Self-driving laboratories to autonomously navigate the protein fitness landscape

Jacob T. Rapp[1], **Bennett J. Bremer**[1], **Philip A. Romero**[1,2,✉]

[1]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI, USA.

[2]Department of Chemical & Biological Engineering, University of Wisconsin–Madison, Madison, WI, USA.

Abstract

Protein engineering has nearly limitless applications across chemistry, energy and medicine, but creating new proteins with improved or novel functions remains slow, labor-intensive and inefficient. Here we present the Self-driving Autonomous Machines for Protein Landscape Exploration (SAMPLE) platform for fully autonomous protein engineering. SAMPLE is driven by an intelligent agent that learns protein sequence–function relationships, designs new proteins and sends designs to a fully automated robotic system that experimentally tests the designed proteins and provides feedback to improve the agent's understanding of the system. We deploy four SAMPLE agents with the goal of engineering glycoside hydrolase enzymes with enhanced thermal tolerance. Despite showing individual differences in their search behavior, all four agents quickly converge on thermostable enzymes. Self-driving laboratories automate and accelerate the scientific discovery process and hold great potential for the fields of protein engineering and synthetic biology.

Human researchers engineer biological systems through the discovery-driven process of hypothesis generation, designing experiments to test hypotheses, performing these experiments in a wet laboratory, and interpreting the resulting data to refine understanding of the system. This process is iterated to converge on knowledge of biological mechanisms and design new systems with improved properties and behaviors. However, despite notable achievements in biological engineering and synthetic biology, this process remains highly inefficient, repetitive and laborious, requiring multiple cycles of hypothesis generation and testing that can take years to complete.

Robot scientists and self-driving laboratories combine automated learning, reasoning and experimentation to accelerate scientific discovery and design new molecules, materials and systems. Intelligent robotic systems are superior to humans in their ability to learn across disparate data sources and data modalities, make decisions under uncertainty, operate continuously without breaks, and generate highly reproducible data with full metadata tracking and real-time data sharing. Autonomous and semi-autonomous systems have been applied to gene identification in yeast[1–3], new chemical synthesis methodologies[4–6] and the discovery of new photocatalysts[7], photovoltaics[8], adhesive materials[9] and thin-film materials[10]. Self-driving laboratories hold great promise for the fields of protein engineering and synthetic biology[11–13], but these applications are challenging because biological phenotypes are complex and nonlinear, genomic search spaces are high-dimensional, and biological experiments require multiple hands-on processing steps that are error-prone and difficult to automate. There are examples of automated workflows for synthetic biology that require some human input and manual sample processing[14,15], but these are not fully autonomous in their ability to operate without human intervention.

In this Article we introduce the Self-driving Autonomous Machines for Protein Landscape Exploration (SAMPLE) platform to rapidly engineer proteins without human intervention, feedback or subjectivity. SAMPLE is driven by an intelligent agent that learns protein sequence–function relationships from data and designs new proteins to test hypotheses. The agent interacts with the physical world though a fully automated robotic system that experimentally tests the designed proteins by synthesizing genes, expressing proteins and performing biochemical measurements of enzyme activity. Seamless integration between the intelligent agent and experimental automation enables fully autonomous design–test–learn cycles to understand and optimize the sequence–function landscape.

We deployed four independent SAMPLE agents to navigate the glycoside hydrolase landscape and discover enzymes with enhanced thermal tolerance. The agents' optimization trajectories started with exploratory behavior to understand the broad landscape structure and then quickly converged on highly stable enzymes that were at least 12 °C more stable than the initial starting sequences. We observed notable differences in the individual agents' search behavior arising from experimental measurement noise, yet all agents robustly identified thermostable designs while searching less than 2% of the full landscape. SAMPLE agents continually refine their understanding of the landscape through active information acquisition to efficiently discover optimized proteins. SAMPLE is a general-purpose protein engineering platform that can be broadly applied across biological engineering and synthetic biology.

Results

A fully autonomous system for protein engineering

We sought to build a fully autonomous system to mimic the human biological discovery and design process. Human researchers can be viewed as intelligent agents that perform actions in a laboratory environment and receive data as feedback. Through repeated interactions with the laboratory environment, human agents develop an understanding of the system and learn behaviors to achieve an engineering goal. SAMPLE consists of an intelligent agent that autonomously learns, makes decisions and takes actions in a laboratory environment to explore protein sequence–function relationships and engineer proteins (Fig. 1a).

The protein fitness landscape describes the mapping from sequence to function and can be imagined as a terrestrial landscape of peaks, valleys and ridges[16]. The SAMPLE agent aims to identify high-activity fitness peaks (that is, top performing sequences) from an initially unknown sequence–function landscape. The agent actively queries the environment to gather information and construct an internal perception of the landscape. The agent must allocate resources between exploration, to understand the landscape structure, and exploitation, to utilize current landscape knowledge to identify optimal sequence configurations. We pose the agent's protein engineering task as a Bayesian optimization (BO) problem that seeks to optimize an unknown objective function and must efficiently trade off between exploration and exploitation[17,18].

The SAMPLE agent uses a Gaussian process (GP) model to build an understanding of the fitness landscape from limited experimental observations. The model must consider the protein function of interest, in addition to inactive 'holes' in the landscape arising from destabilization of the protein structure[19,20]. We use a multi-output GP that simultaneously models whether a protein sequence is active/inactive and a continuous protein property of interest (Methods). We benchmarked our modeling approach on previously published cytochrome P450 data consisting of 331 inactive sequences and 187 active sequences with thermostability labels[21,22]. The multi-output GP showed excellent predictive ability with an 83% active/inactive classification accuracy and, for the subset of sequences that are active, predicts the thermostability with $r = 0.84$ (Fig. 1b).

The GP model trained on sequence–function data represents the SAMPLE agent's current knowledge, and, from here, the agent must decide which sequences to evaluate next to achieve the protein engineering goal. BO techniques address this problem of sequential decision-making under uncertainty. The upper confidence bound (UCB) algorithm iteratively samples points with the largest upper confidence bound (predictive mean plus prediction interval) and is proven to rapidly converge to the optimal point with high sample efficiency[23,24]. However, naïve implementation of UCB for protein engineering is limited, because the inactive 'holes' in the landscape provide no information to improve the model. We devised two heuristic BO methods that consider the output of the active/inactive GP classifier (P_{active}) to focus sampling toward functional sequences. The 'UCB positive' method only considers the subset of sequences that are predicted to be active by the GP classifier ($P_{active} > 0.5$) and selects the sequence with the top UCB value. The 'Expected UCB' method takes the expected value of the UCB score by multiplying by the GP

classifier P_{active} and selects the sequence with the top expected UCB value. We tested these methods by running 10,000 simulated protein engineering experiments with the cytochrome P450 data (Fig. 1c,d). On average, the UCB positive and Expected UCB methods found thermostable P450s with only 26 measurements and required three- to fourfold fewer samples than the standard UCB and random methods. We also tested the BO methods in a batch setting where multiple sequences are tested in parallel and found a slight benefit to running experiments in smaller batches (Supplementary Fig. 1).

The agent designs proteins and sends them to the SAMPLE laboratory environment to provide experimental feedback (Supplementary Video 1). We developed a highly streamlined, robust and general pipeline for automated gene assembly, cell-free protein expression and biochemical characterization. Our procedure assembles pre-synthesized DNA fragments using Golden Gate cloning[25] to produce a full intact gene and the necessary 5′/3′ untranslated regions for T7-based protein expression. The assembled expression cassette is then amplified via polymerase chain reaction (PCR) and the product is verified using the fluorescent dye EvaGreen to detect double-stranded DNA (Supplementary Fig. 2). The amplified expression cassette is then added directly to T7-based cell-free protein expression reagents to produce the target protein. Finally, the expressed protein is characterized using colorimetric/fluorescent assays to evaluate its biochemical activity and properties (Supplementary Fig. 3).

For this work we focused on glycoside hydrolase enzymes and their tolerance to elevated temperatures. We tested the reproducibility of our automated experimental pipeline on four diverse glycoside hydrolase family 1 (GH1) enzymes from *Streptomyces* species (Fig. 1e). The system reliably measured the thermostability (T_{50}, defined in the section Thermostability assay) of the enzymes with an error less than 1.6 °C. The procedure takes ~1 h for gene assembly, 1 h for PCR, 3 h for protein expression, 3 h to measure thermostability, and 9 h overall to go from a requested protein design to a physical protein sample to a corresponding data point.

We added multiple layers of exception handling and data quality control to further increase the reliability of the SAMPLE platform (Fig. 1f). The system checks whether (1) the gene assembly and PCR has worked by assaying double-stranded DNA with EvaGreen, (2) the enzyme reaction progress curves look as expected, and the activity as a function of temperature can be fit using a sigmoid function, and (3) the observed enzyme activity is above the background hydrolase activity from the cell-free extracts. Failure at any one of these checkpoints will flag the experiment as inconclusive and add the sequence back to the potential experiment queue.

Combinatorial sequence spaces to sample protein landscapes

The SAMPLE platform searches a large and diverse protein sequence space by assembling unique combinations of pre-synthesized DNA fragments. Combinatorial sequence spaces leverage exponential scaling to broadly sample the protein fitness landscape from a limited set of gene fragments. We define a combinatorial sequence space using a DNA assembly graph that specifies which sequence elements can be joined to generate a valid gene sequence (Fig. 2a). We designed a glycoside hydrolase (GH1) combinatorial sequence

space composed of sequence elements from natural GH1 family members, elements designed using Rosetta[26], and elements designed using evolutionary information[27]. The fragments were designed to sample broad sequence diversity and were not intended to target or enhance a particular function (for example, thermostability). All designed sequence fragments are provided in Supplementary Data 5. The full combinatorial sequence space contains 1,352 unique GH1 sequences that differ by 116 mutations on average and by at least 16 mutations (Fig. 2b). The sequences introduce diversity throughout the GH1 TIM barrel fold and sample up to six unique amino acids at each site (Fig. 2c).

Autonomous cloud-based design of glycoside hydrolases

We applied SAMPLE with the goal of navigating and optimizing the GH1 thermostability landscape. We implemented our experimental pipeline on the Strateos Cloud Lab for enhanced scalability and accessibility by other researchers[28]. We deployed four independent SAMPLE agents that were each seeded with the same six natural GH1 sequences. The agents designed sequences according to the Expected UCB criterion, chose three sequences per round, and ran for a total of 20 rounds (Fig. 3a). The four agents' optimization trajectories showed a gradual climb of the landscape, with early phases characterized by exploratory behavior and later rounds consistently sampling thermostable designs. There were two instances where the quality filters missed faulty data and incorrectly assigned a thermostability value to an inactive sequence (Agent 1 in round 10 and Agent 3 in round 5). We intentionally did not correct these erroneous data points to observe how the agents recover from the error as they acquire more landscape information. There were a large number of inconclusive experiments as noted by question marks along the bottom of Fig. 3a. A majority of these were the result of inactive enzymes that the agent must test twice to assign as inactive (Fig. 1f). Approximately 9% of the experiments failed, presumably due to liquid-handing errors.

Each agent discovered GH1 sequences that were at least 12 °C more stable than the six initial natural sequences. The agents identify these sequences while searching less than 2% of the full combinatorial landscape. We visualized the agents' search trajectory and found that each agent broadly explored the sequence space before converging on the same global fitness peak (Fig. 3b). All four agents arrived at similar regions of the landscape, but the top sequence discovered by each agent was unique. The thermostable sequences tended to be composed of the P6F0, P1F2 or P5F2, and P1F3 gene fragments, suggesting the corresponding amino-acid segments may contain stabilizing residues and/or interactions. We believe the agents have identified the global fitness peak of the 1,352-member combinatorial sequence space, because all four agents converged to the same peak, and a GP model trained on all data collected by all agents (the unified landscape model discussed in the following section) predicts top sequences similar to those discovered by the agents.

The agents' search trajectory and landscape ascent varied substantially, despite being seeded with the same six sequences and following identical optimization procedures. Agent 3 found thermostable sequences by round 7, whereas Agent 1 took 17 rounds to identify similarly stable sequences. Agent 2 did not discover any functional sequences until round 8. The divergence in behaviors can be traced to the first decision-making step, where the four

agents designed different sequences to test in round 1. These initial differences arose due to experimental noise in characterizing the six seed sequences, which gave rise to slightly different landscape models that altered each agent's subsequent decisions. The stochastic deviation between agents propagated further over the rounds to produce highly varied landscape searches, but these were ultimately steered back to the same global fitness peak.

SAMPLE agents actively acquire landscape information

SAMPLE agents efficiently and robustly discovered thermostable GH1 enzymes. We analyzed the four agents' internal landscape perception and decision-making behavior to reveal how they navigate the protein fitness landscape. We plotted each agent's model predictions for all 1,352 combinatorial sequences over the course of the optimization (Fig. 4a). The agents' perception of the landscape changed over time, and important events, such as observing new stable sequences or erroneous data points, resulted in large landscape reorganization, as indicated by the crossing lines in Fig. 4a. Many eventual top sequences were ranked near the bottom in early rounds.

To obtain an estimated 'ground truth' landscape, we trained a GP model on all sequence–function data from all agents, which we refer to as the 'unified landscape model' (Supplementary Fig. 4). We analyzed how each agent's landscape perception correlates with the unified landscape model and found agents' understanding became progressively refined and improved as they acquired sequence–function information (Fig. 4b). Notably, most agents discovered thermostable sequences by rounds 11 or 12, when their understanding of the landscape was still incomplete, as indicated by a moderate Pearson correlation of ~0.5. We also analyzed the different agents' degree of agreement on the underlying landscape structure (Fig. 4c). All four agents started with correlated landscape perceptions because they were initialized from the same six sequences, but the landscape consistency quickly dropped, with some agents even displaying negative correlations. The early disagreement arose because each agent pursued a unique search trajectory and thus specialized on different regions of the landscape. The correlation between agents' perceived landscapes eventually increased as more information was acquired. Again, it is notable how the agents tended to discover thermostable sequences by rounds 11–12, while largely disagreeing on the full landscape structure. BO algorithms are efficient because they focus on understanding the fitness peaks, while devoting less effort to regions known to be suboptimal. After round 20, we found the four agents were more confident on the top thermostable sequences and had greater uncertainty associated with lower fitness regions of the landscape (Fig. 4d).

The SAMPLE agents designed sequences according to the expected UCB criterion, which considers the thermostability prediction, the model uncertainty and the probability an enzyme is active (P_{active}). We wanted to understand the interplay of these three factors and how they influenced each agent's decision-making. We looked at the sequences chosen in each round and their percentile rank for thermostability prediction, model uncertainty and P_{active} (Fig. 4e). The agents prioritized the thermostability prediction throughout the optimization, and tended to sample uncertain sequences in early phases, while emphasizing P_{active} in the later phases. Agent 3 prioritized P_{active} earlier than the other agents, which seems to be the result of discovering thermostable sequences early and putting less emphasis

on exploration. We also analyzed the agents' final perception of thermostability, P_{active} and expected UCB, and found the agents specialized on different factors resulting from their past experiences (Fig. 4f). Agent 4's expected UCB is dictated by its large P_{active} range, and Agent 2's is determined by its predicted thermostability. Meanwhile, Agent 3 still has considerable landscape uncertainty, as indicated by the high expected UCB points with moderate thermostability and P_{active} predictions.

Human characterization of machine-designed proteins

The SAMPLE system was given a protein engineering objective, reagents and DNA components, and autonomously proceeded to search the fitness landscape and discover thermostable GH1 enzymes. We experimentally characterized the top sequence discovered by each agent to validate the SAMPLE system's findings using standard human protocols. We expressed the enzymes in *Escherichia coli* and performed lysate-based thermostability assays (Methods). We found that all four machine-designed enzymes were substantially more thermostable than the top natural sequence (Bgl3), and the designs from Agents 1 and 4 were nearly 10 °C more stable (Fig. 5a). The human-measured thermostability values and thermostability differences were not as large as observed using our automated experimental set-up, which is a result of the different protein expression and assay conditions. We also tested the enzymes' Michaelis–Menten kinetic properties and found that all designs displayed similar reaction kinetics with wild-type Bgl3 (Fig. 4b) and the other wild-type input sequences (Supplementary Fig. 5). Our protein engineering search did not explicitly consider reaction kinetics, but it seems that the enzyme catalytic activity was maintained by utilizing an activity-based thermostability assay.

Discussion

Self-driving laboratories automate and accelerate the scientific discovery process and hold great potential to revolutionize the fields of protein engineering and synthetic biology. Automating the biological design process remains challenging due to the scale and complexity of biological fitness landscapes and the specialized operations required for wet laboratory experiments. In this work we have developed the SAMPLE platform for fully autonomous protein engineering. SAMPLE tightly integrates automated learning, decision-making, protein design and experimentation to explore fitness landscapes and discover optimized proteins. We deployed SAMPLE agents with the goal of engineering glycoside hydrolase (GH1) enzymes with enhanced thermal tolerance. The agents efficiently and robustly searched the landscape to identify thermostable enzymes that were at least 12 °C more stable than the initial starting sequences. These gains are larger than achieved in other GH1 thermostability engineering work using Rosetta[29] and high-throughput screening[30].

SAMPLE is a general protein engineering platform that can be broadly applied to diverse protein engineering targets and functions. Although we only demonstrated thermostability engineering, the same general approach could engineer enzyme activity, specificity and even new-to-nature chemical reactions. Like directed evolution, the system does not require prior knowledge of protein structure or mechanism, but instead takes an unbiased approach that examines how sequence changes impact function. The greatest barrier to establishing

SAMPLE for a new protein function is the required biochemical assay. The robotic systems used in this work had access to a microplate reader and thus required a colorimetric or fluorescence-based assay. In principle, more advanced analytical instruments, such as liquid chromatography-mass spectrometry or NMR spectroscopy, could be integrated into automation systems to expand the types of protein functions that could be engineered. Finally, we implemented our full experimental pipeline on the Strateos Cloud Lab to produce a cost-effective and accessible system that can be adopted by other synthetic biology researchers.

SAMPLE has the potential to streamline and accelerate the process of protein engineering. The experimental side of the system is the major throughput bottleneck that limits the overall process. A single round of experimental testing takes 9 h on our Tecan automation system or 10 h split over two days (5 h × 2 days) on the Strateos Cloud Lab. At these rates, with continuous operation, the system could get through 20 design–test–learn cycles in just 1–2 weeks. In practice, the process was much slower due to system downtime, robotic malfunctions and time needed for restocking reagents. Our 20 rounds of GH1 optimization took just under six months, which included a single 2.5-month pause caused by shipping delays. Even this six-month duration compares favorably to human researchers, which we estimate may take 6–12 months to perform similar experiments using standard molecular biology and protein engineering workflows. Learning from previous delays, and with better planning, we estimate that SAMPLE could perform 20 design–test–learn cycles in two months using the Strateos Cloud Lab. We estimate the cost to perform a SAMPLE run of 20 rounds with a batch size of 3 is US$5,200 (US$2,400 for the DNA fragments, US$1,300 for all the reagents and US$1,500 for the Strateos Cloud Lab).

We deployed four identical SAMPLE agents and observed notable differences in their search behavior and landscape optimization efficiency. The agents explored distinct regions of sequence space, specialized on different tasks such as classifying active/inactive enzymes versus predicting thermostability, and Agent 3 discovered thermostable enzymes with ten fewer rounds than Agent 1. The initial divergence in behavior arises from experimental measurement noise, which influences the agents' decisions, which then further propagates differences between agents. There is also an element of luck that is compounded with positive feedback: an agent may happen to search in a particular region and come across improved sequences, which then drives the search upward in favorable directions. These observations have interesting parallels with human researchers, where success or failure could be influenced by seemingly inconsequential experimental outcomes and the resulting decisions. The SAMPLE agents explored distinct regions of the landscape and specialized on unique tasks, which indicates a potential to coordinate multiple agents towards a single protein engineering goal. The decentralized and on-demand nature of cloud laboratory environments would further assist multi-agent coordination systems.

Other research groups have developed automated pipelines and semi-autonomous systems for biological systems engineering. Carbonell and colleagues developed an automated design–build–test–learn pipeline that searches over gene regulatory elements such as promoters and operon configurations to optimize biosynthetic pathway titers[14]. They demonstrated their pipeline by performing two design–build–test–learn cycles to optimize

flavonoid and alkaloid production in *E. coli*. Each step of this pipeline utilized automation, but the entire procedure was not fully integrated to enable autonomous operation. HamediRad and colleagues developed an automated design–build–test–learn system to optimize biosynthetic pathways by searching over promoters and ribosome binding sites[15]. They applied their system to enhance lycopene production in *E. coli* and performed three design–build–test–learn cycles. The most notable difference between SAMPLE and these earlier demonstrations is SAMPLE's high level of autonomy, which allowed us to perform four independent trials of 20 design–test–learn cycles each. High autonomy enables more experimental cycles without the need for slow human intervention.

The protein engineering set-up for this initial SAMPLE demonstration was relatively simple compared to most directed evolution campaigns. First, the search space of 1,352 is small and, for some assays, could be fully evaluated using high/medium-throughput screening. The size the combinatorial sequence space is determined by the number of gene fragments (in our case we used 34) and could be scaled massively using oligonucleotide pools. Even a small pool of 1,000 oligos could be split into 250 fragment options for four segments across a gene and could be assembled into nearly four billion (250^4) unique sequences. Another simple aspect of our SAMPLE demonstration was the thermostability engineering goal. Protein thermostability is fairly well understood and there are already computational tools to predict stabilizing mutations with moderate success. SAMPLE is certainty not restricted to thermostability, and similar classes of machine learning models have been used to model complex protein properties such as enzyme activity[20], substrate specificity, light sensitivity of channelrhodopsins[31], in vivo titer in metabolic pathways[32] and adeno-associated virus capsid viability[33], among others. Our initial work demonstrates a generalizable protein engineering platform whose scope and power will continuously expand with future development.

It was notable that our combinatorial sequence space consisted of natural-sequence, Rosetta-designed and evolution-design fragments, but the top designs were composed purely of natural sequence elements. The agents collectively tested seven designs with Rosetta- or evolution-designed fragments, and only two showed any enzyme activity, with very low thermostability. Our unified landscape model (Supplementary Fig. 4) predicts most of these designed fragments to negatively impact the probability an enzyme is active (P_{active}), thermostability, or both. These fragments probably failed because the designs were too aggressive by introducing many sequence changes. Future work could focus on more conservative designs with two to five mutations per fragment and the latest protein design methods (such as ProteinMPNN[34]).

Our combinatorial sequence space was designed to generate sequence diversity in a function-agnostic manner, but we see great future potential of using more advanced design algorithms to tailor the sequence space toward desired molecular functions. CADENZ is a recent atomistic and machine learning design approach to generate diverse, low-energy enzymes for combinatorial assembly of gene fragments[35] and would readily integrate with SAMPLE's gene assembly procedure. SAMPLE's sequence space design provides an opportunity for humans to propose multiple different molecular hypotheses, which the agent can then systematically explore to refine mechanistic understanding and discover new

molecular behaviors. This human–robot collaboration would combine human intuition and creativity with intelligent autonomous systems' ability to execute experiments, interpret data and efficiently search large hypothesis spaces, leading to rapid progress in molecular design and discovery.

The powerful combination of artificial intelligence and automation is disrupting nearly every industry, from manufacturing and food preparation to pharmaceutical discovery, agriculture and waste management. Self-driving laboratories will revolutionize the fields of biomolecular engineering and synthetic biology by automating highly inefficient, time-consuming and laborious protein engineering campaigns, enabling rapid turnaround and allowing researchers to focus on important downstream applications. Intelligent autonomous systems for scientific discovery will become increasingly powerful with continued advances in deep learning, robotic automation and high-throughput instrumentation.

Methods

Benchmarking BO methods on P450 data

We compiled a cytochrome P450 dataset to benchmark the modeling and BO methods. The dataset consists of 518 data points with binary active/inactive data from ref. 22 and thermostability measurements from ref. 21. We tested the multi-output GP model by performing tenfold cross-validation, where a GP classifier was trained on binary active/inactive data and a GP regression model was trained on thermostability data. The models used a linear Hamming kernel (sklearn[36] DotProduct with sigma_0 = 1) with an additive noise term (sklearn WhiteKernel noise_level = 1). For the test-set predictions, we categorized sequences as either true negative (TN), false negative (FN), false positive (FP) or true positive (TP), and for true positives we calculated the Pearson correlation between predicted thermostability and true thermostability values.

We used the cytochrome P450 data to benchmark the BO methods. The random method randomly selects a sequence from the pool of untested sequences. The UCB method chooses the sequence with the largest upper confidence bound (GP thermostability model mean + 95% prediction interval) from the pool of untested sequences. The UCB method does not have an active/inactive classifier and, if it observes an inactive sequence, it does not update the GP regression model. The UCB positive method incorporates the active/inactive classifier and only considers the subset of sequences that are predicted to be active by the GP classifier ($P_{active} > 0.5$). From this subset of sequences it selects the sequence with the top UCB (GP thermostability model mean + 95% prediction interval) value. The expected UCB method takes the expected value of the UCB score by (1) subtracting the minimum value from all thermostability predictions to set the baseline to zero, (2) adding the 95% prediction interval and (3) multiplying by the active/inactive classifier P_{active}. The sequence with the top expected UCB value is chosen from the pool of untested sequences.

We tested the performance of these four methods by running 10,000 simulated protein engineering trials using the cytochrome P450 data. For each simulated protein engineering trial, the first sequence was chosen randomly, and subsequent experiments were chosen according to the different BO criteria. A trial's performance at a given round is the

maximum observed thermostability from that round and all prior rounds. We averaged each performance profile over the 10,000 simulated trials.

We also developed and tested batch methods that select multiple sequences each round. For the batch methods we use the same UCB variants described above to choose the first sequence in the batch, then we update the GP model assuming the chosen sequence is equal to its predicted mean, and then we select the second sequence according to the specified UCB criteria. We continue to select sequences and update the GP model until the target batch size is met. We assessed how the batch size affects performance by running 10,000 simulated protein engineering trials at different batch sizes and evaluating how many learning cycles were needed to reach 90% of the maximum thermostability.

Glycoside hydrolase combinatorial sequence space design

We designed a combinatorial glycoside hydrolase family 1 (GH1) sequence space composed of sequence elements from natural GH1 family members, elements designed using Rosetta[26], and elements designed using evolutionary information[27]. The combinatorial sequence space mixes and matches these sequence elements to create new sequences. The sequences are assembled using Golden Gate cloning and thus require common four-base-pair overhangs to facilitate assembly between adjacent elements.

We chose six natural sequences by running a BLAST search on Bgl3[37] and selecting five additional sequences that fell within the 70–80% sequence identity range (Supplementary Fig. 3). We aligned these six natural sequences and chose breakpoints using SCHEMA recombination[38,39] with the wild-type Bgl3 crystal structure (PDB 1GNX). The breakpoints for the Rosetta and evolution-designed sequence fragments were chosen to interface with the natural fragments and also introduce new breakpoints to promote further sequence diversity. For the Rosetta fragments, we started with the crystal structure of wild-type Bgl3 (PDB 1GNX), relaxed the structure using FastRelax, and used RosettaDesign to design a sequence segment for a given fragment while leaving the remainder of the sequence and structure as wild-type Bgl3. At each position, we only allowed residues that were observed within the six aligned natural sequences. For the evolution-designed fragments, we used Jackhmmer[40] to build a large family of multiple sequence alignment and designed sequence segments containing the most frequent amino acid from residues that were observed within the six natural sequences. The GH1 family's active site involves a glutamic acid catalytic nucleophile around position 360 and a glutamic acid general acid/base catalyst around position 180. As all fragments were designed based on aligned sequences, these conserved active-site residues will all fall within the same fragment position. The Glu nucleophile is present in blocks P1F3, P2F3, P3F3, P4F3, P5F3, P6F3, PrF6 and PcF6. The Glu general acid/base is present in blocks P1F1, P2F1, P3F1, P4F1, P5F1, P6F1, PrF4, PcF4, PrF5 and PcF5.

We designed DNA constructs to assemble sequences from the combinatorial sequence space using Golden Gate cloning. The designed amino-acid sequence elements were reverse-translated using the Twist codon optimization tool, and the endpoints were fixed to preserve the correct Golden Gate overhangs. We added BsaI sites to both ends to allow restriction

digitation and ordered the 34 gene fragments cloned into the pTwist Amp High Copy vector. Each sequence element's amino acid and gene sequence are given in Supplementary Data 5.

Automated gene assembly, expression and characterization

We implemented our fully automated protein testing pipeline on an in-house Tecan liquid-handling system and the Strateos Cloud Lab. The system was initialized with a plate of the 34 gene fragments (5 ng μl^{-1}), an NEB Golden Gate Assembly Kit (E1601L) diluted to a $2\times$ stock solution, a 2 μM solution of forward and reverse PCR primers, Phusion 2X Master Mix (ThermoFisher F531L), $2\times$ EvaGreen stock solution, Bioneer AccuRapid Cell Free Protein Expression Kit (Bioneer K-7260) Master Mix diluted in water to $0.66\times$, AccuRapid *E. coli* extract with added 40 μM fluorescein, a fluorogenic substrate master mix (139 μM 4-methylumbelliferyl-α-D-glucopyranoside, 0.278% vol/vol dimethylsulfoxide (DMSO), 11 mM phosphate and 56 mM NaCl, pH 7.0) and water.

Golden Gate assembly of DNA fragments.—For a given assembly, 5 μl of each DNA fragment were mixed and 10 μl of the resultant mixture was then combined with 10 μl of $2\times$ Golden Gate Assembly Kit. This reaction mix was heated at 37 °C for 1 h, followed by a 5-min inactivation at 55 °C.

PCR amplification of assembled genes.—A 10-μl volume of the Golden Gate assembly product was combined with 90 μl of the PCR primers stock, and 10 μl of this mixture was then added to 10 μl Phusion 2X Master Mix. PCR was carried out with a 5-min melt at 98 °C, followed by 35 cycles of 56 °C anneal for 30 s, 72 °C extension for 60 s, and 95 °C melt for 30 s. This was followed by one final extension for 5 min at 72 °C.

Verification of PCR amplification.—A 10-μl volume of the PCR product was combined with 90 μl of water, and 50 μl of this mixture was then combined with 50 μl $2\times$ EvaGreen. The fluorescence of the sample was read on a microplate reader (excitation, 485 nm; emission, 535 nm) and the signal was compared to previous positive/negative control PCRs to determine whether PCR amplification was successful.

Cell-free protein expression.—A 30-μl volume of the $10\times$ PCR dilution from the previous step was added to 40 μl of AccuRapid *E. coli* extract and mixed with 80 μl of AccuRapid Master Mix. The protein expression reaction was incubated at 30 °C for 3 h.

Thermostability assay.—We used T_{50} measurements to assess GH1 thermostability. T_{50} is defined as the temperature where 50% of the enzyme is irreversibly inactivated in 10 min and is measured by heating enzyme samples across a range of temperatures, evaluating residual enzyme activity, and fitting a sigmoid function to the temperature profile to obtain the curve midpoint. T_{50} represents the fractional activity lost as a function of temperature and is therefore independent of absolute enzyme concentration and expression level.

A 70-μl volume of the expressed protein was diluted with 600 μl of water, and 70-μl aliquots of this diluted protein were added to a column of a 96-well PCR plate for temperature gradient heating. The plate was heated for 10 min on a gradient thermocycler such that each protein sample experienced a different incubation temperature. After incubation, 10 μl of the

heated sample was added to 90 µl of the fluorogenic substrate master mix and mixed by pipetting. The fluorescein internal standard was analyzed on a microplate reader (excitation, 494 nm; emission, 512 nm) for sample normalization, and the enzyme reaction progress was monitored by analyzing the sample fluorescence (excitation, 372 nm; emission, 445 nm) every 2 min for an hour. Any wells with fluorescein fluorescence less than 20% of the average for a given run were assumed to reflect pipetting failure and were not considered when fitting a thermostability curve.

Human characterization of top designed enzymes

Bacterial protein expression and purification.—The designs were built using Golden Gate cloning to assemble the constituent gene fragments, and the full gene was cloned into the pET-22b expression plasmid. The assemblies were transformed into *E. coli* DH5α cells and the gene sequences were verified using Sanger sequencing. The plasmids were then transformed into *E. coli* BL21(DE3) and preserved as glycerol stocks at −80 °C. The glycerol stocks were used to inoculate an overnight Luria broth (LB) starter culture and the next day this culture was diluted 100× into a 50-ml LB expression culture with 50 µg ml^{-1} carbenicillin. The culture was incubated while shaking at 37 °C until the optical density at 600 nm reached 0.5–0.6 and then induced with 1 mM isopropyl β-D-1-thiogalactopyranoside. The expression cultures were incubated while shaking overnight at 16 °C, and the next day the cultures were collected by centrifugation at 3,600g for 10 min, discarding the supernatant. The cell pellets were resuspended in 5 ml of phosphate-buffered saline and lysed by sonication at 22 W for 20 cycles of 5 s on and 15 s off. The lysates were clarified by centrifugation at 10,000g for 15 min.

The enzymes were purified by loading the clarified lysates onto a Ni-NTA agarose column (Cytiva 17531801), washing with 20 ml of wash buffer (25 mM Tris, 400 mM NaCl, 20 mM imidazole, 10% vol/vol glycerol, pH 7.5) and eluting with 5 ml of elution buffer (25 mM Tris, 400 mM NaCl, 250 mM imidazole, 10% vol/vol glycerol, pH 7.5). The eluted samples were concentrated using an Amicon filter concentrator and concurrently transitioned to storage buffer (25 mM Tris, 100 mM NaCl, 10% vol/vol glycerol, pH 7.5). The final protein concentration was determined using the Bio-Rad protein assay, the samples were diluted to 2 mg ml^{-1} in storage buffer, and frozen at −80 °C.

Thermostability assay.—The clarified cell lysate from the protein expression was diluted 100× in phosphate-buffered saline, then 100 µl of the diluted lysate was arrayed into a 96-well PCR plate and heated for 10 min on a gradient thermocycler from 40 °C to 75 °C. The heated samples were assayed for enzyme activity in quadruplicate with final reaction conditions of 10% heated lysate, 125 µM 4-methylumbelliferyl-β-D-glucopyranoside, 0.125% vol/vol DMSO, 10 mM phosphate buffer pH 7 and 50 mM NaCl. The reaction progress was monitored using a microplate reader analyzing sample fluorescence (excitation, 372 nm; emission, 445 nm) every 2 min for 30 min. The reaction progress curves were fit using linear regression to obtain the reaction rate, and a shifted sigmoid function was fit to the rate as a function of temperature incubation to obtain the T_{50} value.

Michaelis–Menten kinetic assay.—The purified enzymes were assayed in quadruplicate along an eight-point twofold dilution series of 4-methylumbelliferyl-β-D-glucopyranoside starting from 500 μM. The assays were performed with 10 nM enzyme, 0.5% vol/vol DMSO, 10 mM phosphate buffer pH 7 and 50 mM NaCl. The reaction progress was monitored using a microplate reader analyzing the sample fluorescence (excitation, 372 nm; emission, 445 nm) every 2 min for 30 min. A standard curve of 4-methylumbelliferone (4MU) ranging from 3.91 to 62.5 μM was used to determine the assay's linear range. The initial rate for each reaction was determined by fitting a linear function to 4MU fluoresence (excitation, 372 nm; emission, 445 nm) at 0-, 2- and 4-min reaction times. The initial rate data were fit to the Michaelis–Menten equation using the scikit-learn[36] curve_fit function to determine the enzyme k_{cat} and K_M.

SAMPLE code execution

A detailed description of the software loop driving SAMPLE is provided in the Supplementary Information under the heading Detailed description of SAMPLE code functionality.

Materials availability

All plasmids used in this project are available upon request to promero2@wisc.edu.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data availability

A more complete set of data including the code to interpret the data is accessible at https://doi.org/10.5281/zenodo.10048592. Source data are provided with this paper.

References

1. King RD et al. The automation of science. Science 324, 85–89 (2009). [PubMed: 19342587]

2. Coutant A et al. Closed-loop cycles of experiment design, execution and learning accelerate systems biology model development in yeast. Proc. Natl Acad. Sci. USA 116, 18142–18147 (2019). [PubMed: 31420515]

3. King RD et al. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 247–252 (2004). [PubMed: 14724639]

4. Caramelli D et al. Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network. ACS Cent. Sci. 7, 1821–1830 (2021). [PubMed: 34849401]

5. Abolhasani M & Kumacheva E The rise of self-driving labs in chemical and materials sciences. Nat. Synth 2, 483–492 (2023).

6. Volk AA et al. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. Nat. Commun. 14, 1403 (2023). [PubMed: 36918561]

7. Burger B et al. A mobile robotic chemist. Nature 583, 237–241 (2020). [PubMed: 32641813]

8. Langner S et al. Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems. Adv. Mater. 32, 1907801 (2020).

9. Li R et al. A self-driving laboratory designed to accelerate the discovery of adhesive materials. Digit. Discov. 1, 382–389 (2022).

10. MacLeod BP et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci. Adv. 6, eaaz8867 (2020). [PubMed: 32426501]

11. Beal J & Rogers M Levels of autonomy in synthetic biology engineering. Mol. Syst. Biol. 16, e10019 (2020). [PubMed: 33331138]

12. Martin HG et al. Perspectives for self-driving labs in synthetic biology. Curr. Opin. Biotechnol. 79, 102881 (2023). [PubMed: 36603501]

13. Carbonell P, Radivojevic T & García Martín H Opportunities at the intersection of synthetic biology, machine learning and automation. ACS Synth. Biol. 8, 1474–1477 (2019). [PubMed: 31319671]

14. Carbonell P et al. An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. Commun. Biol. 1, 66 (2018). [PubMed: 30271948]

15. HamediRad M et al. Towards a fully automated algorithm driven platform for biosystems design. Nat. Commun. 10, 5150 (2019). [PubMed: 31723141]

16. Romero PA & Arnold FH Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. 10, 866–876 (2009). [PubMed: 19935669]

17. Shahriari B, Swersky K, Wang Z, Adams RP & De Freitas N Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE 104, 148–175 (2016).

18. Hie BL & Yang KK Adaptive machine learning for protein engineering. Curr. Opin. Struct. Biol. 72, 145–152 (2022). [PubMed: 34896756]

19. Thomas N & Colwell LJ Minding the gaps: the importance of navigating holes in protein fitness landscapes. Cell Syst. 12, 1019–1020 (2021). [PubMed: 34793698]

20. Romero PA, Krause A & Arnold FH Navigating the protein fitness landscape with Gaussian processes. Proc. Natl Acad. Sci. USA 110, E193–E201 (2013). [PubMed: 23277561]

21. Li Y et al. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat. Biotechnol. 25, 1051–1056 (2007). [PubMed: 17721510]

22. Otey CR et al. Structure-guided recombination creates an artificial family of cytochromes P450. PLoS Biol. 4, e112 (2006). [PubMed: 16594730]

23. Srinivas N, Krause A, Kakade SM & Seeger M Gaussian process optimization in the bandit setting: no regret and experimental design. In Proc. 27th International Conference on Machine Learning 1015–1022 (ACM, 2010).

24. Auer P. Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res. 3, 397–422 (2002).

25. Engler C, Kandzia R & Marillonnet S A one pot, one step, precision cloning method with high throughput capability. PLoS ONE 3, e3647 (2008). [PubMed: 18985154]

26. Alford RF et al. The Rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theory Comput. 13, 3031–3048 (2017). [PubMed: 28430426]

27. Porebski BT, Buckle AM, By E & Daggett V Consensus protein design. Protein Eng. Des. Sel. 29, 245–251 (2016). [PubMed: 27274091]

28. Arnold C Cloud labs: where robots do the research. Nature 606, 612–613 (2022). [PubMed: 35697877]

29. Carlin DA et al. Thermal stability and kinetic constants for 129 variants of a family 1 glycoside hydrolase reveal that enzyme activity and stability can be separately designed. PLoS ONE 12, e0176255 (2017). [PubMed: 28531185]

30. Romero PA, Tran TM & Abate AR Dissecting enzyme function with microfluidic-based deep mutational scanning. Proc. Natl Acad. Sci. USA 112, 7159–7164 (2015). [PubMed: 26040002]

31. Bedbrook CN et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat. Methods 16, 1176–1184 (2019). [PubMed: 31611694]

32. Greenhalgh JC, Fahlberg SA, Pfleger BF & Romero PA Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. Nat. Commun. 12, 5825 (2021). [PubMed: 34611172]

33. Bryant DH et al. Deep diversification of an AAV capsid protein by machine learning. Nat. Biotechnol. 39, 691–696 (2021). [PubMed: 33574611]

34. Dauparas J et al. Robust deep learning-based protein sequence design using ProteinMPNN. Science 378, 49–56 (2022). [PubMed: 36108050]

35. Lipsh-Sokolik R et al. Combinatorial assembly and design of enzymes. Science 379, 195–201 (2023). [PubMed: 36634164]

36. Pedregosa F et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

37. Perez-Pons JA et al. A β-glucosidase gene (*bgl3*) from *Streptomyces* sp. strain QM-B814. Molecular cloning, nucleotide sequence, purification and characterization of the encoded enzyme, a new member of family 1 glycosyl hydrolases. Eur. J. Biochem. 223, 557–565 (1994). [PubMed: 8055926]

38. Endelman JB, Silberg JJ, Wang Z-G & Arnold FH Site-directed protein recombination as a shortest-path problem. Protein Eng. Des. Sel. 17, 589–594 (2004). [PubMed: 15331774]

39. Voigt CA, Martinez C, Wang Z-G, Mayo SL & Arnold FH Protein building blocks preserved by recombination. Nat. Struct. Biol. 9, 553–558 (2002). [PubMed: 12042875]

40. Wheeler TJ & Eddy SR nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489 (2013). [PubMed: 23842809]
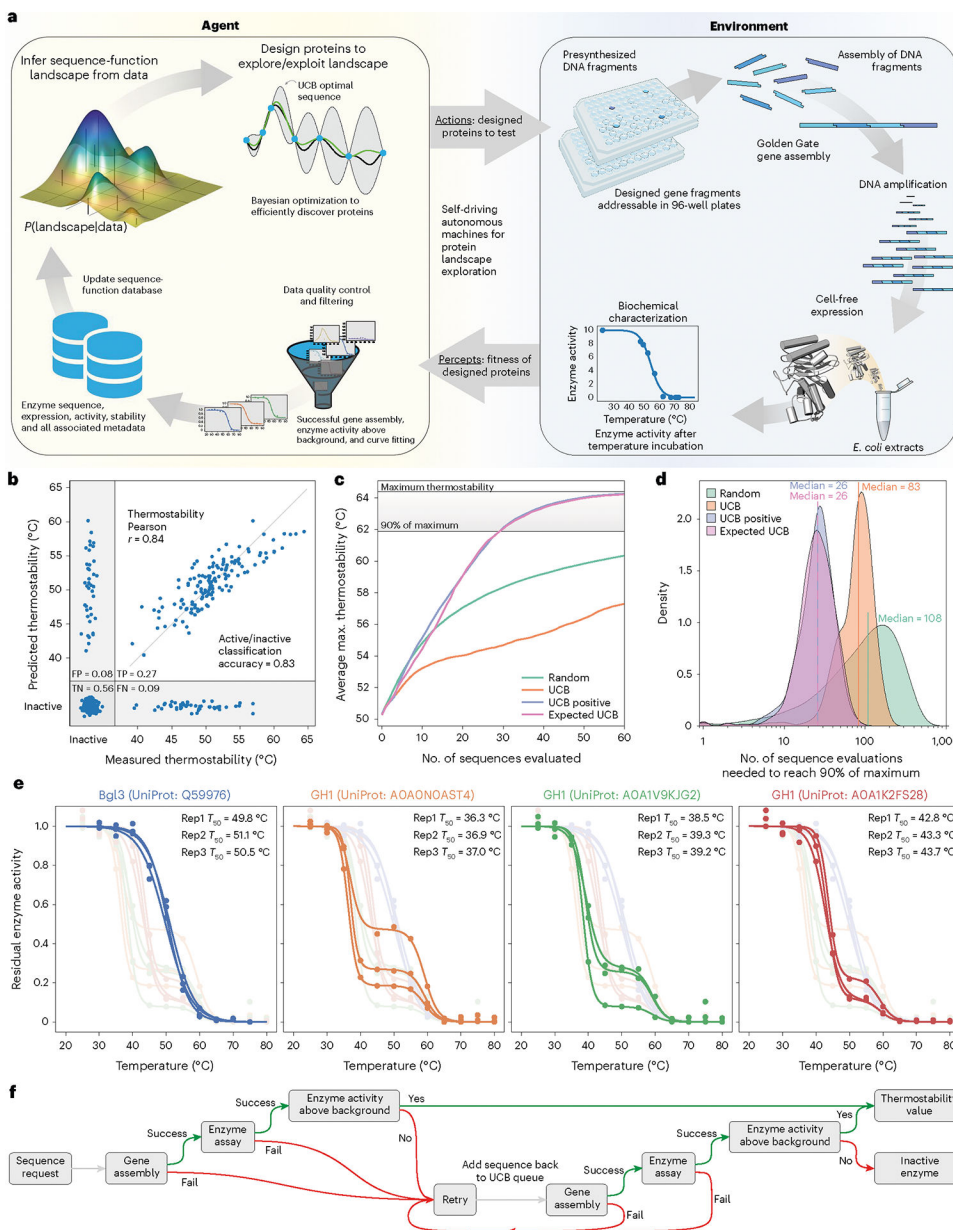
Fig. 1 |. SAMPLE is a fully autonomous system for protein engineering.
a, SAMPLE consists of an intelligent agent that learns sequence–function relationships and designs proteins to test hypotheses. The agent sends designed proteins to a laboratory environment that performs fully automated gene assembly, protein expression and biochemical characterization, and sends the resulting data back to the agent, which refines its understanding of the system and repeats the process.**b**, The multi-output GP model classifies active/inactive P450s with 83% accuracy and predicts P450 thermostability with $r = 0.84$ using tenfold cross-validation. **c**, The performance of four sequential design strategies using P450 sequence–function data. The lines show the maximum observed thermostability for a given number of sequence evaluations, averaged over 10,000 simulated protein engineering trials. **d**, The number of evaluations needed for the design strategies to

discover sequences within 90% of the maximum thermostability (>61.9 °C) using 10,000 simulated protein engineering trials. **e**, The reproducibility of the fully automated gene assembly, protein expression and thermostability characterization pipeline on four diverse GH1 enzymes from *Streptomyces* species. The curves' small shoulder centered around 60 °C is the result of background enzyme activity present in the *E. coli* cell extracts. **f**, The pipeline has multiple layers of exception handling and data quality control for failed experimental steps.
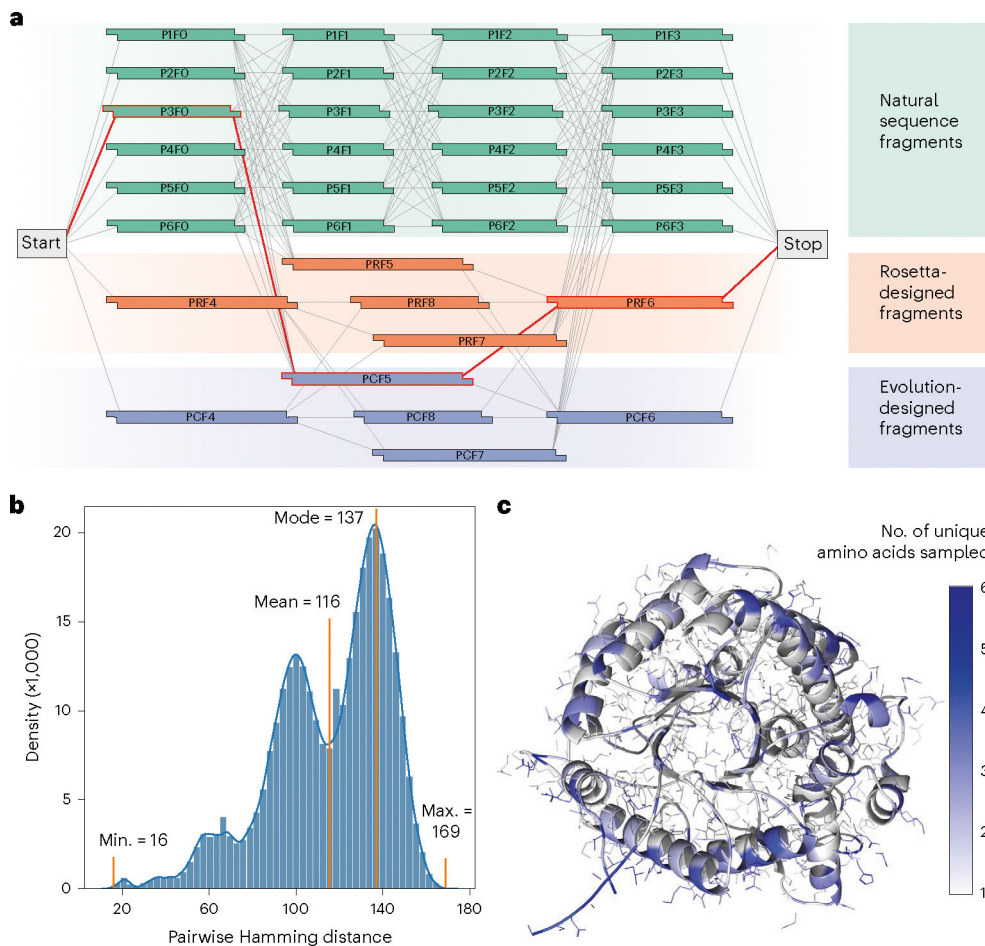
Fig. 2 |. GH1 combinatorial sequence space.
a, A DNA assembly graph defines which sequence elements have compatible overhangs and can be joined to produce a valid gene sequence. Any path from the Start codon to the Stop codon (for example, the red line) is a full gene sequence that can be assembled using Golden Gate cloning. Our GH1 sequence space has a total of 1,352 paths from Start to Stop representing unique protein sequences. **b**, Sequences within the designed GH1 sequence space differ by 116 amino-acid substitutions on average and by at least 16 amino acids. **c**, Sequences within this space sample amino-acid diversity across the protein structure. Sampling diversity is scattered across the protein structure and not focused on a particular domain. The structural illustration is adapted from Protein Data Bank ID 1GNX (β-glucosidase from *Streptomyces* sp).
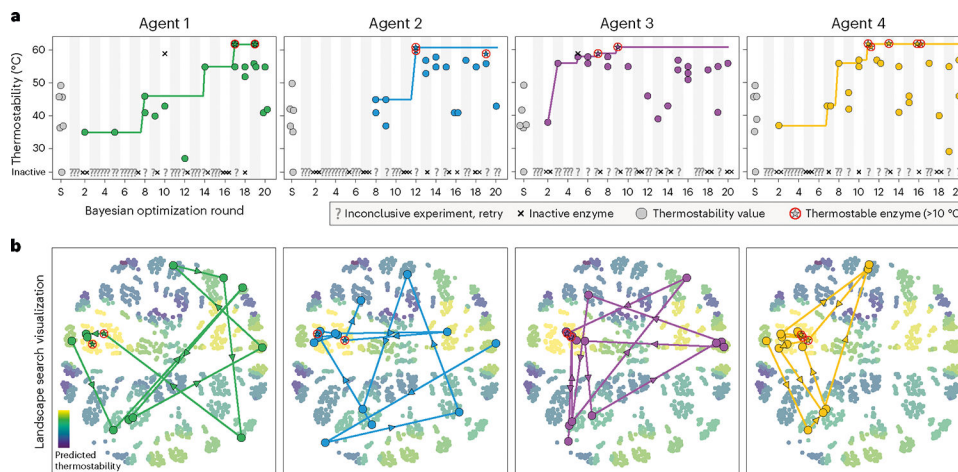
Fig. 3 |. Autonomous exploration of the GH1 landscape.
a, Protein optimization trajectories of four independent SAMPLE agents. Inconclusive experiments, defined as in Fig. 1f, are marked with a '?'. There were two instances of inactive sequences that were incorrectly classified as active enzymes with thermostability values (Agent 1 in round 10 and Agent 3 in round 5). **b**, Visualization of the landscape search. The 1,352 possible sequences were arranged using multidimensional scaling and colored according to their predicted thermostability from the unified landscape model. The center left yellow cluster corresponds to the landscape's fitness peak. The search trajectory is plotted as the most stable sequence from each round.
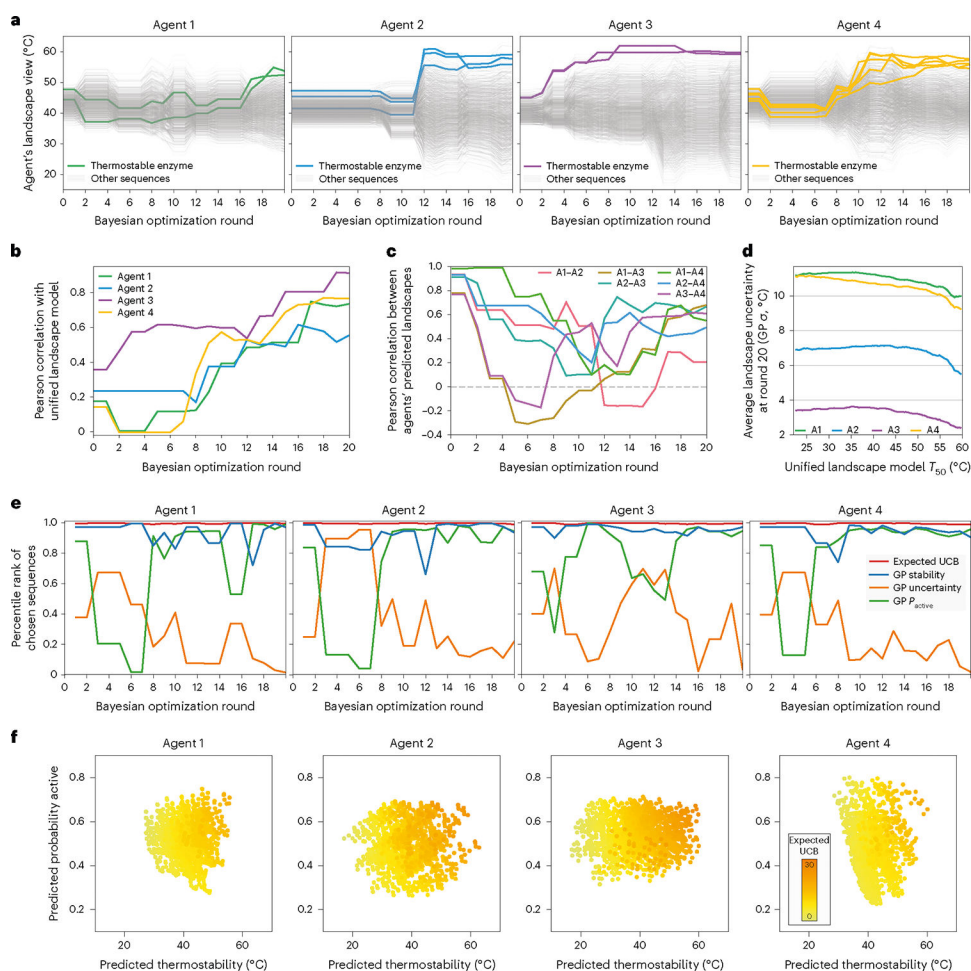
Fig. 4 |. SAMPLE agents' landscape search behavior.
a, Agents' landscape perception over the course of the optimization. The light gray lines show the agents' thermostability predictions for all 1,352 sequences, and the bold colored lines show sequences that were ultimately discovered to be thermostable by each agent. **b**, Pearson correlation between the agents' predicted thermostability landscape and the unified landscape model that incorporates all data retrospectively. **c**, Pearson correlation between different agents' thermostability landscapes over the course of the optimization. **d**, Average model uncertainty as a function of the landscape thermostability. The GP uncertainty (sigma) was averaged over all sequences falling within a 10 °C sliding window across the full T_{50} range predicted by the unified landscape model. **e**, The chosen sequences' percentile ranks for four key factors: the expected upper confidence bound (expected UCB), the thermostability model's mean prediction (GP stability), the thermostability model's predictive uncertainty (GP uncertainty) and the active/inactive classifier's predicted probability a sequence is active (P_{active}). The percentile ranks were averaged over the three sequences in the batch. A percentile rank approaching one indicates the chosen sequences were exceptional for a given factor. **f**, The agents' view of the landscape after the 20 rounds of optimization, with expected UCB overlaid to highlight which factors are contributing to the expected UCB.
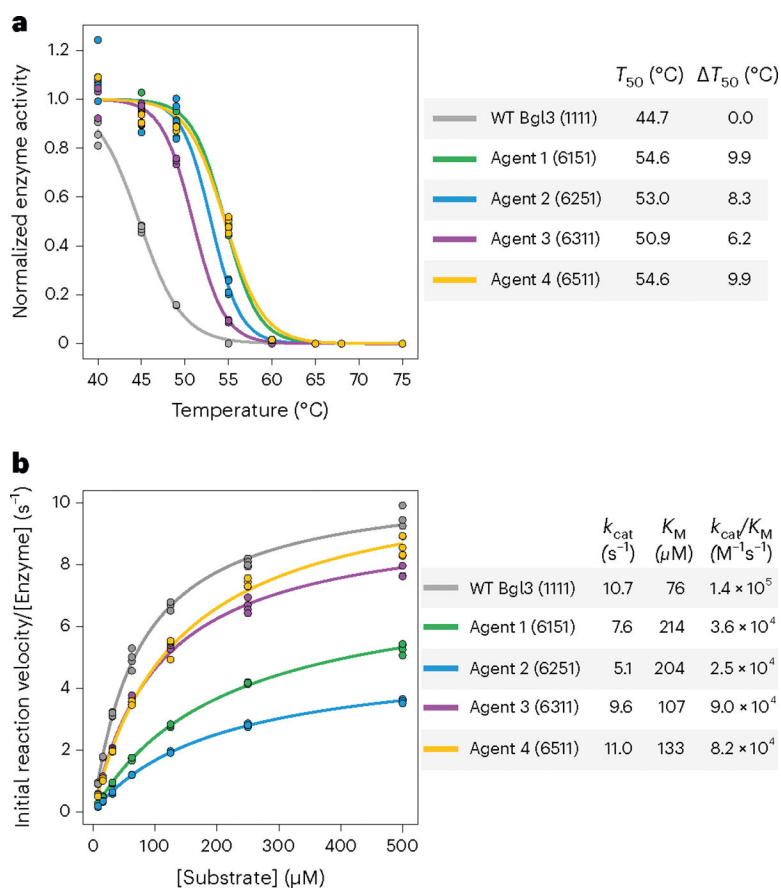
Fig. 5 |. Thermostability and kinetic properties of the designed GH1s.
a, Enzyme inactivation as a function of temperature. Each measurement was performed in quadruplicate, and shifted sigmoid functions were fit to the average over replicates. The T_{50} parameter is the midpoint of the sigmoid function and is defined as the temperature where 50% of the enzyme is irreversibly inactivated in 10 min. The enzyme variant is specified by the sequence of its four constituent fragments, for example 6151 corresponds to P6F0-P1F1-P5F2-P1F3. **b**, Enzyme reaction velocity as a function of substrate concentration. Each measurement was performed in triplicate, and the Michaelis–Menten equation was fit to the average over replicates to determine the kinetic constants. Bgl3 is the most active wild-type input sequence, and the kinetics for the other wild-type sequences are provided in Supplementary Fig. 5