constructs were transformed into *E. coli* BL21(DE3) for expression. Transformants were cultured in LB medium with 100 mg L^{-1} ampicillin at 37 °C until an OD_{600} of 0.6–0.8 was reached. Expression was induced with 0.1 mM IPTG, and cultures were incubated at 16 °C for 16 hours. Cells were harvested by centrifugation (8,000 rpm, 10 min, 4 °C) and washed with 50 mM phosphate buffer (pH 7.0). For purification, cell pellets were resuspended in 50 mM Tris buffer (pH 7.0) and lysed by ultrasonication. The lysate was cleared by centrifugation (10,000 × g, 5 min, 4 °C). The supernatant was applied to a gravity-flow column packed with Ni^{2+}-charged resin (High Affinity Ni-charged Resin FF, GenScript) pre-equilibrated with lysis buffer (50 mM Tris, 10 mM imidazole, 300 mM NaCl, pH 7.5). The column was washed with a wash buffer (50 mM Tris, 30 mM imidazole, 300 mM NaCl), and the purified H-proteins were eluted using an elution buffer (50 mM Tris, 300 mM imidazole, 300 mM NaCl, pH 7.5). Excessive imidazole was removed by ultracentrifugation. Purity was confirmed by SDS-PAGE.

7.9 Characterization of pinalGFP

7.9.1 Computational screening and design refinement

The initial 1,000 sequences from Pinal were processed through a multi-tiered computational pipeline. Quality filtering was based on: a ProTrek folding propensity score > 15; and AlphaFold3-predicted structural quality (pLDDT > 80, PAE < 7). The sequences were further filtered for length (215–245 amino acids), predicted solubility (Protein-Sol > 0.7), and thermodynamic stability (PyRosetta ref2015 energy score < -50 REU), yielding 80 high-confidence candidates. To ensure diversity, these candidates were clustered into 20 groups using K-means on PCA-reduced features (BLOSUM62 similarity, molecular weight, hydrophobicity). A representative sequence was selected from each cluster based on a weighted score of solubility (0.6) and stability (0.4). For targeted engineering, the 20 representatives were structurally aligned with avGFP (PDB: 2WUR) using TM-align to identify positions corresponding to the Ser65-Tyr66-Gly67 chromophore. The TYG motif was manually introduced using PyMOL, and variants were re-assessed for stability and structural integrity, resulting in 15 candidates for experimental validation(Supplementary Table 8.).

7.9.2 Fluorescence characterization

Whole-cell fluorescence was screened in 24-well plates. After 20 hours of induction at 30 °C, cells were pelleted, resuspended in PBS (pH 7.4), and transferred to 96-well plates for measurement on an Agilent

Microplate Reader (excitation 460 nm, emission 510 nm). For purified proteins, excitation and emission spectra were recorded from 300–520 nm and 460–650 nm, respectively. Quantum yields (QY) were determined relative to the avGFP standard (QY = 0.79) using the equation:

$$QY_i = \frac{\int_{460\text{nm}}^{650\text{nm}} (F_i)\, d\lambda \cdot A_{280\text{avGFP}}}{A_{280\text{avGFP}} \cdot \int_{460\text{nm}}^{650\text{nm}} (F_{\text{avGFP}})\, d\lambda} \cdot QY_{\text{avGFP}} \tag{3}$$

7.10 Characterization of pinalPETase

7.10.1 Computational screening and design refinement

Candidate sequences were first filtered for the presence of the conserved SMGGG motif. Putative hydrolases were further selected based on the canonical spacing of the catalytic triad (Ser-Asp-His), with Asp located at positions 43–49 and His at 71–81 relative to the SMGGG motif. Structures were predicted using AlphaFold3, and the geometry of the catalytic triad was validated by measuring Cα distances (S–D: 10.5 ± 0.2 Å, D–H: 4.7 ± 0.3 Å, S–H: 8.0 ± 0.2 Å). After excluding sequences with predicted signal peptides (SignalP 6.0), semi-flexible molecular docking was performed using AutoDock Vina between the candidates and the PET tetramer analogue 2-HE(MHET)$_4$. The pose with the lowest binding energy was selected for final candidate consideration.

7.10.2 Preparation of PET film

Amorphous PET films (Treat-PET) were prepared by dissolving commercial GfPET in 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP) at 60 mg/mL. 4 mL of this solution was cast onto a glass Petri dish, and after overnight solvent evaporation, the film was incubated in 75% ethanol for 2 h before being peeled and cut for assays.

7.10.3 Thermal shift assay by nanoDSF (PR Panta)

Protein thermal stability was measured on a PR Panta instrument (NanoTemper Technologies) using nano-differential scanning fluorimetry (nanoDSF) in chemically inert glass capillaries (NanoTemper, Cat. #K003). Each sample (10 µL, 5–250 $\mu\text{g}\,\text{mL}^{-1}$ in the final formulation buffer) was loaded by capillary action into the standard chip. The temperature was ramped from 25 °C to 95 °C at a constant rate of 5 $°\text{C}\,\text{min}^{-1}$, while intrinsic fluorescence was recorded at 330 nm and 350 nm every 0.2 °C ($\geq$20 data points min^{-1}).

Melting temperature (T_m) was derived from the first derivative of the ratio $F_{350\,\text{nm}}/F_{330\,\text{nm}}$ versus temperature.

7.10.4 Esterase activity and plastic hydrolysis activity test

Esterase activity was measured using p-nitrophenyl butyrate (pNPB). Reactions containing 80 μL of 100 mM KH_2PO_4-NaOH buffer (pH 8.0), 10 μL of 2 μM enzyme, and 10 μL of 10 mM pNPB were monitored for p-nitrophenol (pNP) formation by measuring absorbance at 405 nm over 5 minutes on a microplate reader. Plastic hydrolysis activity was assessed by incubating 1 mg of Treat-PET film with 1 μM enzyme in 600 μL

of 100 mM KH_2PO_4-NaOH buffer (pH 8.0) at 40 °C. Soluble degradation products (MHET and TPA) were quantified by HPLC.

7.11 Characterization of pinalADH

7.11.1 Derivatization and HPLC analysis of acetaldehyde

To quantify acetaldehyde formation from ethanol, we employed a 2,4-dinitrophenylhydrazine (DNPH) derivatization strategy. Briefly, reaction mixtures were quenched with 5 mL of DNPH solution (0.2 mg mL^{-1} in 85% phosphoric acid/acetonitrile) to ensure > 100-fold molar excess of reagent. Samples were incubated at 60 °C for 20 min to allow complete formation of the corresponding hydrazone derivative. Following centrifugation and filtration (0.22 µm), 5 µL aliquots were injected onto an HPLC system (C18 column, 4.6 × 150 mm, 2.7 µm; 37 °C). Separation was achieved using a water–acetonitrile gradient at 0.8 mL min^{-1}, and detection was performed at 360 nm.

7.11.2 PinalADH activity test

The activity of designed ADHs was determined by quantifying the production of acetaldehyde from ethanol. The 100 μL reaction mixture contained 100 mM ethanol, 500 μM NAD^+, and 200 nM purified enzyme in Tris-HCl buffer (pH 8.0). Reactions were incubated at 37 °C for 120 minutes. The resulting acetaldehyde was then derivatized and analyzed by HPLC, with a characteristic retention time of 7.292 minutes.

7.12 Characterization of pinalH

7.12.1 Reconstituted Glycine Cleavage System assay

The function of designed H-proteins was assessed in a reconstituted GCS. The reaction mixture (50 mM Tris-HCl, pH 7.5) contained 200 mM NH_4HCO_3, 60 mM DTT, 1 mM THF, 30 mM formaldehyde, 5 μM P-protein, 3 μM T-protein, 1 μM LplA, 0.5 mM lipoic acid, 1 mM ATP, 1 mM $MgCl_2$, and was initiated by adding 40 μM of the respective designed H-protein. The reaction was carried out at 42 °C for 12 hours, with samples taken periodically for analysis.

7.12.2 Quantification of glycine

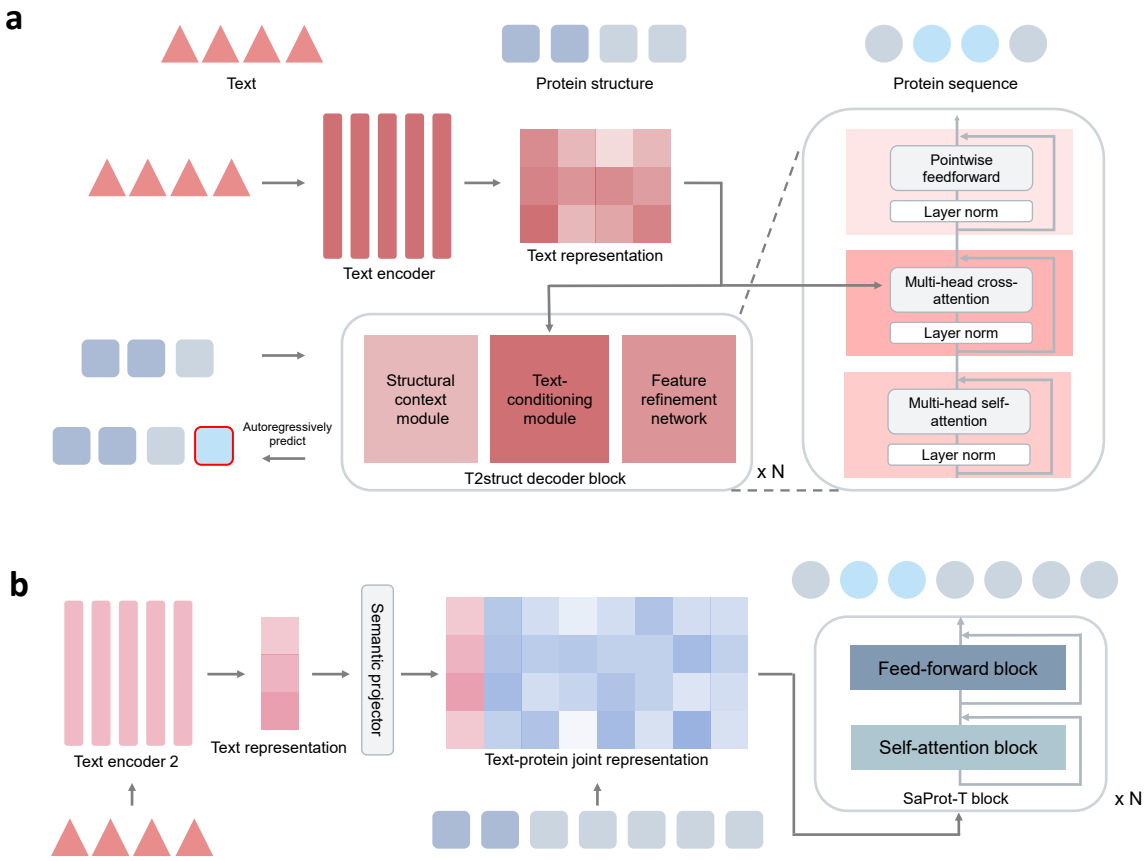
Glycine concentration was determined by HPLC following pre-column derivatization with dansyl chloride. 40 μL of the reaction mixture was mixed with $NaHCO_3$ and dansyl chloride solution, incubated at 30 °C for 30 min, and then acidified. The dansyl-glycine derivative was quantified by HPLC on a Shim-pack GIST C18 column, with detection at 254 nm. Results were cross-verified using an alternative OPA derivatization method on an Agilent InfinityLab Poroshell HPH-C18 column.

8 Data availability

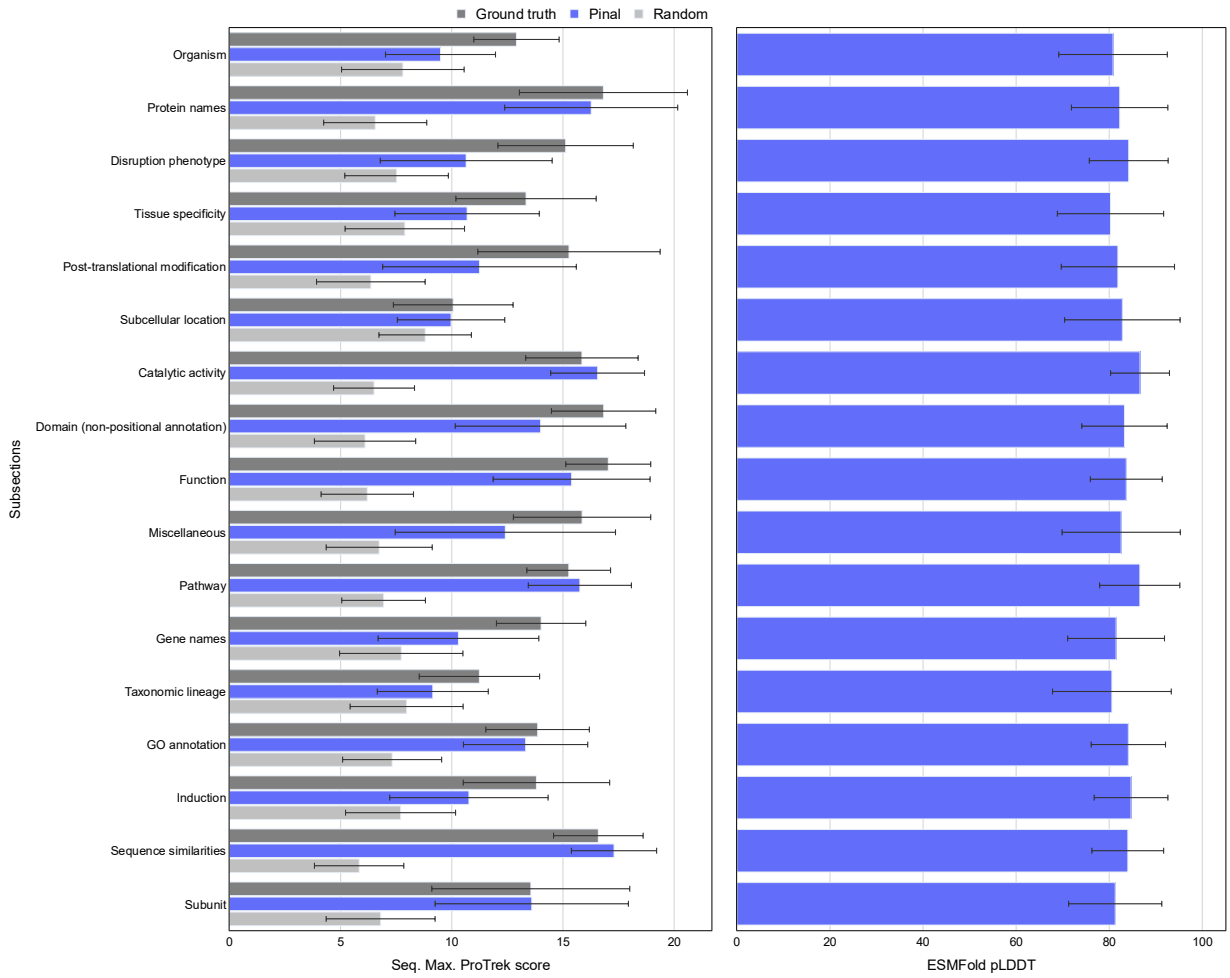
The pre-trained Pinal weights can be downloaded from https://huggingface.co/westlake-repl/Pinal.

9 Code availability

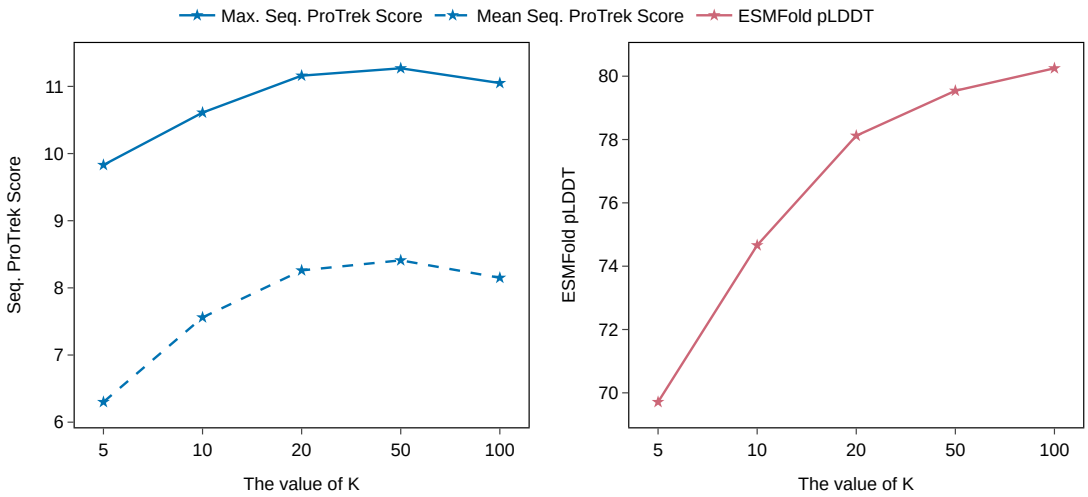
Pinal is open-sourced under the MIT license. The code repository is available at https://github.com/westlake-repl/Denovo-Pinal. The Pinal web server is located at http://www.denovo-pinal.com/. The SaProt-T/O (the second component of Pinal) web server is located at http://113.45.254.183:9527/.



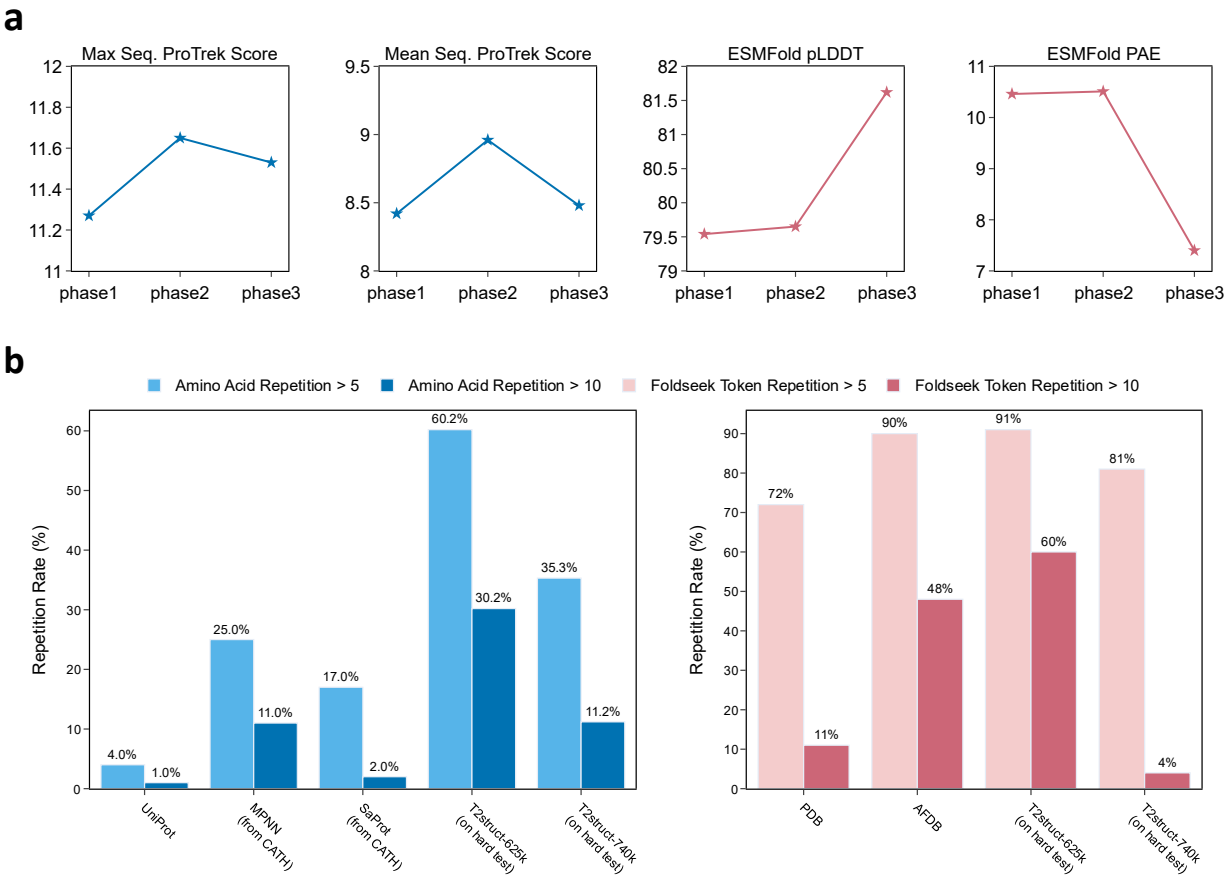
Extended Data Fig. 1: **Detailed architectures of the Pinal generative modules. a**: the T2struct module, which translates natural language descriptions into structural token sequences. A text encoder first generates a context-rich text representation. This representation is then used to guide the T2struct decoder block, which autoregressively predicts the sequence of 3Di structural tokens. The detailed view (right) shows the internal composition of each decoder block, which consists of three main sub-modules: a structural context module to process the generated structural sequence, a text-conditioning module to integrate guidance from the text representation, and a feature refinement network for further processing. This entire block is repeated N times. **b**: the SaProt-T module, which designs an amino acid sequence based on a structural backbone and a text prompt. A semantic projector maps the text representation into the same embedding space as the structural tokens. This projected text vector is prepended to the structural token embeddings to form a text-protein joint representation. This joint representation is then processed by the SaProt-T block, repeated N times, to autoregressively predict the final amino acid sequence.



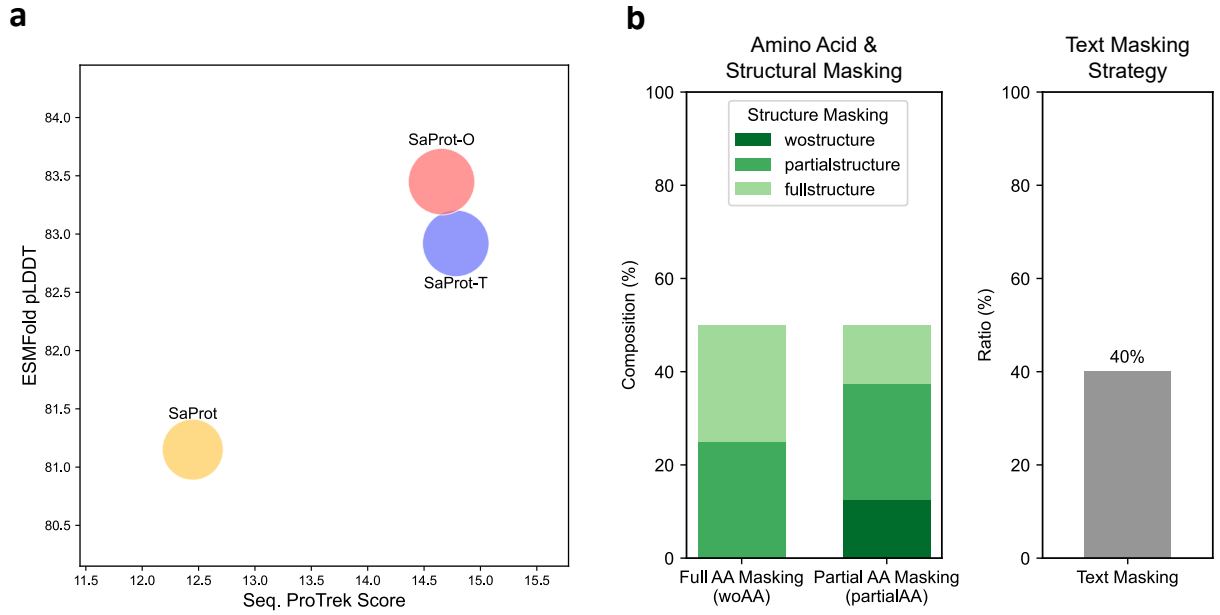
Extended Data Fig. 2: **Detailed performance analysis across diverse protein function descriptions.** The figure evaluates Pinal's design performance across 17 distinct functional categories curated from Swiss-Prot. Left panel: Analysis of textual alignment, measured by the maximum sequence ProTrek score. The performance of Pinal's top-ranked design (blue) is compared against a random baseline (light grey) and the ground-truth protein (dark grey) for each category. Pinal consistently and significantly outperforms the random baseline across all categories and approaches ground-truth performance, demonstrating robust language understanding. Right panel: Predicted structural plausibility of the corresponding Pinal-designed proteins, measured by ESMFold pLDDT. The results indicate that Pinal generates proteins with high predicted foldability (mean pLDDT is typically > 80) across all functional description types, from concrete categories like "Protein names" to more abstract ones like "Disruption phenotype".



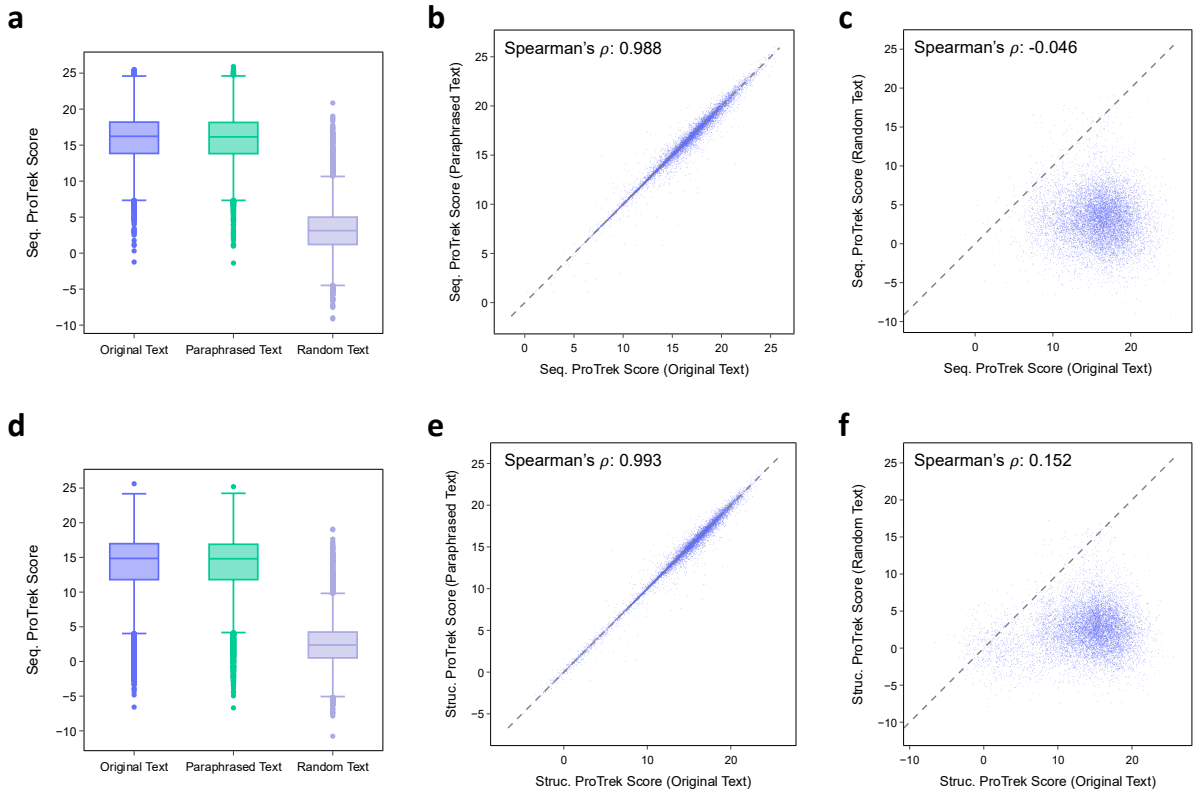
Extended Data Fig. 3: **Determining the optimal number of sampling candidates (K).** Performance of the top-ranked Pinal designs on the "hard" test set as a function of the number of sampled candidates (K). For each value of K, we generate K candidates, rank them by their joint log-probability, and then evaluate the top-ranked design. Left panel: textual alignment, measured by Max. and Mean Seq. ProTrek score. Right panel: predicted structural plausibility, measured by ESMFold pLDDT. Both textual alignment and foldability improve significantly as K increases from 5 to 50. Further increasing K to 100 yields only marginal gains (or a slight decrease in ProTrek score) while substantially increasing computational cost. Based on this analysis, K=50 is used as the standard setting for all experiments, offering a robust balance between design performance and inference efficiency.



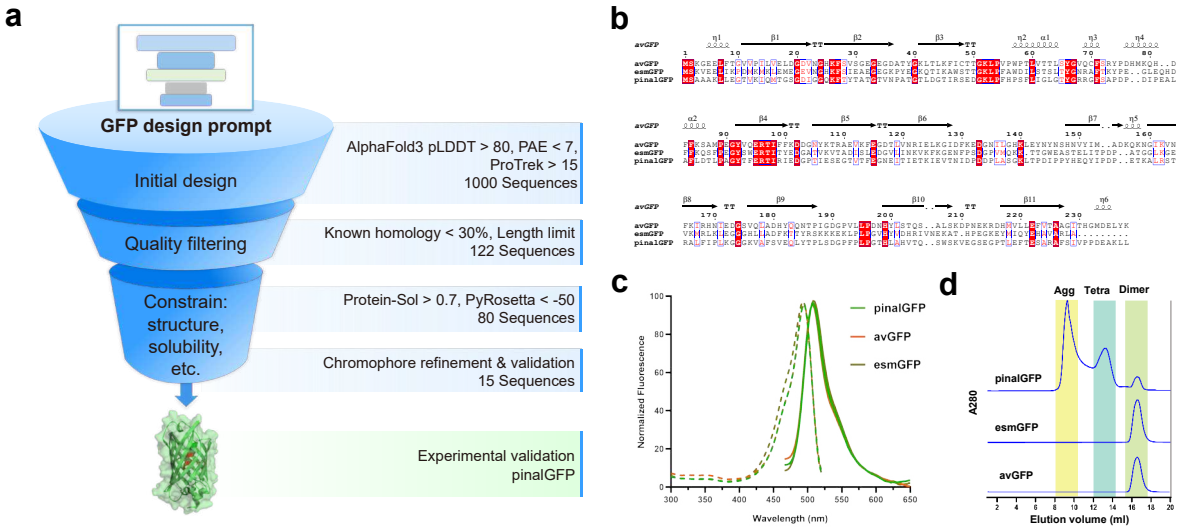
Extended Data Fig. 4: **The three-phase training improves model performance and mitigates sequence repetition. a**, Performance of the 15.5B T2struct model on the "hard" test set across three distinct training phases. Phase 1 (320k steps) represents initial training; Phase 2 (625k steps) incorporates data filtering by ProTrek score; Phase 3 (740k steps) is a final fine-tuning step on higher-quality data filtered to exclude low-pLDDT scores and high token repetition. The plots demonstrate that the third phase significantly improves protein foldability, while incurring only a minor, acceptable trade-off in textual alignment (ProTrek score). **b**, Analysis of the sequence repetition phenomenon that motivated the third training phase. The left panel shows that the model after Phase 2 exhibited a high rate of amino acid repetition (>5) (e.g.,"AAAAA") at 60.2%, far exceeding that of natural proteins (UniProt, 4.0%). The final model after Phase 3 substantially reduces this rate to 35.3%. The right panel illustrates the effectiveness of the Phase 3 training data-filtering strategy. The high rate of Foldseek token repetition in the Phase 2 model's outputs (91%) correlates with the high rate found in the AFDB training data (90%). The refined training strategy after Phase 3 successfully reduces the token repetition rate to just 4%, thereby substantially resolving the Foldseek structure token repetition issue.



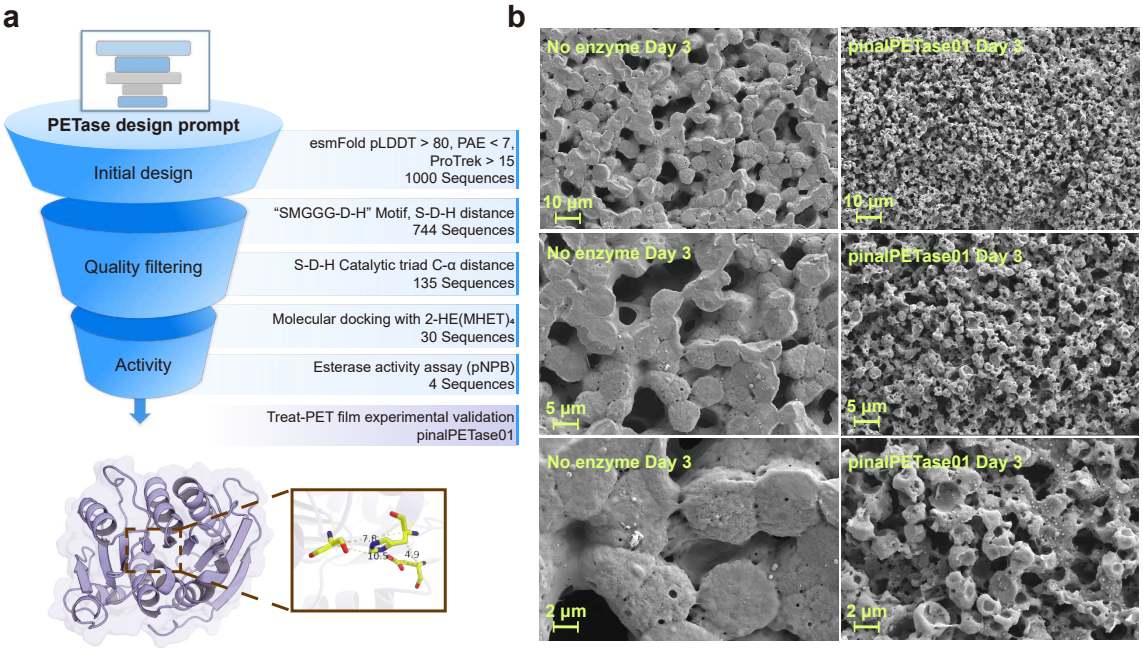
Extended Data Fig. 5: **Text conditioning is crucial for the sequence design module. a**: Ablation study comparing the performance of the original SaProt (650M parameters) with our text-conditioned models, SaProt-T and SaProt-O (760M parameters). Performance is evaluated on both textual alignment and predicted foldability. The results demonstrate that incorporating text conditioning (SaProt-T and SaProt-O) significantly improves both alignment and foldability compared to the baseline SaProt model, which is conditioned on structure alone. The size of each bubble corresponds to the number of model parameters. **b**: Composition of the omni-modal training strategy for the SaProt-O model. This strategy involves a varied combination of masking applied to the amino acid (AA) sequence, the structural tokens, and the text prompt to train a flexible, multi-purpose design model. The left panel details the distribution of structural token masking (no, partial, or full masking) under two different AA masking conditions. The right panel shows that the text prompt is masked in 40% of the training examples.



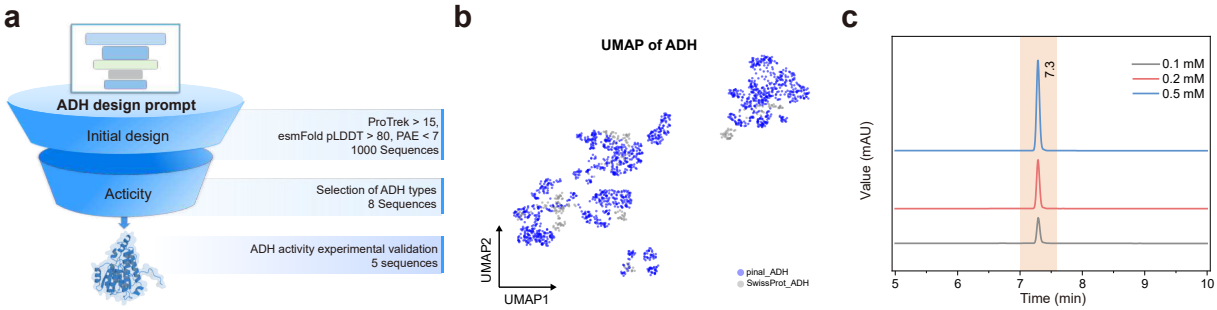
Extended Data Fig. 6: ProTrek score is a robust and discriminative metric for both text-sequence and text-structure alignment. The figure validates the ProTrek score's performance using a test set of 1,000 protein-text pairs from Swiss-Prot. The top row (a-c) evaluates the text-to-sequence score (Seq. ProTrek), while the bottom row (d-f) evaluates the text-to-structure score (Struc. ProTrek). (a, d) Box plots show the distribution of ProTrek scores for proteins paired with their original descriptions, LLM-paraphrased descriptions, and randomly assigned (mismatched) descriptions. For both sequence and structure, the scores for original and paraphrased text are high and statistically similar, while scores for random text are significantly lower, demonstrating the metric's discriminative power. (b, e) Scatter plots comparing the scores of original versus paraphrased descriptions for each protein. The strong linear relationship, confirmed by a Spearman's rank correlation coefficient (ρ) of >0.98 for both sequence and structure, demonstrates that the metric is robust to variations in wording and captures core semantic meaning. (c, f) Scatter plots comparing the scores of original versus randomly assigned descriptions. The lack of any significant correlation ($\rho \approx 0$) further confirms that the ProTrek score is a valid metric that reliably distinguishes between relevant and irrelevant protein-text pairings. The original ProTrek study also demonstrated that its score selectively responds to biologically significant keywords while ignoring non-functional terms. Building on this capability, a recent benchmark paper[56] adopted ProTrek as a key metric for evaluating the alignment between protein sequences and their textual descriptions.



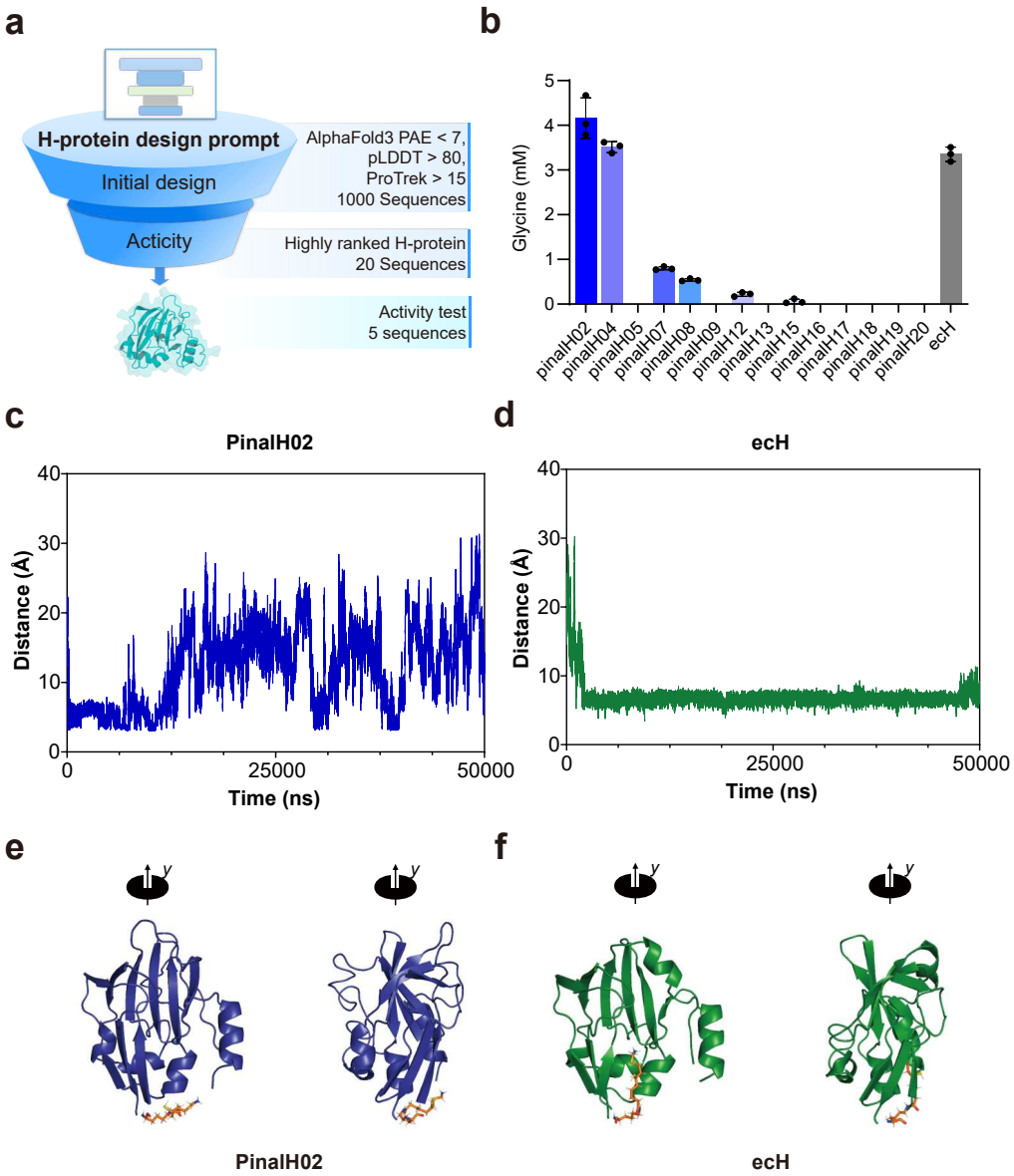
Extended Data Fig. 7: **Computational design, screening, and characterization of pinalGFP.** **a,** Flowchart illustrating the multi-tiered computational filtering strategy applied to initial Pinal-generated sequences. The numbers indicate the count of candidate sequences remaining after each screening step. This pipeline culminated in the selection of 15 structurally diverse candidates for experimental validation. **b,** Multiple sequence alignment of pinalGFP, esmGFP, and avGFP. The alignment highlights the significant sequence divergence of pinalGFP from both the natural and the other synthetic variant. **c,** Size exclusion chromatography profiles of the purified GFP variants on a Superdex 200 column. The elution volumes indicate their oligomeric states in solution: avGFP and esmGFP exist predominantly as dimers, whereas pinalGFP displays a more complex profile with fluorescence activity associated with the dimeric fraction. **d,** Normalized fluorescence excitation (dashed lines, λ_{em} = 510 nm) and emission (solid lines, λ_{ex} = 460 nm) spectra. The similar spectral profiles among the functional variants confirm successful chromophore maturation despite their sequence and oligomeric differences.



Extended Data Fig. 8: Computational design and microscopic validation of Pinal-designed PETases. a, Flowchart of the multi-tiered computational pipeline for PETase design and selection. An initial pool of 1,000 sequences was generated from a text prompt and progressively filtered based on the presence of key motifs ("SMGGG-D-H"), correct catalytic triad geometry, and molecular docking scores with a PET analogue (2-HE(MHET)$_4$). This process yielded four candidates for experimental testing, with 'pinalPETase01' selected as the top variant. The inset shows the predicted structure of 'pinalPETase01' with the catalytic triad residues highlighted. **b,** Scanning electron microscopy (SEM) images revealing the surface morphology of amorphous PET films. The control film incubated without enzyme for 3 days (left column) shows a relatively smooth surface. In contrast, the film treated with 'pinalPETase01' for 3 days (right column) exhibits extensive surface erosion, pitting, and degradation, confirming the enzyme's catalytic activity on the polymer substrate. Images are shown at increasing magnifications (top to bottom). Samples were sputter-coated with gold and imaged using a Zeiss Sigma 300 SEM.



Extended Data Fig. 9: **Computational design pipeline and analytical method for Pinal-designed ADHs. a,** Flowchart of the computational screening and selection process for ADH candidates. An initial pool of 1,000 sequences generated from the text prompt was filtered based on textual alignment (ProTrek score > 15) and structural confidence (pLDDT > 80, PAE < 7). From this, eight diverse candidates representing different ADH subfamilies were selected, leading to five successfully validated enzymes. **b,** UMAP of alcohol dehydrogenase (ADH) sequences. A total of 1,638 Pinal-designed ADH sequences (blue) that met the screening criteria were combined with natural ADH sequences from Swiss-Prot (gray, EC 1.1.1.1). The sequences were aligned using BLASTp, and the E-values were extracted to generate a similarity matrix. UMAP was used to reduce the dimensionality of the data, demonstrating that the Pinal-designed sequences effectively cover a large portion of the natural ADH family's sequence space, indicating their functional relevance. **c,** Representative HPLC chromatograms for the standard curve of derivatized acetaldehyde, the product of the ADH-catalyzed reaction. The standard was run at concentrations of 0.1 mM, 0.2 mM, and 0.5 mM, showing a consistent retention time of 7.3 minutes, which was used for product quantification.



Extended Data Fig. 10: **Computational design, screening, and mechanistic analysis of Pinal-designed H-proteins. a,** Flowchart illustrating the computational design and screening pipeline for H-protein candidates. An initial pool of 1,000 sequences generated from the text prompt was filtered using structural (pLDDT, PAE) and textual alignment (ProTrek score) metrics, resulting in 20 high-confidence candidates for experimental validation. **b,** Initial activity screening of 14 soluble Pinal-designed H-proteins compared to the native *E. coli* H-protein ('ecH'). Glycine synthesis was measured after a 2-hour reaction in a reconstituted Glycine Cleavage System (GCS). Data are shown as mean ± s.d. (n=3). **c-f,** Molecular dynamics (MD) simulations comparing the conformational dynamics of the lipoamide arm in the top-performing 'pinalH02' and the native 'ecH'. **c,** Trajectory of the distance between the Cδ atom of Glu27 in 'pinalH02' and the nitrogen atom of the lipoamide arm over 500 ns. The plot shows that the arm frequently samples extended conformations far from the protein core. **d,** A representative snapshot of the 'pinalH02' conformation at 200 ns, illustrating the solvent-exposed lipoamide arm. **e,** Corresponding distance trajectory for 'ecH', plotting the distance between the Cδ atom of Glu12 and the arm's nitrogen atom. The arm remains stably sequestered close to the protein body. **f,** A representative snapshot of 'ecH' at 200 ns, showing the aminomethyl lipoate arm locked within the protein's hydrophobic cavity.

Methods References

[67] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**, 1–67 (2020).

[68] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

[69] Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).

[70] Ahmad, S. *et al.* The uniprot website api: facilitating programmatic access to protein knowledge. *Nucleic Acids Research* gkaf394 (2025).

[71] Su, J., Zhou, X., Zhang, X. & Yuan, F. Protrek: Navigating the protein universe through tri-modal contrastive learning. *Nature Biotechnology (accepted for publication* 2024–05 (2024).

[72] Bairoch, A. The enzyme database in 2000. *Nucleic acids research* **28**, 304–305 (2000).

[73] GLM, T. *et al.* Chatglm: A family of large language models from glm-130b to glm-4 all tools (2024). URL https://arxiv.org/abs/2406.12793. 2406.12793.

[74] Zhou, X. *et al.* Decoding the molecular language of proteins with evola. *bioRxiv* (2025).

[75] Chen, T., Xu, B., Zhang, C. & Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).

[76] Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. *Zero: Memory optimizations toward training trillion parameter models*, 1–16 (IEEE, 2020).

[77] Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

[78] Su, J. *et al.* *Saprot: Protein language modeling with structure-aware vocabulary* (2024).

[79] Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* eads0018 (2025).

[80] Yang, G. *et al.* Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466* (2022).

[81] Verkuil, R. *et al.* Language models generalize beyond natural proteins. *BioRxiv* 2022–12 (2022).

[82] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[83] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

Supplementary Information for Toward *De Novo* Protein Design from Natural Language

Fengyuan Dai[1†], Shiyang You[2†], Yudian Zhu[1], Yuan Gao[4], Lihao Fu[4], Xibin Zhou[1], Jin Su[1], Chentong Wang[5], Yuliang Fan[1], Xiaoxiao Ma[2], Xianjun Deng[2], Letong Yu[2], Hui Qian[1], Yan He[5], Yitao Ke[5], Chenchen Han[1], Xing Chang[5], Liangzhen Zheng[6], Sheng Wang[6], Yajie Wang[1], Anping Zeng[1], Shunzhi Wang[3], Tong Si[4*], Jianming Liu[1*], Hongyuan Lu[2*], Fajie Yuan[1*]

[1]School of Engineering, Westlake University, Hangzhou, 310024, China.
[2]Bioscience and Biomedical Engineering Thrust, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511400, China.
[3]Department of Biochemistry, University of Washington Seattle, WA 98195, USA.
[4]State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences Shenzhen, 518055, China.
[5]School of Life Sciences, Westlake University, Hangzhou, 310024, China.
[6]Shanghai Zelixir Biotech Co., Ltd., Shanghai, 201203, China.

*Corresponding author(s). E-mail(s): tong.si@siat.ac.cn; liujianming@westlake.edu.cn; hongylu@hkust-gz.edu.cn; yuanfajie@westlake.edu.cn;
†These authors contributed equally to this work.

Supplementary Materials Guide:

Supplementary Method: Detailed Bioinformatics and Experimental Protocols

Supplementary Fig. 1

Supplementary Table 1-11

1 Detailed Bioinformatics and Experimental Protocols

1.1 GFP: Computational Screening and Targeted Engineering

An initial pool of 1,000 GFP candidate sequences generated by Pinal was subjected to a rigorous multi-tiered computational screening pipeline by applying stringent quality filters, including a ProTrek folding propensity score above 15, and high structural confidence from AlphaFold3 (average pLDDT > 80 and PAE < 7). This set was further refined to 80 high-confidence candidates based on optimal physicochemical properties, such as sequence length (215–245 amino acids), predicted solubility (Protein-Sol > 0.7), and thermodynamic stability (PyRosetta ref2015 energy score $<$ -50 REU). To ensure structural diversity, we performed K-means clustering on PCA-reduced biophysical features to select 20 representative designs. For targeted functionalization, these representatives were structurally aligned to avGFP (PDB: 2WUR), and the essential Ser65-Tyr66-Gly67 (TYG) motif was manually engineered into the homologous positions using PyMOL. Following a final assessment of structural integrity (MolProbity clashscore < 21) and stability, 15 distinct candidates were ultimately selected for experimental validation.

1.2 PETase: Bioinformatics Screening Workflow

Candidate PETase sequences were first filtered for the presence of the conserved "SMGGG" motif. Putative hydrolases were further selected based on the canonical spacing of the catalytic triad (Ser-Asp-His), with Asp at positions 43–49 and His at 71–81 relative to the motif. Structures were predicted with AlphaFold3 and validated for correct catalytic triad Cα distances. After removing sequences with signal peptides (SignalP 6.0), semi-flexible molecular docking was performed using AutoDock Vina between the candidates and a PET tetramer analogue, 2-HE(MHET)$_4$. The poses with the lowest binding energy were selected as final candidates.

1.3 H-protein: Reconstituted Glycine Cleavage System (GCS) Assay

The activity of designed H-proteins was measured in a reconstituted GCS. The reaction mixture (50 mM Tris-HCl, pH 7.5) contained 200 mM NH_4HCO_3, 60 mM DTT, 1 mM THF, 30 mM formaldehyde, 25 μM P-protein, 3 μM T-protein, 1 μM LplA, 0.5 mM lipoic acid, 1 mM ATP, and 1 mM $MgCl_2$. The reaction was initiated by adding 40 μM of the respective H-protein and incubated at 42°C for 12 h.

1.4 Glycine Quantification by HPLC

Glycine concentration was determined by HPLC after pre-column derivatization with dansyl chloride. A 40 μL aliquot of the reaction mixture was mixed with 160 μL of 0.2 M $NaHCO_3$ and 200 μL of dansyl chloride (5.4 mg·mL^{-1} in acetonitrile) and incubated at 30°C for 30 min. The reaction was quenched and acidified with 600 μL of 0.12 M HCl. The dansyl-glycine derivative was separated on a Shim-pack GIST C18 column and detected at 254 nm.

Software and Web Servers used in this article

Software	Web Servers
Pinal	http://www.denovo-pinal.com/
SaProt-T/O (Pinal)	http://113.45.254.183:9527/
AlphaFold3	https://alphafoldserver.com/
Foldseek	https://github.com/steineggerlab/foldseek
MMseqs2	https://github.com/soedinglab/MMseqs2
ProTrek	https://github.com/westlake-repl/ProTrek
Protein-Sol	https://www.novopro.cn/tools/prot-sol.html
PyRosetta	https://www.pyrosetta.org/
BLAST+	https://blast.ncbi.nlm.nih.gov/Blast.cgi
TM-align	https://zhanggroup.org/TM-align/
PyMOL (v3.1.1)	https://pymol.org/
MolProbity	https://molprobity.biochem.duke.edu/
ChimeraX (v1.10)	https://www.cgl.ucsf.edu/chimerax/
IQtree2	https://github.com/iqtree/iqtree2/
CD-HIT (v4.8.1)	https://github.com/weizhongli/cdhit/
Clustal Omega (v1.2.4)	http://www.clustal.org/omega/
iTOL (v7)	https://itol.embl.de/
Autodock Vina (v2.0)	https://github.com/ccsb-scripps/AutoDock-Vina
BioRender	https://www.biorender.com/

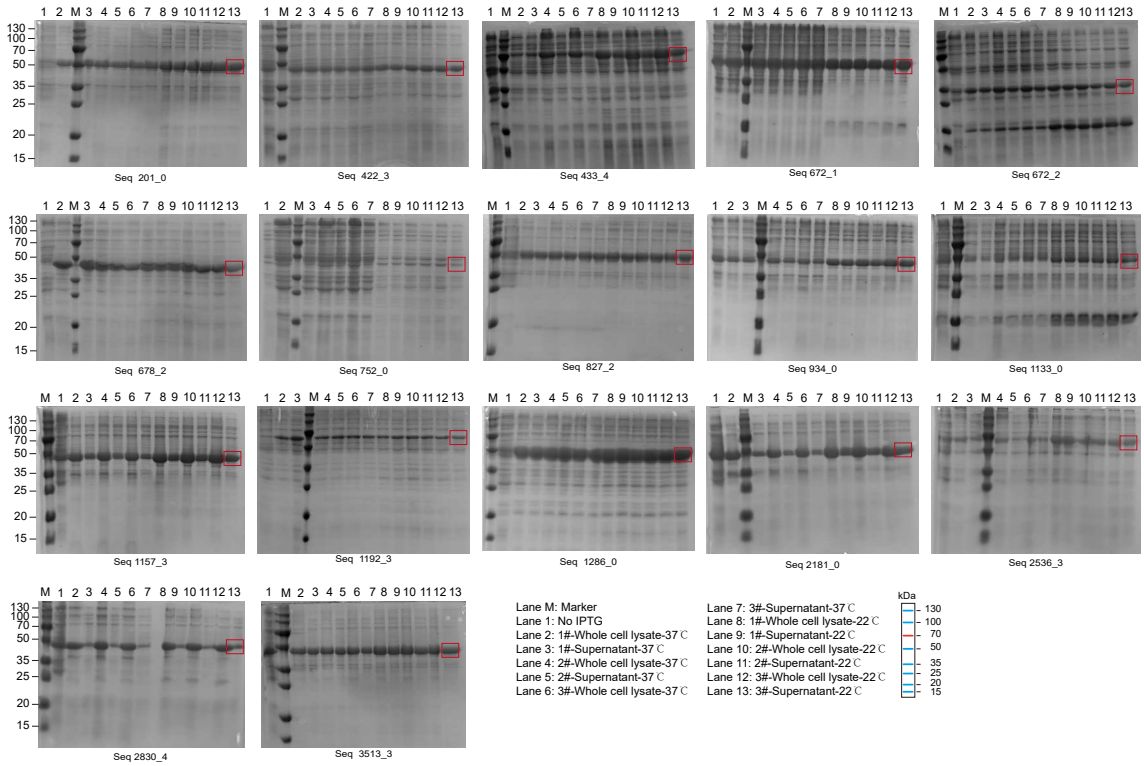


Fig. 1: **Experimental validation of the soluble expression of novel Pinal-designed proteins.** SDS-PAGE analysis of the 17 successfully expressed novel proteins from a set of 20 computationally selected candidates. These proteins were designed by Pinal to have low sequence (<30%) and structural (<0.5 TM-score) similarity to any known proteins, demonstrating a high soluble expression rate of 85% for truly novel designs. For each designed protein (identified by its sequence ID), three independent colonies were tested for expression in *E. coli* BL21(DE3). Protein expression was induced with IPTG and conducted at two different temperatures (37°C and 22°C). The lanes for each gel are as follows: M, molecular weight marker (kDa); 1, uninduced whole-cell lysate; 2-7, whole-cell lysate and corresponding supernatant for three biological replicates induced at 37°C; 8-13, whole-cell lysate and corresponding supernatant for the same three replicates induced at 22°C. The red boxes highlight the target protein band in the supernatant fraction, confirming successful soluble expression. Recombinant proteins with N-terminal SUMO and Strep-tag II were expressed in *E. coli* BL21(DE3). Plasmids (5 ng) were transformed into 50 µL competent cells, heat-shocked at 42 °C for 60 s, recovered in LB, and plated on LB agar containing 50 µg mL^{-1} kanamycin. Single colonies were grown overnight in 4 mL YP medium with kanamycin. Cultures were diluted 1:50 into fresh medium, grown to $OD_{600} \approx 0.6$, and induced with 0.1 mM IPTG at 22 °C or 37 °C for 20 h. Cells were harvested, lysed in PBS, and analyzed by SDS–PAGE.

Table 1: A list of the 46 key protein classes.

DNA/RNA polymerases	Intermediate filaments	GPCRs
Restriction enzymes	Growth factors	Receptor tyrosine kinases
Reverse transcriptases	Cytokines	Cell adhesion molecules
DNA repair enzymes	Chemokines	Heat shock proteins
Topoisomerases	Hormones	Molecular chaperones
Nucleases	Neurotransmitter receptors	Protein disulfide isomerases
antibody	G-proteins	Peptidyl-prolyl isomerases
Fluorescent Proteins	Second messengers	Luciferases
Actin	Transcription factors	β-galactosidase
Tubulin	Histones	Alkaline phosphatase
Collagen	RNA-binding proteins	Horseradish peroxidase
Keratin	DNA-binding proteins	Globular proteins
Fibronectin	Chromatin remodeling factors	Fibrous proteins
Laminin	Nuclear receptors	Motor proteins
Myosin	Ion channels	
Spectrin	Transporters	

Table 2: **Dataset statistics.** We have excluded proteins whose 3Di structures are not in Alphafold/UniProt database. Each dataset is assigned a specific sample weight for training.

Dataset	**Proteins**	**Protein-Text Pairs**	**Text Tokens**	**Sample Weight**
SwissProt-Annot	545K	14M	40M	7
SwissProt-Aug	544K	4M	1.2B	2
SwissProt-LLM	544K	9M	2.7B	4
TrEMBL50-ProTrek	41M	412M	12B	13
TrEMBL50-ProLLM	40M	530M	143B	52
InterPro	113M	791M	1.1B	15
AFDB	214M	–	–	7

Table 3: Hyperparameters for SaProt and T2struct models.

Model name	Total Params	Encoder Params	Decoder Params	Text Length	Protein length	Learning rate	Total training steps
SaProt							
SaProt-T	760M	–	–	512	1024	5e-6	80,000
SaProt-O				768	1536		605,000
T2struct							
T2struct-224M	224M	110M	114M	768	1024	1e-4	320,000
T2struct-1.2B	1.2B	335M	992M				
T2struct-15.5B	15.5B	4.8B	10.7B			1e-5	740,000

Table 4: Examples of descriptions from various data sources.

Example Description (from SwissProt-Annot for *P95368*)

The GO term related to this protein regarding molecular function envelopes shikimate 3-dehydrogenase (NADP+) activity. In this specific protein, position 1 to 269 is marked by a polypeptide chain designated as Shikimate dehydrogenase (NADP(+)). Through molecular interaction, the residue at 89 becomes linked with shikimate. Pertinent to biological process, the GO qualifier for this protein captures aromatic amino acid family biosynthetic process. The GO term of this protein covers amino acid biosynthetic process when considering biological process. In terms of molecular affinity, the shikimate is targeted by the residue at 64. Under the scientific classification, the source organism for this protein aligns with Neisseriaceae. In the interaction between molecules, the residue at 17 to 19 attaches to shikimate. A bond is formed between the residue at location 154 to 159 and the NADP(+). Member of the shikimate dehydrogenase family.

Example Description (from SwissProt-Aug for *P95368*)

This protein aligns with the shikimate dehydrogenase family. The sequence similarities suggest that it shares a common evolutionary history and structural or functional characteristics with other members of this family. The shikimate dehydrogenase family is known for its role in the metabolic pathways involving the shikimate pathway, which is critical for the biosynthesis of various compounds, including aromatic amino acids and other secondary metabolites.

Example Description (from SwissProt-LLM for *P95368*)

This protein is engaged in a critical enzymatic operation within the metabolic pathway of the organism it is associated with. Specifically, it functions as a shikimate dehydrogenase, which is a part of the shikimate dehydrogenase family. Its primary enzymatic operation is the catalysis of the reversible NADPH-linked reduction of 3-dehydroshikimate (DHSA) to produce shikimate (SA).
This reaction is a key step in the biosynthesis of aromatic amino acids, which are essential components in various biological processes. The protein's activity is linked to the chorismate biosynthetic process, which is a pivotal step in the production of aromatic amino acids, including tryptophan, phenylalanine, and tyrosine. The chorismate itself is a precursor in the synthesis of these amino acids, as well as in the production of other important molecules like the pigment chlorophyll.
The catalytic activity of this protein is also dependent on the binding of NADP+, which is a coenzyme that plays a crucial role in redox reactions. The reaction involves the conversion of NADP(+) to NADPH, which is then utilized in subsequent metabolic pathways. This operation is significant for maintaining the cellular redox balance, which is vital for the proper functioning of the cell.
In summary, the protein's enzymatic operation is the reversible reduction of 3-dehydroshikimate using NADPH, a process that is integral to the synthesis of aromatic amino acids and other key metabolites in the organism.

Example Description (from TrEMBL50-ProTrek for *A0A1F3YBL3*)

On the matter of biological process, the GO term specific to this protein involves bacteriocin immunity.

Example Description (from TrEMBL50-ProLLM for *A0A1F3YBL3*)

This protein exhibits a multifaceted identity within various biological contexts. Initially, it is referred to as B-type flagellar protein FliS, which suggests its involvement in the assembly and function of bacterial flagella, a crucial appendage for motility. The flagellum is a complex structure that allows bacteria to move through liquid environments, and the B-type flagellum is distinct from the A-type flagellum, which typically forms a cilium in eukaryotic cells.
Additionally, this protein has been identified under different names that point to its roles in various biological processes. It is recognized as Immunity protein HalI, which implies a potential role in bacterial defense mechanisms against phages or toxins. As Flagellar secretion chaperone FliS, it suggests a function in the proper folding and secretion of flagellar components.
In the context of the flagellar hook-basal body complex, the protein is termed Flagellar hook-basal body complex protein FliE. This suggests that the protein is part of a complex involved in the structural integrity and function of the bacterial flagellum. The protein is also associated with Uncharacterized protein YfhL and Putative protein RIG, which indicates that while its function is known to some extent, there are still aspects of its role that remain to be fully understood.
Furthermore, the protein is identified as Uncharacterized protein aq_aa15 and Cloacin immunity protein, which could mean that it is a part of a more specialized system within the Cloacin genus of bacteria or that its function is not fully elucidated in the broader scientific community.
The recurring mention of the polypeptide chain spanning from position 1 to 110 in various contexts suggests that this region is conserved and may be pivotal in the protein's structure and function, regardless of its various designations. This sequence is associated with different roles, from flagellar assembly to immune responses, indicating that this protein is a versatile player in bacterial physiology.

Example Description (from UniProtKB for *A0A0A1H0A8*)

Organic hydroperoxide resistance protein famiy.

Table 5: LLM prompt for data augmentation.

Prompt for generating protein descriptions for the SwissProt-Annot dataset

You are an expert biologist who is good at paraphrasing sentences from biological area while keeping their semantic meaning the same. Now given the sentence quoted by [], you have to try your best to create as much as possible variants of the sentence. Remind these sentences should have the same meaning. Here's the original sentence: [sentence]. Please give me as much as 10 variants of the sentence with the same meaning. Note that you should keep variants diverse as possible as you can. Here's the format:
1. paraphrased sentence 1
2. paraphrased sentence 2
...

Prompt for generating protein descriptions for the SwissProt-Aug dataset

You are an AI biology assistant capable of analyzing a single protein. You are provided with several sentences that describe the same protein you are examining. Using the information provided, you should analyze the protein and answer any related questions in detail. Rather than directly referencing the site information, use it to conduct a thorough analysis of the protein in natural language, discussing aspects such as the functions of specific sites, their relationships to other proteins, and their contributions to the overall functions of the protein. When incorporating details from the provided information, focus on analyzing the protein without explicitly stating the sources of your information. Always refer to the protein as "this protein" and maintain an analytical perspective throughout your response. If a question cannot be answered based on the available details, do not reference the source of the information. Here are the relevant sentences about the protein.

{Descriptions from SwissProt-Annot}
{Question}

Table 6: Protein sequences and associated descriptions.

ID	Sequence	Text Prompt
index_1286_0	MSDFFGPPEEELGGPKNPELKAAI AALNANPNEENEQALLEALKDARF LVPVTFDKELEEEDGKVVLDEDTK LSFLLLQNPDGEKYLPAFTSWEEL KAWDPDGKYRPVVATFEDLAKLL EKDPSIAGIALNPFGDGFVLPREAL EAILNGKPWVEEEEEVEIQVGEPK QPPEEFIEALREFLDQRPEIKAAYL KLMVQDGEESLLVVLDGEGDLQPL FEELGEAVKPAAGPLRLDILPADSP LAQQLTGNEEPFYKRKK	This protein is multifaceted, as indicated by various names associated with its function within the bacterial Type VI secretion system (T6SS). It is referred to as Protein SseB, which suggests its role in the secretion system's effector complex. The protein is also identified as TssB1, TssE1, TssA1, TssK1, TssC1, and TssF1, each of these names pointing to its involvement in different components of the T6SS. The T6SS is a complex machinery that bacteria use to deliver effectors into host cells, which can manipulate host cell functions to aid in the bacterial infection process. The common thread among these designations is the protein's contribution to the T6SS, a key virulence factor for many pathogenic bacteria. Specifically, TssB1 and SseB are related to the sheath component of the secretion system, which is crucial for the assembly and function of the secretion needle and baseplate. TssE1, TssA1, TssK1, TssC1, and TssF1 indicate different roles within the system, including the baseplate, effector translocation, and possibly other regulatory functions. The protein's primary function appears to be in the delivery of effector proteins into host cells, which can alter host cell physiology to promote bacterial survival. This role is essential for the bacteria's ability to establish infection and evade the immune response. The requirement of SseB for the correct localization of other components, such as SseC and SseD, underscores its importance in the assembly and functioning of the T6SS at the bacterial cell surface.
index_422_3	MAKGKSLQIFLVDGSPDGIRTAEIS NRTGKALVFPRSKLSKLKGREEFK QSGVYFLFGKDEDGEKLLYIGEAE AYKNGKGVLKRIRQHLKNDPELK GFDEVAIITSSDNNLNKASIKYLEY RFYELAKEAGRYKIRNGNLPKKPK LSEEDQEEMEEFLEDLKIILPGLGY DIFDPKEEEESPESEEEEEEEEE	In the context of the described protein, we can identify several modular domains, each with its specific location and category: 1. **GIY-YIG Domain**: This domain is located at amino acid positions 103 to 152. The GIY-YIG domain is a type of zinc-binding domain that is found in various nucleic acid binding proteins. It is known for its role in recognizing and binding to DNA or RNA. The protein in question contains this domain, which suggests a function related to nucleic acid manipulation or recognition. 2. **Uncharacterized Protein YBL059W Chain**: The polypeptide chain from position 1 to 193 is recognized as Uncharacterized protein YBL059W. This chain does not belong to a well-defined domain family and, as the name suggests, its function is not well-characterized. It may interact with other proteins or domains within the protein, or it could be part of a larger functional complex. 3. **Bacterial GIY-YIG-like Domain**: The protein also contains a bacterial GIY-YIG-like domain, which further emphasizes its potential role in nucleic acid interactions. This domain is typically found in proteins involved in DNA metabolism, such as those that participate in DNA repair or recombination. Regarding the category of each domain: - The GIY-YIG domain and the bacterial GIY-YIG-like domain are both zinc-binding domains. These domains are characterized by the presence of a zinc ion coordinated by specific amino acids, which is critical for their structure and function. The protein's domains are integral to its proposed function as a terminase, which is an enzyme involved in the replication of DNA and RNA viruses. The terminase is responsible for adding a poly(A) tail to the 3' end of the viral genome, a crucial step in the replication process. The presence of the GIY-YIG domain and the potential zinc-binding properties suggest that this protein may be involved in nucleic acid processing, possibly in a viral replication cycle, although the exact role and the nature of its interactions with other proteins would require further study.

ID	Sequence	Text Prompt
index_678_2	MSSIIKPFRALLPAKGLASKVASPP YDVIDEEEREELLAKSPYNFLRIDL PELELEENPENFDPEENPPFPEPKE NLEELIEEGILVRDPKPAFYVYQQE RNGRTQTGIIGCVSVDAYDSGRIK KHEKTLPEKEEDRVELIEALEANF GPIFLTYPEDPEIDDLIDEITKGKP LIDFTDEDGVRHRLWRIDDPEEIEE LQELFNKIPAIYIADGHHRAAAAN VRDERRAENPDYPEDADYDYFLA VLVPEEQLPPLYYYHH	This protein, designated as ORF187 and also referred to as the Immunity factor for TNT homolog and Toxin RelE, serves a critical role in the immunity response of the organism Borrelia bissettiae. It is an integral component of an LXG toxin-immunity module, which is a system designed to neutralize the toxic effects of specific toxins produced by the same strain. Primarily, the function of this protein is to counteract the RNase activity of a cognate toxin, BC_0920, thereby preventing damage to the bacterial cell. This neutralization activity is particularly significant in the context of bacterial biofilm formation, as LXG toxins can contribute to the disruption of biofilm structure and function. By neutralizing these toxins, the immunity protein helps to maintain the integrity and stability of the biofilm. The immunity protein is part of a larger system involving six LXG toxin-immunity modules in this strain. These modules are involved in promoting kin selection, mediating competition within biofilms, and driving spatial segregation of different bacterial strains. This suggests that LXG toxins may help to prevent internal conflict among strains, thereby avoiding a form of warfare within the biofilm community. This protein's function is also linked to intercellular competition. When the operon encoding the toxin is disrupted, it can be disadvantageous to the bacteria. However, overexpression of the cognate immunity protein can restore growth in competition with wild-type strains, indicating its role in competitive defense. The protein's domain, characterized as COMM and spanning positions 112 to 186, is crucial for its function. It is involved in mediating the interaction with the cognate toxin YobL, which is neutralized upon expression in E.coli. Importantly, this protein does not exhibit immunity protein activity against other LXG toxins, suggesting that its protective function is highly specific to the cognate toxin. Furthermore, the protein is likely to interact with the cognate toxin YwqJ, inhibiting its toxic activity. This interaction is key to the overall function of the LXG toxin-immunity module, as it helps to maintain the balance between toxin production and immunity within the bacterial population. In summary, this protein is a pivotal player in the immunity response of Borrelia bissettiae, providing protection against specific toxins and contributing to the bacterial competitive strategy in biofilm environments.
index_1192_3	MSDKKEFAKEFYERLKEEADEEVE VSFDEEEYTLSIGDKTIGLSNLAEQ LEEAPEEEHEEIIEQYLEMIEKAIGL EEEEEEFDLEKDIYPVIRSKSFVEE KEYISEPHAGDLVVTYAIDTEETY RFLTEEEFEKLPLTKEELKEIALEN LKKKAPLEYEKQDIGEGETLYFLI GDGYAASRILLPEFLEEIEEEAGTE LVVFVPHQDVLLVYGYKDEETLEE MAEMTLEMFEEGGIPITDLSFTYE DGELEPVFIFKKEKEEEE	This protein, which is a component of an LXG toxin-immunity module, serves a crucial role in the bacterial immune system by providing immunity against specific toxins. Its primary functions include: 1. **Neutralization of Cognate Toxins**: The protein effectively neutralizes the toxic effects of cognate toxins, such as YobL and YqcG, by directly interacting with them. This interaction is specific to the cognate toxin, meaning that the protein does not confer immunity against non-cognate toxins, which is a key feature of the immune system's specificity. 2. **Kin Selection and Biofilm Competition**: The protein contributes to kin selection, a process where bacteria preferentially compete with close relatives over more distant ones. This function is significant in biofilm formation, where it mediates competition between strains and promotes spatial segregation, which can help avoid intra-strain warfare and the resultant release of virulent toxins. 3. **Intercellular Competition**: The immunity protein is involved in intercellular competition during biofilm formation. The disruption of the operon encoding this protein can put the bacteria at a disadvantage in such competitions, while overexpression of the cognate immunity protein can restore growth in competition with wild-type strains. 4. **Immunity to Specific Toxins**: The protein imparts immunity to certain naturally sensitive host strains against toxins such as pediocin PA-1/ACH, suggesting its role in providing protection against a range of threats. 5. **Prevention of Early Activation of Toxins**: The protein also has a function in preventing the early activation of toxins like Tas1, which could otherwise lead to harmful effects. In summary, this protein is a critical part of the bacterial immune response, ensuring that the organism is protected against specific toxins, maintains competitive advantages within biofilm environments, and contributes to the overall health and survival of the bacterial community.

ID	Sequence	Text Prompt
index_827_2	MKKKIPTYKINDFPGKEGNQDIEV SSLEEHLKLAPHIEEPHRHDFYLILY VTKGSGTHTIDFETYPIKPGSLCFV RPGQVHSWEFSDDLEGTVIIFTED FLSKYFPNFSIEDFSFFNNSDSSPVL QIPEEDLPEFESLFEQLEEEYKSND PYKDEILASLLYQLLLKINRLYSSQ SNISNNSLELVRQFKQLVEENFKTE KQVSFYADKLNITVKTLNEITKKIT GKTPSQLIQDRIILEAKRLLYYSDLS IAEIAYLLNFNDSSHFSKKFKKTGG MSPKERRKKLK	This protein exhibits several DNA-binding domains, which are crucial for its interaction with DNA and, by extension, its regulatory functions. The primary DNA-binding domain within this protein is characterized by the presence of the h-T-H motif, a common structural feature in many DNA-binding proteins. Specifically, the h-T-H motif is found in several overlapping regions throughout the protein, suggesting that the domain may be quite stable and possibly engage in multiple DNA-binding events simultaneously. The following are the placements of these h-T-H motif-containing domains: 1. From position 204 to 225, this domain is identified, which is consistent with the descriptions provided. 2. The domain is also noted to be present from position 203 to 224, which slightly overlaps with the previous span but maintains the core motif. 3. Another instance of the h-T-H motif is located from position 201 to 222, with a minor overlap in the amino acid indices. 4. At positions 251 to 274, another region of the protein is marked as having the h-T-H motif, which is again slightly overlapping with the previously mentioned domains. 5. The domain is also described from position 252 to 275, with a very slight shift in the end position. 6. At position 200 to 221, there's another mention of the h-T-H motif, which overlaps with the regions described earlier. 7. Lastly, the domain from position 249 to 272 is also noted, showing a broad consistency with the earlier mentioned domains. In addition to the h-T-H motif, the protein is characterized by a domain from position 187 to 285, described as being of the HTH araC/xylS-type. This type of domain is often found in regulatory proteins and is known for its ability to bind to specific DNA sequences. The similarity with the araC/xylS-type suggests that this protein might be involved in regulatory processes, such as gene expression control. The consistency in the description of the DNA-binding domains across the various spans indicates that these domains are likely to be functionally important, and their presence in multiple, overlapping regions could contribute to the protein's robust DNA-binding capabilities. This could be significant for processes such as transcriptional regulation, where precise and multiple interactions with DNA are often required.
index_3513_3	MKSLKELLNLSEEELKETKVHIISD PENNPLLKAYLNGDLEKYEKLWK KRNAFGDKKYILVFLKEKGSNEYT FLGLYEIISSEKTENGYKYEVELLD EFEELEGRLIVEYKPPRRSTYRNLE KYKDKLIVEGILEEPIKLPEFPGYE KLVLSYDQLKEIIENNNESWKNALS SVKGIYLITDTKTGKQYIGSAYGG NGLWGRWKNYVKDGHGGNKKLK KLIEKNPEYKKNFRFSILEVAPDNA TEEDILKREKEYYKLLLTKKKGLN LN	This protein, designated as ORF319, is a probable endonuclease, with a significant role in intron homing. To deduce the specific cofactor that aids this protein in its operational duties, we can consider the broader context of endonucleases and their functions. Endonucleases are enzymes that cleave DNA or RNA within the sequence of the molecule, which is a critical step in various cellular processes such as DNA repair, recombination, and gene regulation. Given the role of ORF319 in intron homing, which is a process involving the movement of introns within the genome, the cofactor likely plays a crucial role in facilitating this movement. For proteins involved in nucleic acid processing, a common cofactor is magnesium ions (Mg2+). Magnesium ions are known to stabilize the active site of many endonucleases and are essential for their catalytic activity. They can help in orienting the nucleotides correctly for cleavage and can also modulate the enzyme's affinity for its substrate. Another potential cofactor could be a metal ion like zinc (Zn2+), which is also known to be involved in the activity of certain endonucleases. Zinc ions can participate in stabilizing the active site of the enzyme and in the catalytic reaction itself. Since ORF319 is similar to endonucleases of group I introns of fungi and phage, it is possible that it also relies on a cofactor that is typical for these enzymes. For instance, some group I intron endonucleases require divalent cations, such as Mn2+ or Mg2+, to activate the enzyme. Given that the protein is encoded within intron 1 of COX1 and is involved in mobile group II introns, it's also plausible that the protein could have additional cofactors or cofactor-binding motifs that are specific to its intron-encoded nature or its role in intron movement. To definitively identify the particular cofactor that aids ORF319 in its operational duties, experimental analysis would be required to determine the conditions under which the protein exhibits its endonuclease activity. This might involve testing the activity in the absence or presence of various potential cofactors and monitoring the effects on the enzyme's function.
index_201_0	EGWTVKTLILNSATYRQSSKVNPE KLEIDPENRLLARQNRRRLEAEMI RDQALAVSGLLSPKIGGPSVKPYQ PEGIWENSNSKAKWKQSKGEDLY RRGLYTFWKRTFPYPSMLAFDAP DRNVCCVRRERTNTPLQALVLLND PTFVEAARALAERILKEAKTDEERI EYAFRLALSRLPSAEELAVLLQSLQ QQLAQYQANPEAAEALLAVGETK RPADLDPVELAAWTNVANVLLNL DEFITKE	The gene responsible for synthesizing the protein in question is identified by the ORF Name WH7805_09909. This gene is one of several that have been associated with the protein synthesis process, as indicated by the other gene names provided: EN45_078720, EN45_076310, NAS141_03721, NIES39_K04640, UT29_C0001G0139, NIES39_K04650, and DVH24_017674. However, the specific gene named WH7805_09909 is highlighted as the primary identifier for the gene involved in the synthesis of this protein.

ID	Sequence	Text Prompt
index_558_4	MSELSPLERLLERLRREREEAEAR GKEYKPTWGFVIYRTVYPEDDEK WEEFLELLKEETEEELRSSPGLEW LLDYLEWTVIEDPEKLDGASKEQV REHFREWVAEELNSPSGPPPSEGL DPDSSPRYRVCLQIDEECLDHFLSN LEDGKLKPDADPFVYAVEADWDE EEDENADEDEPYVGWMKVPIRDL FDELYLNLGFEDFEDLREDADEDG VYRPDGDD	The gene responsible for coding this protein is identified through various ORF (Open Reading Frame) names. These names include TRAVEDRAFT_139011, AA0117_g3624, VD0003_g7577, AA0117_g5322, BDW47DRAFT_120076, BDW47DRAFT_120077, and BDW47DRAFT_134043. These ORF names suggest that the gene could be found in different organisms and strains, as indicated by the associated organisms' names such as Galerina marginata (strain CBS 339.88), Ceriporiopsis subvermispora (strain B), and Ustilaginoidea virens. The diversity in the ORF names and associated organisms suggests that this gene may have been identified across different studies and genetic databases, potentially highlighting its importance in various biological contexts.
index_1157_3	MKKKLLLAALLLLAASAAQAHEFW LEPEQFQVSASGSATLQLRVGEET EHQPFPLRPERIASFSVVGPDGKK TLPAEQLAGSIPLGQPGLHVVSLET TPRPITLEAEKFNEYLKEEGLTDV LAARKAAGKAAQPAKEQRYAKAL VGVGDGKGADKPLGLPIEIVALAN PYGADLSDGLTVQVLYDGKPLAN VKVTVFEKSADGWVKVSTVKTDA NGRATLPVEPGHRYLLSAVKMRP AADPAKADWLSYFASLTFFAPK	In the context of the mature form of the protein, the polypeptide chain corresponds to the region following the removal of the propeptide. Specifically, the mature protein's polypeptide chain is recognized as the Putative protein-disulfide oxidoreductase RP025, spanning from position 22 to 272. This region is of particular interest as it is where the protein's functional activities are likely centered, given its classification as a protein-disulfide oxidoreductase. This area is responsible for the protein's disulfide bond formation and reduction, which are critical in many biological processes, including protein folding and stability. The mature form does not include the propeptide, which is situated between positions 247 to 272, suggesting that this segment is not involved in the protein's mature function and is shed during the maturation process.
index_1133_0	MWLLNTETLELKEFDDPESVPYAI LSHTWGDEEVTFQDMINGTNKDS AGYKKIKKFCELARERYGIEWIWI DTCCIDKSSSAELSEAINSMFRWYQ LAEVCIVYLSDVPAPSDPSDLDSPE FEEAFRNSRWFTRGWTLQELIAPR SVEFFDKDWQLIGTKRSLADLISEI TGIDADVLRGEAPLSDFSVAERMS WAANRETTREEDRAYCLLGIFDVN MPLIYGEGDKAFRRLQEEIMKRIP DDSIFAW	The gene responsible for coding this protein has been identified through various ORF (Open Reading Frame) names. These names are VFPFJ_04709, VFPFJ_04705, VFPFJ_04707, VFPFJ_04702, VFPFJ_04704, and VFPFJ_04701. These ORF names suggest that the gene in question is likely part of a larger genomic context within the organism being studied. Each ORF name might correspond to a specific region or variant of the gene, potentially reflecting different alleles or isoforms of the protein. The use of these distinct ORF names may indicate that the gene has multiple transcripts or that it has been studied in various genetic backgrounds or conditions.
index_672_1	MRLLNTRTLELEEFLDDDIPPYAIL SHTWEDEEVSFQDIHPEEKRKKKK GYAKIKRCCELALKDGLEYVWVD TCCIDKTSSAELSEAINSMYRWYQ KAEVCYAYLSDVSSDDDPETVLPK CRWFTRGWTLQELIAPSSVEFFNK DWEEIGTKDSLADLISEITGIPIDVL RHEKPLSSVSVAERMSWAANRETT RVEDRAYCLLGIFDVNMPTIYGEG ENAFRRLQEEIIKRSDDQSIFAWGD PSSPSDLSSSPDDLLLL	This protein, which is part of the Heterokaryon incompatibility protein family, is associated with a complex biochemical process that is pivotal in the organism's biology. Specifically, the protein contains a domain that is annotated as a glycosyltransferase domain, which is characteristic of enzymes that catalyze the transfer of glycosyl groups from a donor to an acceptor molecule. The glycosyltransferase activity in this protein suggests that it may be involved in glycosylation processes, where sugars are added to other molecules, such as proteins or lipids. This modification can significantly impact the structure and function of the glycoprotein or glycolipid, potentially affecting its stability, solubility, and interaction with other proteins or cellular components. Given the context provided, this protein's catalytic function is likely related to the synthesis or modification of certain toxins. The mention of mutated toxins being expressed as transgenes in yeasts that cause lethality in lepidopteran larvae suggests that the glycosyltransferase activity may be part of a pathway that modifies or synthesizes these toxins. This modification could be crucial for the toxin's efficacy, allowing it to evade the host's immune system or enhance its ability to disrupt cellular processes in the larvae. Furthermore, the protein's impairment of pathogenicity on wheat lines with the Snn1 gene implies a role in interactions with the plant's defense mechanisms or perhaps in the protein-protein interactions that are essential for the pathogen's virulence. The glycosyltransferase domain may be involved in these interactions by modulating the protein's structure to facilitate recognition or binding to specific receptors or targets. The down-regulation of this protein's expression during fruiting body formation suggests that its function may be particularly relevant during certain stages of the organism's lifecycle, possibly indicating that glycosylation or toxin modification is a critical aspect of the organism's reproduction or environmental adaptation. In summary, the biochemical process that this protein likely catalyzes involves the glycosylation of molecules, potentially modifying toxins to enhance their biological activity and contributing to the organism's ability to interact with its environment, either as a pathogen or in other ecological roles.

ID	Sequence	Text Prompt
index_934_0	MRLLNTETLELEEFLGDIPPYAILS HTWEDGEEVTFQDLQEGKGESKS GYAKIKRCCELARKDGYEYVWIDT CCIDKSSSAELSEAINSMFRWYQEA EVCYAYLSDVPSIPISEENEEELEA AFRKSRWFTRGWTLQELLAPRSV EFFSKDWQLLGTKSSLADLISEITG IPVEALRGSPLSSFSVAERMSWAAS RETTRVEDMAYCLLGIFDVNMPLI YGEGENAFRRLQEEIMKSSDDHSIF AW	The protein in question originates from the organism Fusarium acaciae-mearnsii. This species is a fungal pathogen that has been identified in association with the protein sequence.
index_2830_4	MLLLTKKELEELLEESNDLPDWQK EAFEWFKKKMTNKTNPFPCIFGK KAYKKDQLRYAFAEQGDPDAAEE IAEALPLLPEISPYSSLVVFFEDLGE DATIEDYEQLFWDTLKEVHALDPV PWPEDIPKDPEHPLWEFSFNGEPY FIIGMTPAHKKRKSRRFPYLMLVF NPRSQFEYLREKGRYEKTKKKIRK RLDKLDSVPLPPLLGSYGSKENSE WKQYFLPDEDKPCPFK	In analyzing this protein, several potential errors or areas of confusion can be identified: 1. **Protein Length Ambiguity**: The protein, referred to as Protein YecM, is described to have a polypeptide chain spanning positions 1 to 188. However, it is also noted that this protein may be missing up to 130 C-terminal residues compared to its orthologs. This discrepancy raises questions about the true length of the protein and how this affects its structure and function. It could lead to confusion regarding the complete sequence and the role of the truncated portion of the protein. 2. **Original Misclassification**: The protein was initially thought to encode for OrgA and OrgB, which suggests that its function and role within the cellular context were initially misunderstood. This misclassification could have implications for how the protein is studied and its current classification as a putative immunity protein component. 3. **Functionality of C-Terminal Region**: The protein's function is linked to blocking the toxic effects of the C-terminus of cognate toxin RhsC. The fact that this protein is missing up to 130 C-terminal residues could potentially affect its ability to perform this function, as the C-terminus is crucial for toxin interaction and neutralization. 4. **Sequence Similarities**: The protein shares sequence similarities with E.coli YfiI and P.aeruginosa RluD. This raises questions about the conservation of function across different organisms and the significance of these similarities in the context of the protein's role in a toxin-immunity protein module. 5. **Interactions with Other Proteins**: The protein interacts with YeeF but not with YobL, which might be due to the differences in their C-termini. The interaction with YeeF is thought to inhibit the toxic activity of the toxin. The stoichiometry of this interaction (2:2) suggests a precise requirement that could be difficult to achieve, leading to potential confusion in experimental designs and interpretations. 6. **Gene Name and Protein Synthesis**: The nucleotide sequence yubF corresponds to the synthesis of this protein. However, without a clear understanding of the complete protein sequence and its structure, the relationship between the gene sequence and the protein product might be difficult to interpret fully. 7. **No Obvious Phenotype on Deletion**: The deletion of the yeeF-yezG operon does not lead to a visible growth phenotype. This could be confusing, as it suggests that the function of this protein might not be essential for growth under certain conditions, or it could be that its role is masked by other cellular processes. Overall, these ambiguities and potential errors highlight the importance of further research to clarify the protein's structure, function, and interactions within its cellular context.
index_1157_2	MKKLLLLLLLLLLLSSAAAQELNCQ VQVNAPQVNADREVLRSLQKAITD FLNTTKWTDEEFEEEERIQCNILIN LTEQPSENTFKGTLQVQSSRPVYG TNYSTPLLNFRDKNFEFEYIEFQPL EFNPNTFVSNLTSILAFYAYLILGL DYDSFSLNGAKPYFEKAKNLVNNA QNAGAPGWSDSNVKVNRYWLIEN LLNPNFADVREAMYDYHREGLDR FSEDKEKAKKAIKEAIPKLKKAKR ARPGSYLLKALLSSKEEELRRIGGK KNINN	In the context of the mature form of the protein, the polypeptide chain corresponds to the region following the removal of the propeptide. Specifically, the mature protein's polypeptide chain is recognized as the Putative protein-disulfide oxidoreductase RP025, spanning from position 22 to 272. This region is of particular interest as it is where the protein's functional activities are likely centered, given its classification as a protein-disulfide oxidoreductase. This area is responsible for the protein's disulfide bond formation and reduction, which are critical in many biological processes, including protein folding and stability. The mature form does not include the propeptide, which is situated between positions 247 to 272, suggesting that this segment is not involved in the protein's mature function and is shed during the maturation process.

ID	Sequence	Text Prompt
index_2181_0	MRLLNTRTLELEEFFGDDIPPYAIL SHTWGDEEVTFQDMINGTAESKK GYAKIQECCELAREDGYEYVWVD TCCIDKSSSAELSEAINSMYRWYQK AEVCYAYLSDVPSDDDPRAEGSLF RKSRWFTRGWTLQELIAPSSVEFF SQEWEEIGTKRSLEQLISEITGIDID VLRHERSLSDFSVAERMSWAANRE TTRVEDRAYCLLGIFDVNMPLIYG EGDNAFRRLQEEIMKKSDDHSIFA W	The protein in question is associated with multiple gene names, which suggests that it may be encoded by more than one gene. The gene names MAJ provided include_08936, MAJ_07524, MAJ_07522, UVI_02036180, MAJ_09462, and MAJ_07523. Additionally, there are two similar gene names, MAJ_09464, which is likely a typographical variation of MAJ_09462. It is common for proteins, especially those involved in complex biological processes such as heterokaryon incompatibility, to be encoded by multiple genes, which might contribute to different variants or isoforms of the protein with distinct functions or roles within the cell. In the context of the information provided, it is not possible to determine a single specific gene name that uniquely codes for this protein without additional information about which gene is the primary or the most relevant for the protein's function. If we were to prioritize based on the number of mentions, MAJ_09462 (or its possible typographical variation MAJ_09464) might be considered a strong candidate, as it is mentioned twice along with its corresponding protein name, Heterokaryon incompatibility protein S. However, without further context or experimental data, we cannot definitively state that this is the gene that codes for the protein.
index_2536_3	MSDPYEELREKLDAHPVPVPDTPE LREILKLLFTPEEAALAAKLPFKPL TVEEAAKLTGLPEEEVEALLERMA EKGLVLRTEKDGERRYMLAPFVV GIFEFQFMMKRDLGERGKELAELF KEYLEDGEFAEFLLAGEVPMMRV VPVESAIPPDQEVLPYEKVKEHIER ADKIAVAECFCRKQKELLGKDCGK PLETCMALNRGAEYYIEQGFAREI TKEEALEILDEAEEAGLVHQVDNS QEELGFICNCCGCCCGLGKAALLL GAGGAMAASTFTADVDPEKCVCC GACVAKCVPGAIEFDEEKKKATV TDDCCLGCGCCVTVCPGFAITIER RDGEEEPPPPSEELLKKLIKAKRA AAA	In the final product of this protein, the polypeptide chain is formed by the sequence of amino acids that make up the entire protein structure. Specifically, the polypeptide chain extends from position 1 to 373, as indicated by the Chain description. This chain is recognized as an uncharacterized protein, which suggests that while its structure is known, its specific function or role within biological processes may not be fully understood yet. The sequence from position 1 to 373 is likely the primary structure of the protein, which then folds into a three-dimensional structure, potentially involving the domains and motifs described. For instance, the domain from position 287 to 318 is characterized as a 4Fe-4S ferredoxin-type domain, which is known to be involved in electron transfer reactions. This domain is repeated, with the motif at position 306 to 331 being classified as Cx9Cx9RCx2HK, indicating a pattern of repeats and possible functional significance. The region from position 1 to 373 also encompasses the binding site at position 320, where the residue forms a bond with the [4Fe-4S] cluster, suggesting that this site is crucial for the protein's function, possibly related to iron-sulfur cluster assembly and electron transfer, as indicated by the mutagenesis data. The mutation from cysteine to serine at position 321 affects iron-sulfur cluster assembly, which underscores the importance of this region in the protein's activity. The protein is part of the NapF family, and its gene sequence corresponds to the nucleotide sequence rnfC, which is indicative of its genetic origin and potential involvement in cellular processes related to iron-sulfur cluster metabolism or electron transport. In summary, the polypeptide chain from position 1 to 373 is the part of the protein that forms the backbone, upon which the protein's structure and function are based, including the iron-sulfur cluster interaction and its potential role in electron transfer.
index_672_2	MRLLNTRTLELEEFIGDNIPPYAIL SHTWEDEEVTFQDMQDPDRESKK GYAKIKRCCEQALKDGLEWVWID TCCIDKTSSAELSEAINSMYRWYRE AEVCYAYLSDVSGDGFSEEEFRKS RWFTRGWTLQELIAPRSVEFFSKD WGRIGTKRSLEEEIHEITGIPIEALR GASLSDFSVEERMSWAAKRETTRE EDMAYCLLGIFDVNMPLIYGEGDK AFRRLQEEIMKSSDDDSLLSSLLEL LSSSSSSSSSGPPPDDDSLLSSSSSSSS	This protein, which is part of the Heterokaryon incompatibility protein family, is associated with a complex biochemical process that is pivotal in the organism's biology. Specifically, the protein contains a domain that is annotated as a glycosyltransferase domain, which is characteristic of enzymes that catalyze the transfer of glycosyl groups from a donor to an acceptor molecule. The glycosyltransferase activity in this protein suggests that it may be involved in glycosylation processes, where sugars are added to other molecules, such as proteins or lipids. This modification can significantly impact the structure and function of the glycoprotein or glycolipid, potentially affecting its stability, solubility, and interaction with other proteins or cellular components. Given the context provided, this protein's catalytic function is likely related to the synthesis or modification of certain toxins. The mention of mutated toxins being expressed as transgenes in yeasts that cause lethality in lepidopteran larvae suggests that the glycosyltransferase activity may be part of a pathway that modifies or synthesizes these toxins. This modification could be crucial for the toxin's efficacy, allowing it to evade the host's immune system or enhance its ability to disrupt cellular processes in the larvae. Furthermore, the protein's impairment of pathogenicity on wheat lines with the Snn1 gene implies a role in interactions with the plant's defense mechanisms or perhaps in the protein-protein interactions that are essential for the pathogen's virulence. The glycosyltransferase domain may be involved in these interactions by modulating the protein's structure to facilitate recognition or binding to specific receptors or targets. The down-regulation of this protein's expression during fruiting body formation suggests that its function may be particularly relevant during certain stages of the organism's lifecycle, possibly indicating that glycosylation or toxin modification is a critical aspect of the organism's reproduction or environmental adaptation. In summary, the biochemical process that this protein likely catalyzes involves the glycosylation of molecules, potentially modifying toxins to enhance their biological activity and contributing to the organism's ability to interact with its environment, either as a pathogen or in other ecological roles.

ID	Sequence	Text Prompt
index_433_4	MSSTSSPSPAGPVGAEIRPFRALRY NPEKIADLSDVVAPPYDVISPEEQE RLYARHPYNIIRLILGKEEPDDSEE NNRYTRAAELLEEWLEEGVLVQD PEPAIYVYRQEFTLDGKEVTRTGF IAAVKLEPFGTGAVLPHEQTLAKP KADRLKLIEATRANLSPIFLLYRDP ERAVARILEEVTAGEPLIEVTDEDG VTHRLWRISDPADIAAIQELMADK SLLIADGHHRYETALNYREEHGDT SPGADYVLATLVNLDDPGLPIPPIH VVRLLGPEDDDELLLEELLAGLFE VEVAADEEELLAALAAQGPGAFAL RFADSGTLLLLRLDDEAALAELLP DLPEALASVDDALLAALLLPLLLGD DAEELEEGGRLVYVRDEEELAALL AEGDAAFFLLNPPPIEVVVEAAGE GGPQQQKYTYFPPKRLTTLFFRPL L	The organism that serves as the source of this protein is Deinococcus erythromyxa. This organism is identified in the context of the protein sequence provided.
index_752_0	MRLLNTRTLELEEFLGDPPPYAILS HTWGDEEVTFQDMQNGTYKGKA GYEKIKRCCEQALRDGLEYVWIDT CCIDKSSSAELSEAINSMYRWYRDA AVCYAYLADVPSDDDPETVLPKC RWFTRGWTLQELIAPRSVEFFNKD WELIGDKASLADLISEITGIDADVL NGEKPLSSASVAERMSWMANRET TRVEDIAYCLLGIFDVNMPLIYGEG DKAFRRLREI	This protein, which plays a crucial role in the expression of heterokaryon incompatibility and sexual functions, offers several intriguing insights regarding its structure, function, and regulation: 1. **Genetic Identity**: The protein is encoded by the gene ORF Name VFPFJ_04705, suggesting a specific genetic heritage within the organism. This genetic identity is essential for the proper expression of the protein's functions, which are integral to the organism's reproductive biology. 2. **Functional Significance**: As a component of the heterokaryon incompatibility system, this protein is involved in preventing the fusion of genetically distinct fungal hyphae. This is particularly important in fungi like the Oat take-all root rot fungus, which can lead to significant crop losses if not controlled. Its role in sexual functions suggests that it may also be involved in the mating process, which could have implications for the genetic diversity and adaptation of the fungus. 3. **Disruption Phenotype**: The disruption of this protein impairs the production of victorin or any derivative or intermediate. Victorin is a known antifungal compound, so the loss of this protein might result in increased susceptibility to fungal infections. This highlights the protein's role in defense mechanisms against pathogens. 4. **Expression Regulation**: The protein's expression is influenced by various environmental factors and regulatory proteins. It is repressed by glutamine and at alkaline pH, which could be a response to certain nutritional or environmental stresses. Conversely, it is highly induced under nitrogen starvation and acidic pH conditions, indicating its role in stress responses. The negative regulation by vel1 further suggests complex regulatory pathways that may be involved in fungal development and survival. 5. **Structural Analysis**: The protein's chain, spanning from position 1 to 680, is recognized as Heterokaryon incompatibility protein 6, OR allele. This implies that the protein has a defined structure that is crucial for its function. Analyzing this structural information could provide insights into how the protein interacts with other molecules and performs its functions. 6. **Biotechnological Implications**: The use of lichens containing usnic acid and the mention of chondramides, which are secondary metabolites with antifungal activity, suggest that the protein and its associated pathways might be of interest in biotechnology. These compounds could potentially be developed into antifungal agents, and understanding the protein's role could lead to more effective control strategies for fungal diseases. 7. **Potential for Toxicity**: The protein's link to usnic acid, which is toxic to the liver and can cause contact allergy, might suggest that the protein itself or its metabolites could have similar properties. This could be a concern in therapeutic applications or when using the protein as a model for studying fungal biology. In summary, this protein is a multifaceted molecule that has a significant impact on fungal biology, including sexual reproduction, defense against pathogens, and stress response. Its complex regulation and potential biotechnological applications underscore the importance of further study into its mechanisms of action and structure.

ID	Sequence	Text Prompt
index_3236_1	MSAPTAPVGVRPFRGVRYDPAKV GDLSRVVAPPYDVIDPALQERLAA RAPHNVVRLELGKEEPGDPPDDN RYTRAAALLAEWLADGVLVRDPE PALYVYRQTFTVDGRTYTRRGFIA AVRLGDFAAGGAVLPHEHTLSGPK ADRLALLRATRANFSPVFALYDDP GGAVAALLDELAAGPPDLEVTDD DGVRHRLWRVDDPADIAALAAAF AGRPLYIADGHHRYETALAYRDER RAETGAGADAPYDYVLAMLVNLP DPGLVLPTHHRVVRGDDDALAAA AAAAVFEVVEVADAAALAAALAA PGGGARRRALRLGDGRFVVLRLR RPGAAAAAALAGGDRPPLDALAL HLGLLLALLGLSAEAAAAHLELVR DAAEAAAAAAAGGARAAAFPLPT EMDLAAAVAAAGAPLPKKKTYFY PKPLSGGLIRRL	The gene responsible for producing the described protein is identified by the ORF (Open Reading Frame) name GA0070610_0181. This gene encodes the protein in question, which is associated with a specific protein family and has a conserved domain structure. The ORF name serves as a unique identifier within the genomic context of the organism from which the protein is derived.

Table 7: Pinal prompt texts for different protein types.

Protein Type	Prompt Text
ADH	Please design a protein that is an alcohol dehydrogenase.
GFP	Green Fluorescent Protein (GFP) is a light-emitting protein originally isolated from the jellyfish Aequorea victoria. When exposed to blue or ultraviolet light, its unique structure—a beta-barrel encasing a self-formed chromophore—absorbs energy and re-emits it as bright green fluorescence. Unlike many fluorescent markers, GFP requires no external additives (e.g., substrates) to glow, making it ideal for non-invasive live-cell imaging.
PETase	PET hydrolases are a class of enzymes capable of catalyzing the degradation of polyethylene terephthalate (PET), a widely used plastic found in bottles and packaging. This enzyme class includes cutinases, lipases, and esterases, which hydrolyze the ester bonds in PET through a conserved catalytic triad (Ser–His–Asp/Glu). PETase is a well-characterized PET hydrolase that specifically targets and breaks down PET polymers.
Hprotein	Design a H protein. H protein acts as an essential mobile shuttle within the glycine cleavage system, transferring the methylamine intermediate between the P and T proteins, thereby enabling glycine breakdown or reversed glycine synthesis. Simultaneously, it serves as a crucial carrier protein intermediate in the protein lipoylation pathway, facilitating the attachment of essential lipoic acid cofactors to other key metabolic enzymes. Its covalently bound lipoamide arm is central to both carrier functions.

Table 8: Amino acid sequences of fluorescent proteins.

ID	Sequence
pinalGFP	MSAAAKLLEGTVKIQMTGSGDIGGQKFTYTATGTVNPATGTLDGTIRSEDGKLPFHPSFLIGLGTYGRRGFSA PDPDIPEALAFLDTLPAGYTFERTIRIEDGPTIESEGTVTFEGNEITIETKIEVTNIDPDDPLASGKLTPDIPPYHE QYIPDPETKALRSTRALFIPLKGGGKVAFSVEQLYTPLSDGPFPLPGTHLAHVTQSWSKVEGSEGPTLEFTESA RAFSIVPPDEAKLL
esmGFP	MSKVEELIKPDMKMKLEMEGEVNGHKFSIEAEGEGKPYEGKQTIKAWSTTGKLPFAWDILSTSLTYGNRAFT KYPEGLEQHDFFKQSFPEGYSWERTITYEDGATVKVTADISLEDGVLINKVKFKGENFPSDGPVMQKKTTGW EASTELITPDPATGGLKGEVKMRLKLEGGGHLLADFKTTYRSKKKEKLPLPGVHYVDHRIVNEKATHPEGKE YMIQYEHAVARLA
esmGFP-b8	MSKVEELIKPEMKMKLEMEGEVNGHKFSIEAEGEGKPYEGKQTIKAWSTTGKLPFAWDILSTSLTYGFRMFT KYPEGLEEHDYFKQSFPEGYSWERTITYEDGATVKVTSDISLEDGVLINKIKFKGTNFPSDGPVMQKKTTGWE PSTELITPDPATGGLKGEVKMRLKLEGGGHLLADFKTTYRSKKKEKLPLPGVHYVDHTIRNEKAPHPEGKEY VVQYETAVARLA
avGFP	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSR YPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNY NSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHM VLLEFVTAAGITHGMDELYK

Table 9: Amino acid sequences of ADHs.

ID	Sequence
pinalADH01	MSLKGKNVLFVGGLGGIGLATCRALVKKNLKNLAILDIVENPEAVEELKALNPSVKVTFVKCDVTKPESIKAAF EEVKEKFGHLDVLVNGAGILDDKNIEKTIAVNLTGLINTTLAALPLMDKRKKGKGGVIVNVASVAGLEPFPAV PVYCASKHGVVGFTRSLGHPFYYELTGVKVIAVCPGITETPLLKNLGKNPLFDEFKELVEKLLKLKKQTPEEC GKHIAKVIETAENGSIWKSDKGKLSELELPPWWKPPKS
pinalADH02	MNLNEFYVPGKNLFGAGSIKELAEELKGKGYKKALLVTDKFLVKLGLADKLKALLAEAGIETVVFDGVEPNP KDTNVMEGLAVFKENGCDFIIALGGGSPHDCAKAIALVATNGGSIRDYEGVDKSKKPCLPLIAINTTAGTGSEM TRFAVITDEERHHKMVIVDKHVTPKIAVNDPELMVGMPPGLTAATGMDVLTHAIEAYVSTGANPITDALALQ AIRLIAKNLPKAVKNGSDLEAREQMHYAQMLAGMAFNNALLGVHHAAAHALGGFYNPPHGAGNVVVLVVVM YYNAAVVPEKFAIIAKAAGLDVDGDSEEEAAEKAAEVVKDLLKDVPIPLKLSGGGVKKEDIIPLAELAKKDACA ATNPRPSSLEEIIDIYEEAF

Table 9 – *Continued from previous page*

ID	Sequence
pinalADH03	MSLEGKTIVVTGASSGIGLATAKLLQERGAEVIGVDIVAPDFEVAQFIQADLSTPEGVEAALAQLPEQIDGLVN NAGVGPSAPAELVLAVNLLALVALTEALLPRVPPGGSIVNTSSNAGRLWRDDPEELEELLAAETPEELEAYLAA NPIPKEEAYAFSKALVIALTRRLALPLFRERGVRVNAVAPGLVDTPILDGFVEAMGEEAVAALLALQPRLAQPE DVANVIAFLASDESRWITGQVIFVDGGLSLLLR
pinalADH04	MKAAVLEEFGKPLEIVEVPKPEPKGEQVLIKVEAAGVCHSDVHIWTGKYGGWDLEEDFGFKLPFTLGHEVAG VVEAVGPEVEGWKVGDRVAVYPWIGCGDCRYCRSGEENLCENGRWIGLTVDGGFAEYVLVPDARYVVKIEN LSPEEAAPITCAGVTVYRAVKEANPTSSDTVAILGAGGGLGTLAVQYAKALSDAKVIAIDIRDEGLELAKEMGA DVVINSATEDVVEEVKEITKGRGVDAVLDFVGSPATWEKAPKLLAPGGTLTLVGVGGPPPPPPVLLAVGGGK IIKGSYYGNRRDLEEVEEFVKGGVPPPPEIETIPLEEAAEGLEKLLLGKIIGRYVLEP
pinalADH05	MKINEFYVPGKNLFGAGSIKELAEELKGKGYKKALLVTDKFLVKFGLADKLKALLAEAGIETVVFDEVEPNPK DTNVMEGLAVFKENGCDFIIALGGGSPHDCAKAIALVATNGGSIRDYEGVNKSKKPCLPLIAINTTAGTGSEMT RFAVITDEERHHKMVIVDKHVTPTIAVVDPELMVGMPPGLTAATGMDVLTHAIEAYVSTGASPITDALALQAI RLIAKNLPKAVKNGSDLEAREQMHYAQMLAGMAFNNALLGVHHAAQHQLGGFYNPPHGAGNVVVLVVVMY YNAAVVPEKFAIIAKAAGLDVDGDSEEEAAEKAAEVVKDLLKDVPIPLKLSGGGVKKEDIIPLAENAKKDACA ATNPRPSSLEEIIDIYEEAY
pinalADH06	MKAAVLEEYGEPLEIEDVPDPELGPGQLLVKVKACGVCHSDLHVANGDWLDVLPSPPPKLPLVLGHEVAGVV EEIGEGVEGFKVGDRVVVYPWIGCGHCRYCRSGRETLCENPGWIGITVDGGYAEYVLVPDPRYVVKFEDLSF EEAAPITCAGVTVYKALKEANLTPDDWLLILGAGGGLGHLAVQYAKALTGAKVIAVDPREEGRELAKEMGAD HVLEEIDAEAIKELTGGRGADAVLDFSGNPAVLDAALDLLRPPGRLVLVGLLGGPGPDDNLLLLILGLKIITGLL VGTRRDAEEVLDFVKGGIVKSIVSVRPPLEEAAEAYYDLLLGKITGRVVLEVP
pinalADH07	MAFSVANKNVIFVAGLGGIGFDTSKQLVKKNLKNLVILDRIEKPAAIEELKALNPKVKVSFYPYDVTVPLAETV ALLAKIFSELKTVDVLVNGAGLLDDHQIERTIAVNYTGLINTTTAILNFWDKRKGGPGGVICNICSVTGFNAIYQ VPVYSGTKHAVVNFTSSLAKLAPITGVTAYAINPGITRTPLVHHFNSWLDVDPKIAENLLEGPTQTSLQCGKNF VKAIESNKNGAIWKLDLGKLEPVEWTKQWDKGN
pinalADH08	MSLTNKNIIFVAGLGGIGLDTSKELVKKNLKNLVILDRIDNPAAIAELKAINPKVKVSFYPYDVTVPVAETTKLL KSIFDKIKTVDVLVNGAGILDDHQIERTIAVNYTGLINTTTAILNFWDKRKGGPGGVIVNICSVTGFNAIYQVPV YSASKAAVVNFTNSLAKLAPITGVTAYAINPGITKTPLVHHFNTWLDVDDKAAENLLEHPTQTSLQCAKNFVK AIEANKNGAIWKLDLGKLEPIEWTKQWDSNI

Table 10: Amino acid sequences of PETases characterized in this study.

ID	Sequence	T_m (°C)
pinalPETase01	MAPGPYSVTTETIPAPSSQPGFTSALLYRPVGSGPFPGVAISPGLGGSSASSYAWLGEHLASHGFVVIAVN TNSPLDTPQTRAQQLLAALDYLKSQPASYRVDTSRLGVLGHSMGGGGVLAAAAADPDLAAVVPLAPWT PSGAPPPLTFPVLMIAGDADTITPPSAVRAFYNSLPSGVDKAYVELKGASHFDFYSDPLPRGYATAWFKR YLDGDTDAAALFEPNGELTTSPRVASYRNYSGPP	61.7
pinalPETase02	MAADPPLASGSYSVSSQSVSSLQAGGGFGGGTIYYPTASGTFGVVAISPGFTGSNSSIAWLGKHLASHGFV VLTINTNSALDSPSTRANQLSAALDYLANYRGGGLSNVTADANRLGVMGHSMGGGGALIAAQQRPDIKA VVPLTPWSTSTNFSNVKAPTLIIGAENDLIAPIALHARPFYNSLPANVPKAYVEIKGATHFDPVLPSGQNV VGKYSTAWLKLYLDGDDAYKALFTGPGCPLRQDPDVARYKTNLPP	NA
pinalPETase03	MTGGGPGDYYYPTGGGPFPAVVFSPGFLATPSSYSWLGTHLASHGFIVAMPNHYPGPNPDYSSAALAAI DELKALNADSSSPLYGKIDTSRIGAGGHSMGGGGALAAAAVLGSRIKAVVAFSPWQGGTSFSGRKAPVLF IVGTNDLIAPPSRVQSLYNTLPSPKAYATIVGGTHFLPNDGSPAGPVAVAWLKWQLKNNSAYKALFTGSN PAIGTDSNWTNCRYKP	90.0
pinalPETase04	MSALLQKYAQPGPYAVTSESLASVGLDGTLYYPTEKGDSFPAVVFSPGFTGSNSAYSWLGEHLASHGIVV LIPQFSSRFDSLATRVAQLQKALDWLKKNAGNIGGIKVDTNRIGVMGHSMGGGAALALAEQDPDIKAVVA FAPWYSPSVDYSALKAPVLLIHGTADTVVPPSNSEKLYEALQPSTAPKALVEIKGADHFAFNDLLGKDHR VINAYTLAWLKLFLKNDKSYLPLLKQPQPGVSRLKTNAALLAL	89.3
pinalPETase05	MSAAAQALLAPGPYAVTTESLSSVGLSGTLYYPASGGPFPAIAFSHGFTGSQSGYAWLGEHLASHGFIVIA PDTETGLFPSHYNRAADLSKAIDYLIKQNATPGDPLYGKVDTSRLGVMGHSMGGGGSLLAASRDSRIKAA VPLTPWQLNTNFSGIKAPVLIIGAENDTIAPIALHSRPFYNSLPANVPKAYVELAGAGHFGFVGSGGPVSEL IKLYTTAWFKLYVKGEEEAAPLLAGEAAKANPDVARYKTNGGLSPPPASS	89.3
pinalPETase06	MAAPPPYLQPGPYSVTSQTVSSSQAPGFGGGTLYYPTASGTFPGVAISPGFTGSQSAYSWLGKHLASHGY VVLAINTNSIFDTPSNRANQLLAALDYLKSGPVSGGVDPSNAKLGVMGHSMGGGGTLIAAQQLGSAVKAV IPLAPAPVSADYSNTKAPVLILHGTNDTVVPISGSDTLYNRLSSPNVDRAYVEIDGAGHFDPTNFGYYSHL VAVPWFDAYLKDQTEAASLFCPTPTLASEPGVASYRTNPTSLASSP	60.5
pinalPETase07	MADDYLAPKYAVTTQSISSLQAPGFGGGTIYRPANLSGKFPVIAFSPGFLASNSNIAWLGEHLASHGYIVIA PNTYSSLDQPSSRANQLIAAIDYLKAQNADPSDPLYGKIDTSRLGVMGHSMGGGGVLVAAQNDSSIKAAV PLTPWNTNTNFSGIKAPTLIIGLQLDLIAPIALHSRPFYNSLSSTVDKAYVEIKNGTHFSPYFPGGYGELSKR YSTAWLDRYLKNDPSAYCDLYGDYAKYNGLLEKYKTNGPPPPPSPCSS	NA
pinalPETase08	MADPSQTGSYAVATQTISSSQAPGFGGGTLYYPTDSGSFGAIAISPGFTGSQSSIAWLGEHLASHGFVVLTI NTNSSTDTPQARANQLSAALDYLKTGAPSAAKSAVDTSRLGVMGHSMGGGGTLIAATQRPDLKAAVPLT PWNTNFNFSGIKTPTLIIGGENDTIAPIALHGKPFYNSLPANLDKAYVELAGATHFPYIQYPALYNRYTTA WLGRYVKNQDSAASYLKTSTDSDVVRYKSNLSNC	60.6
pinalPETase09	MATSDYAAYYQPGGGFEVTSLPLPPDLSPFGAIYYPADLSDAPNGLPVVVFSPGFTATPSLGYDWLGEHL ASHGFVVIFPQAGGPLDNVETRAGQLLDALDLLKELAQGGDPELLALLKLLLGLPPPPDFSKYVDLSRVG VTGHSMGGGGALVAGLSDDRVKTVVALAPWPGPVTEEELAKGGPILLIHGTSDTIVLYSNSSGFYNKLPS GVDKALVTIKGGSHFLFIDPPPPWYPTDVLAPAVAWLDRYLKNDSPAAYAIGGGGGDGLSSPDVTFESK	53.1
pinalPETase10	MATPPLLGYQPKPGPYKVTSEEYTVPGNPNASADVYYPKNPPGPPGEKLPLVILSPGFGGSKSNLAYLGE HLASHGFVVATINATGLTPVGGRVAAIKAVLKQAKDPGSPLYGKVDTNRVAMTGHSMGGGASISLANNP PLGLKAIIPLAPWISSPLPADSKTSVLVIAAENDLVASNSSVDALYNSASSNNDAAYVEIKGAGHFGPYLDP GPARTLSTAWLKYYLKGNSEYKPLFFGKNCLLCTDPDISKYKSNNPK	NA
pinalPETase11	MLPPLCLLDAGLPPPTPPLPGKEFPYEVTSKTFYDPENGVTADVTYPTEPGKTFPLVVFSHGFGSSVASY GGLPEHLASHGFVVVAPDHPDGGGRSPLPPLEALLLAFDPKNWQQRVKQLKAVLDLLENGKLPLPVKVD TSRIGVMGHSMGGGTALALAARDPRVKAVVPLAPWFPRPPPKLKPTLIIHGTNDTVVPPSQVEKLYNKL KGDKALLVVKGGTHFLFLEPNGLDILSLIPPWLDYYLKNDPEYKDYFFGPKGPCCNSPLVEKYKSSNNLSS	NA
pinalPETase12	MASSSTPYSGTYPVGTTTVTSPVSITLYYPSTTSGNNAPIASGAKFPVIVFGPGFNAPLSAYSWLLEYLASR GFIVAATSGTSTSDFLNAIDYLESQNSTPGDPLYGKLNGKVGAMGHSMGGGAALTAASQDSRFKAVATLS PGKPSGTYNGTGPVLILHGENDTVVPPSSNGRPLYNKLKPTYNPTYVEYPGAGHAFFLFPSPATD	87.5
pinalPETase13	MSEEELLASPTGKLFEVKSKTLYYPLENPTKILGTIYYPKNPEKGPFGLIVFLPGFGGSGTNYSWLGEHLA SHGYIVATPNGGGGGSVEENLKQINEAIKLANKDPSLKGYVDTNRLGVMGHSMGGGGAIRLSKANPRIKA VILFAGGASPSDLAYQKNIPVLIIHGKNDTVVPYSESKELYNLLKGDKALLSFPNAGHFDPFEGDNEVYLYL LTLWLDKYLRNDEEAKEKLKNPEELNKKYNLTYKYKKK	NA

Table 10 – *Continued from previous page*

ID	Sequence	T_m (°C)
pinalPETase14	MAALLAKYSQPGPYEVKSLSLSEIGLDGTVFYPTEKGPFPLIAFSHGFTGSASSYAWLLEHLASHGFVVVA PDTETGLNPNHQARADNLLKAIDYVLKLPELPNIDTSRIGVMGHSMGGGSTIIAAADPRVKTIIPLYPVGV QDVFPRLKKITTPVLIIAGSNDTIAPPATNQLPFYEALKGPKALVEIKGATHFFPPGQPNPELLGYTTAWL KAYLKNDPELLEGYKPPEGLTLESKP	67.9
pinalPETase15	MANVGTRTLTVTDPARNRSLTVTIYYPAETAGENAPFAGDPSKPYPVVALAHGFLASPDHGYSYLGEYL ASRGYIVAAPETNTSSASPNHKNRGEDLLLAVKYVLELNSDPSSPFYGKIDESRVGAIGHSMGGGAALLAA RSPLVKAAVLLAPYGKGLGSDFSNVKVPTLFIVGSKDTIAKYPNSEKIYNSLQEPKYLLTIKGANHCQFLD GSSLLCKLLGDGPPTIDDEEQREIISKYVLAWLDQYLKGDSSAKKYLEEGDLDGGGDVKVK	76.9
pinalPETase16	MASPAETYLAPGPYAVKSESLSSYGFNGTLYYPVATGKYPAIAFSHGYTGSPSSVAWWGEYLASHGFVVI APDTPTSPLGDTHEDFADQLSKALDYLAKQNADPSSPLYGKVDTSRLGVMGHSMGGGGSLYAAATDPRI KAAVDLTPGWGGGSFSASNITAPTLVIGGENDEIVPIATHGKRFYNQLPDSVPKAYVELKGVTHITIFAGP NPAVRDLTVAWLKRYLDGETAYAPFFCSDDSSNPDIKKYKSQNCPS	82.9
pinalPETase17	MAPPYGTTPLYPVGTLNYSAPRSFGGPPVTGTLYYPATTAGENAPIATDGPYRVVAFSPGFLGSPSAYR WLGEHLASHGYVVAMPPSGSLGPTQLSAALDYLKAQNADPSSPFYQKVDTSRVGVTGHSMGGGGTYRL ALTNPRVKAAVALSPGNIPGSPPSTPIKVPALVIGGSNDLVAPPSLHGKKFYNRATGYAEYIEIPNGDHFTP LGNGPLALALLLRLFGDDPEAAAEIKANGLKLPPGLVSRYKEKTPPPSS	NA
pinalPETase18	MAKKGPYEVTSESLSSYGLNGTLYYPTEDGPFPAIAFSPGFTGSQSSIAWLGKHLASHGYVVITINTSSILD QPSTRAKELLKAAKLLKKWNADSGSPFYQKIDTSKVAVMGHSMGGGGTLQAAAQGGVDVAVPLAPWNL NLSWAKKVKVPVLIMSGTNDTIAPIALHSDPFYNLLPANVPKAYVEIKGAGHFAFLDGNKYEIQAPYTTA WLKYYLKGDAAAKEALENNTDPDLAKFKVNSNKKEPKK	NA
pinalPETase19	MADPTAPGPYPVGTRDVTVGGGSSGVSGRVYYPATTAGENAPVASGKFPVVAFSHGFLGSQTSSPWLG EHLASRGYIVAAPDTETGLIPNPDTDAPARGQQLSAALDYLAAQNTTSGSPFYGHVDTSRVGVMGHSMG GGGAVQAALADSRFKTVVLLAPYPPPGGGGLSSLTIPVLVISGTNDGVATYSNVQALYDSSASPKRLLTIA GANHCGFLDPGAGPPDPGPGPISPAEQHKVTNPYTTAFFDAYVKGDTTAAAALPTTSPYASTPYVPSVT LRR	NA
pinalPETase20	MALSAAAQYYLTPGPYATTSETDTSLPFPKRVDIYYPTDLASGGKLPVIVFAPGFLASNSSYAWLGEHIAS HGYVVAAPNDPLAQSGSQVRAAVDYLIAQNADPSSPLYGKVDTSRVGATGHSMGGGGTYNAACQPGSR VKAAVPLYPGGSLALNLPCTNPVPTLFIAGGNDLIVPPSWARALYERAPSPKYFLEIPGAGHFFFNLSPDP GLELAKLLTTAWFDRYLKGDADAKTFLENEAKALLGGWIFESKP	NA
pinalPETase21	MASATPPSPLALLGQPGPYAVKSESLASIGLSGTLYYPVAGVEKFPAIVVSPGFGGNAEDWAALLGTQLA SHGFVVVAPDDSPANGSEFKKAIEYLKKQNSNPGSPLYGKVDTSRIGVTGHSMGGGGSLLAAARGPPGLK AIVPLWPGSPGFPPSPNITVPTLIIHGTADEVVPPSKHSDPFYDAVGSPEKALLEIKGGGHFGGPGALLGG EDGVLGQAFVAWFDRYLKGDQEARAALKNLSPENLAKLNAGGNPWTIEVGPPPA	53.4
pinalPETase22	MADPSATGPYAVTSASVSTSSAGLGAATLYYPTASGTFPAVALSHGFGTGATNLAYLARHLASHGFVVLS IGTSTLDGGPARANQLLAAIDALLARNGGGGPGIKVDPSRLGVMGHSMGGGGALYAAALDSRIKAAVPM SPGSGGPASAYNTTAPTLIIVGSNDTVVPPSRGQALYNALPSSVPKAYVELNGATHISIYTSDVRPLAVAW LKRYLDNDAAAAPYFANPSLYGNSPVTTTKSNPPLP	NA
pinalPETase23	MAAVTPPPPPVTPPNTGGIPDYGQPGPYTVQSADLNGGGIKGTVYYPTNGTFPVVILSHGLTGDQTSYA WLAKHLASHGYVVVTPNSGGGGGSGGNDHNGRADDIRTVINYLKSGNPALAGKADTSKVGLMGHSMGG GASLLAAQQDKRFKTVVLLAPWILDSSWFNKPHPPLLIIGASNDTVAPIAQHAKAFYNGVKDKSNVVYLLI TGPPATHFFFFGDSGTGNPAALNAVKGYSTAWFDYYLKGDPAAAAYLLDGGGNNALPGGLVYVVPPP	92.2
pinalPETase24	MAPPGGGTGPYPATSVTGGGATIYYPTELTDKKLPVVVFSPGFTGSPANYAWLGEELASYGYIVVTPNS STLNGGLNGPKALDAVLAANATPGSPLYGKIDLDRIAIAGHSMGGGGALAAAKADPRIKTVIPLAPVNLTP LSDLGDKPVLIIHGDADTVVPYSTNALPLYNAAGPGKYLLTIPGASHFEPGGPGPSPGPYDDVNDPVVPW LDWYLKGDEAAKAKLNTPPNDPPLWTYEYKRGS	NA
pinalPETase25	MAVTSTSGSSFGVSGTIYYPTSVPSGVKAGVIAFSPGFLASPSSYSWLGKHLASRGYVVVAPNVSLIDNHSA RANQLLKAIDALIKLNNDSSSPLYQKVDTSKLGVMGHSMGGGGTLLAAQGNSRIKTVIPLAGGQLDTSPLK NITVPVLIIHGTNDLIVPVSLSDNAYNALPSNTDKAYAERKGADHVDPITNGGGPDYNKVSTAWLDRYLK GNTAAKAKLCGSNSVLNNSGWTNYKSNCCK	73.3

Table 10 – *Continued from previous page*

ID	Sequence	T_m (°C)
pinalPETase26	MADTSTPPVDPSDTSNFYSQLAAPGPYAVTSESLSSYGVNATIYYPTEDPNGKKKFPIIAFGNGFLASNSSYPYLGEHLASHGYVVVGVNSSTPDPSVVQNAIDYVIKQNNDSSSPFYGKIDTSRIGVMGHSMGGGAALRAAADSRIKAVVPLTPWNLGLTDYSSVTQPTLIIGADLDLVAPYSVQQKLYDNLGSPKKAYVTIKGAGHFSIYLPPGPTRGLATAWFDRYLKNDEEAKAYFKPNGLLDQSPYWTNVGSKNLG	61.3
pinalPETase27	MSALLKKYAQPGPYAVKSISSATGGGPGDLYYPADASDVTGTLPVILFSPGFLGSSSGNAWLLEHLASHGFVVIAPDDPGLDGPATRAAQLQAALDYLKSLNLPFKVDTSRVGLMGHSMGGGASLLLAAQGPPVKAIIPIAGWVGSPSPAAKIKTPTLIIGGDADTVVPIAAHGDPFYNNATNTKVVLAEIPGAPHFGGGDWTPAAAYATAWFKYYLDGETDYKEFLTTSPGPAANGGLDIKTNNLN	93.9
pinalPETase28	MAPPPGTGPYPVGSLTVTIPPGSGRSRDLPTTVWYPAATSGENAPIDTSGGPFPLVAFSHGFLLSNSNYSWLLEAIASHGYIVAAPDTVGSDLLDSLDLSADRPQDLQAALDYVKAQNSTPGSPFYGKIDTSRLGLAGHSMGGGGAIQLASQPNNVKAVVLLAPYLLSGDLENAKKVKVPVLLIGASNDTIATPSKVQALYNNLPAPKAFVTIKGANHYQFTDWGGSLDPGGGSISPEEQQEITATYLVAWFDYYLKADSSAKKYDLDDPPPKTPGVSSLEVVDLSS	65.9
pinalPETase29	MSALAAKYAAPGSYAVTSESLSSVGLSGTIYYPTSGGPFPVIAFSHGYTGSASSMAWLGEHLASHGFVVIAPNTQTGLIPNHSAMAADLLKAIDYLLAQNADPSSPLYGKVDTSRLGVSGFSMGGGATLLAAQDPRVKAAVPLWPGLPSGCDTSSPGKPVPTLIIHGTNDPVVPVSIHGEAFYNALPASTDKAYAELAGMGHALPSDPAQLNAIKALTTAWFDRYLKNNPEAAAAFCGGSSCYLCTNPLVSRYKSNCPP	67.1
pinalPETase30	MAGIDDAALLDASGGSGLGGGPYSVTSYTDPSLPGVTIYYPTTASANGYPAVVLVPGASGSGTGYSWLGEMLASNGYVVVAPNGGLDDGTRVAQIRAAIDYLESQNGKSGPLGGKIDTSRIGVTGHSMGGGAALAAAADPRIKAIIPLYGYWLLSSYSNVKAPTLFIHGSNDTVVTLSNSQKAYNSIPATLDKAYLELKGAGHVTPVLTPNPVIAYYGVAWLDRYLKNNTAADLLLKPGGPPPPaGVASYKTNLPS	65.5
*Is*PETase	MQTNPYARGPNPTAASLEASAGPFTVRSFTVSRPSGYGAGTVYYPTNAGGTVGAIAIVPGYTARQSSIKWWGPRLASHGFVVITIDTNSTLDQPSSRSSQQMAALRQVASLNGTSSSPIYGKVDTARMGVMGWSMGGGGSLISAANNPSLKAAAPQAPWDSSTNFSSVTVPTLIFACENDSIAPVNSSALPIYDSMSRNAKQFLEINGGSHSCANSGNSNQALIGKKGVAWMKRFMDNDTRYSTFACENPNSTRVSDFRTANCS	47.0

NA: No significant data detected during thermal stability testing.

Table 11: Amino acid sequences of H-proteins characterized in this study.

ID	Sequence
pinalH01	MLRLRLVLRRRRRPVRRRRRASYSPDSPDKSKFPYSLLKNGPASVLFTKEHEWIAIDKDGLAVIGITEYAQESLGDVVFVELPEVGTTVEAGDELGAVESVKAASDLYMPVDGEILEVNEALSDKPALVNKDPYGKGWLIKAKADEGVDPAKLEGLLSEEDYEKLLEEEE
pinalH02	MLRLLLRLRRRPRSLSRVRLTSRSPNTPDSDVFPFSLLSDGPSKVYFTKEHEWLAVDEDGTAVVGITEYAQEALGDVVFVDLPEVGTKLSAGDELGAVESVKAASDLYMPVSGEVLEVNEALEDSPGLVNEDPYGKGWLIKMKVAEGVDPENVEGLLDEEEYEKLLEEEH
pinalH03	MLSLLRSLLRGFLLLARTRRAYSAADLRFAKTHEWLRLSEDGKTGTVGISDFAQSLLGDVVFVDLPKPDTKLTKGEEFGVVESSKAPASDLYSPVSGEVVEVNEALEDSPSLVNEDPEEAWIFKIKLSNPSELDNLMTEEAYEELCKEEE
pinalH04	MLRRLLLRFLRRRRLSSSSARRFSPDDPDPSKPPYSQFKNGPVAVYFTKEHEWIAVDSDGKAVIGITEFAQEALGDVVFVELPEVGTKLQAGDELGAVESVKAASDIYMPVDGEILEVNEALEDNPSLVNESPYGKGWIVKMKVAEGVDPEKIEGLLSLEEYEKLLKEEE
pinalH05	MLRSFRLRRLRRPLRAARRRRFSSLLSPDKSSFPFSLFKDGPSKVYYTKDHEWIAVDADGTAVVGITKYAQEQLGDVVFVELPEVGDSLKADDSLGAVESVKTASDLYMPVDGEVLEVNEALEDKPELVNKSPYGKGWLIKMKLAEGVDPSKVEGLLSEEEYEKLIKEDE
pinalH06	MLRRLLLLLRGRRLRFLRLRSASASDLKFTKTHEWVQLDDDGKTGTVGISDFAQSLLGDVVFVELPETDTKLKQQETFGTVESVKAPSAELYSPVSGEVVEVNEELSDNPGLVNSSPYEAWIAKIKLSDPSELDKLMTEEEYEKLCKEEV

Table 11 – *Continued from previous page*

ID	Sequence
pinalH07	MAEDMIEIKGCKFPKNLYYDVENHVWVRKEEDGTVTVGITDVAQAELGDIVFVDIKEEGKKVKKGKSIAVVESVKWAEDIPSPVSGEIVEVNEALEENPELVNESPYGEGWIVKIKPSDLEADLKELVSGEEAIDALKEEIEERGIDPEAIVEEHEEKEEE
pinalH08	MSKKITFKTEVDHCWIELEGDVVTVGLSPYAQEQLGDIVFIELTEKETVKAGDTLAVVESVKAASDVYAPVSGEIVERNEEVEDEPELLNSDDEEKNWIVKLTDVDQEELAALPKA
pinalH09	MVEIKGCEFPEDLYYYIEDGEEIWVKKEDDDIVRVGITLYAQSLLGDLAFVRLKKVGKEVKKGKSVAVVESVKWAGPVPAPVSGEIVEVNEEVEKDPELVNKDPYGEGWIVIMKVSDPSELEKLLSGEDAISYLKEELEELGIEPTCH
pinalH10	MLRALLLRLLGLSVSSSQSRSLSAAALKFSETHEWISLEEDGKVGTVGISDYAQELLGDVVFIDLPEEDTKLSAGEEFGVVESVKAPASELYAPVSGTVVEVNEELSESPGLVNENPYEAWIFKVKPSNPSELENLMDEDEYEKLIESE
PinalH11	MLSALFRLRRSFLLFLRRSRTFSLSSLKYAESHEWVKLSEDGKVGTVGLSKFAQELLGDIVFVELPETDTKLKAGKEFGVVESVKAPSAELYSPLSGEVTEVNEDLEESPGLVNSSPYSAWLVKIKPSDPSELEDLMDEEEYEEFCKEEE
pinalH12	MTENKIEIKGCKFPKDLYYDVERHIWVRKEEDGTVTVGITDVAQAELGDIVFVELKKEGKKVKAGKSVAVVESVKAAEPIPSPVSGEIVEVNEALEDNPELVNEDPYGEGWIVKIKPSDLEEDLKKLVTGEEAIEAVKEEIEEEGIDAKALAEELEEEEEG
pinalH13	MVEVKGCEFPEDLYYYVEDGEEVWVKKLDEDTVRVGITDYAQSLLGDLVFIRLKKVGKEVKAGKSIATAESVKWVGPVKAPVSGEVVEVNEEVEEDPELVNKDPYGKGWIVKIKVSDPSELENLKTGEEAIEWLREEIEELGIEPECE
pinalH14	MLRLLRRLLSSLLLTGLARRSFSAANLLFSKSHEWVKLEEDQKVGTVGISEFAQELLGDVVFVDLPEPDTKLKKGQEFGVVESVKAPASDLYSPVSGEVVEVNEELEDSPGLVNSSPYESWIVKIKVSDPSELESLLSQEAYEELCKEEE
pinalH15	MAEVKGVKFPDDRLYYVDEEKHIHVWVKLEGDEAVVGLTDIAQALLGDIVFVELKKVGKEVKAGKPIAVVESVKWAEPVKSPVDGEVVEVNEELEDSPELVNEDPYGEGWIAKVKPADPEAAKKKLLSPEEAAKKIEPLAEELGVGCC
pinalH16	MVEIKGCKVPEDLYYYFKDDEIVWLKKEDDNTYKVGITDYAQALLGDIVFIDLKKVGKEVKAGKSVATVESVKWVGPIKSPVSGEIIEVNEEVEENPSLVNEDPYEKGWIVKVKVSDESELEKLLKGEEAIEYLRKLVEEKGIKGKAEE
pinalH17	MAEIKGCEFPEDLYYIIEGGEEIWVKKLSDDTVRIGITDYAQALLGDLVFVSVKKVGKEVEAGKSIATAESVKWAGPIPAPVTGEIVEVNEEVEENPSLINKDPYGKGWVAKIKVSDPSELENLLSGEKALEFLKEEIEELGIEPECE
pinalH18	MAEIKGCEVPEDLYYYIDEDKIVWLKKEDDNTYLVGITDYAQALLGDIVFIRVKKVGKKVKAGKSAATVESVKWVGPVKSPVSGEIVEVNEEVVENPSLINEDPYGEGWIAKVEVSDESELNNLLSGEEAIEYLKELIEEKGLEVKVVE
pinalH19	MAEIEGCEYPEDLYYSFEGDEEVWVRKESDDTVRVGVTDYAQALLGDLVFVELKKVGKEVKAGKSVAVAESVKWAGPVPAPVSGEVVAVNEEVIENPSLINKDPYGKGWIVKIKVSDPEELEQLKSGEEALEYYKELLEEEGIECKCE
pinalH20	MVEIKGCEIPENLYYYVEGLEEIWIKKIEDDVVTVGITDYAQELLGDLVYVEIKKVGKEVKEGKSLAVAESVKWAGPVPAPVSGEVVEVNEEVVKNPSIVNKDPYGKGWIVKMKISDPSELENLLSGDEAKKYYEEEIEELGIEGKGE