Cell-Free Protein Synthesis as a Method to Rapidly Screen Machine Learning-Generated Protease Variants

Ella Lucille Thornton,* Jeremy T. Boyle, Nadanai Laohakunakorn, and Lynne Regan*

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) tools have revolutionized protein structure prediction, engineering, and design, but the best ML tool is only as good as the training data it learns from. To obtain high-quality structural or functional data, protein purification is typically required, which is both time and resource consuming, especially at the scale required to train ML tools. Here, we showcase cell-free protein synthesis as a straightforward and fast tool for screening and scoring the activity of protein variants in ML workflows. We demonstrate the utility of the system by improving the kinetic qualities of a protease. By rapidly screening just 48 random variants to initially sample the fitness landscape, followed by 32 more targeted variants, we identified several protease variants with improved kinetic properties.

Fast workflow to improve protein function:

KEYWORDS: cell-free protein synthesis, protein design, machine learning, protease, enzyme activity screen
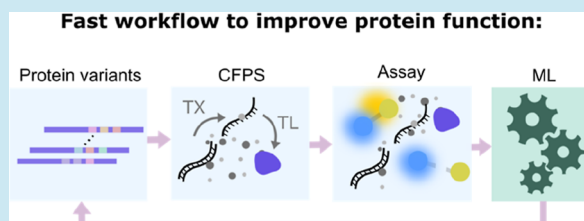
INTRODUCTION

The success of machine learning (ML) methods depends on the quality of the data on which an algorithm is trained. The vast amount of high-quality structural data in the protein database (PDB) underlies the remarkable success of the ML program AlphaFold in predicting protein structure from sequence.[1] To successfully apply ML methods to other areas of protein science, such as the design of function, high-quality data is vital.[2,3]

Protein design space can be imagined as a landscape, populated with many different protein sequences each associated with a different "fitness". The user-defined fitness could be any desirable characteristic such as binding affinity, catalysis rate, yield, or solubility (Figure 1).[4] Identifying the protein with the highest fitness within this landscape can be facilitated by using a ML approach to sample and learn from high-quality data from different sequences.[5] Employing statistical methods embedded in ML workflows helps the user to avoid common "traps" in the navigation of a protein fitness landscape, such as getting stuck at local fitness optima where a variant is better than original, but not the absolute best.[6,7] It enables more efficient sampling of the wider fitness landscape to improve the likelihood of identifying the absolute best variant possible.

Collecting functionality data for protein variants to explore the protein fitness landscape is often achieved by the time-consuming process of in vivo expression, purification, and characterization. But as ML-driven protein mutagenesis projects typically require 10−100's of variants for the model to be effectively trained, a higher-throughput approach is required.[8]

Many different experimental approaches have been previously employed to find tobacco etch virus (TEV) protease variants
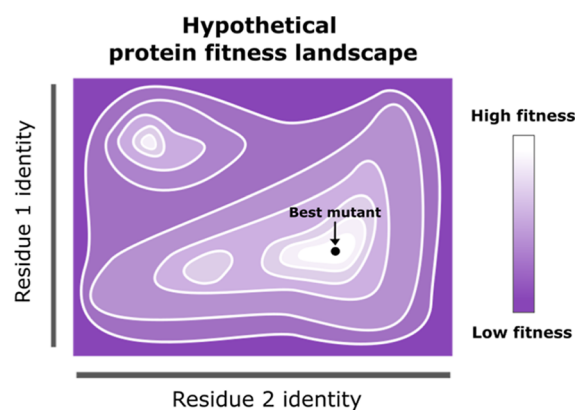
Hypothetical protein fitness landscape

Figure 1. Diagram representing a hypothetical protein fitness landscape, where sequence identity is mapped to fitness. In this hypothetical scenario, two residues (residue 1 and 2) are selected for mutagenesis in the protein and can each be mutated to any other amino acid, represented on the "x and y axes" of this space. Areas shaded with lighter colors represent protein variants with higher "fitness", where fitness can refer to any protein functionality, for example, affinity, catalysis rate, yield or solubility. Labeled with an arrow is the best hypothetical variant.
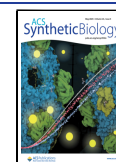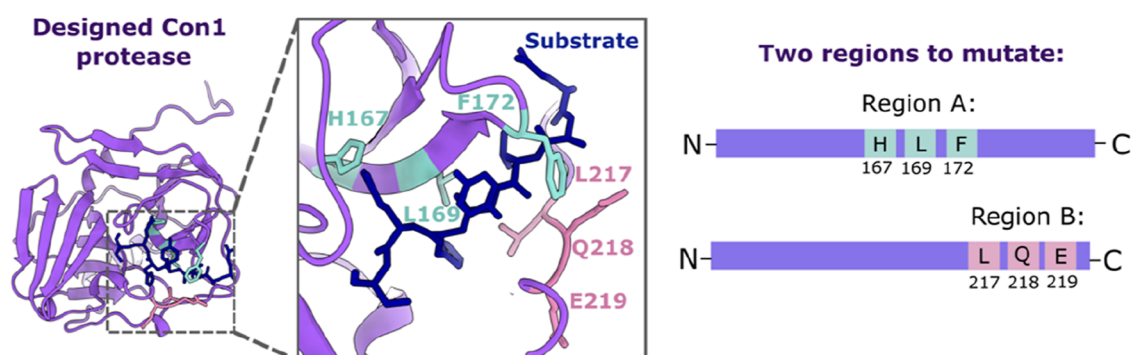
Figure 2. Residues to mutate within the Con1 protease were identified by computational alanine scanning. Six residues were selected and split into two regions to be mutated simultaneously: region A (blue) or B (pink).
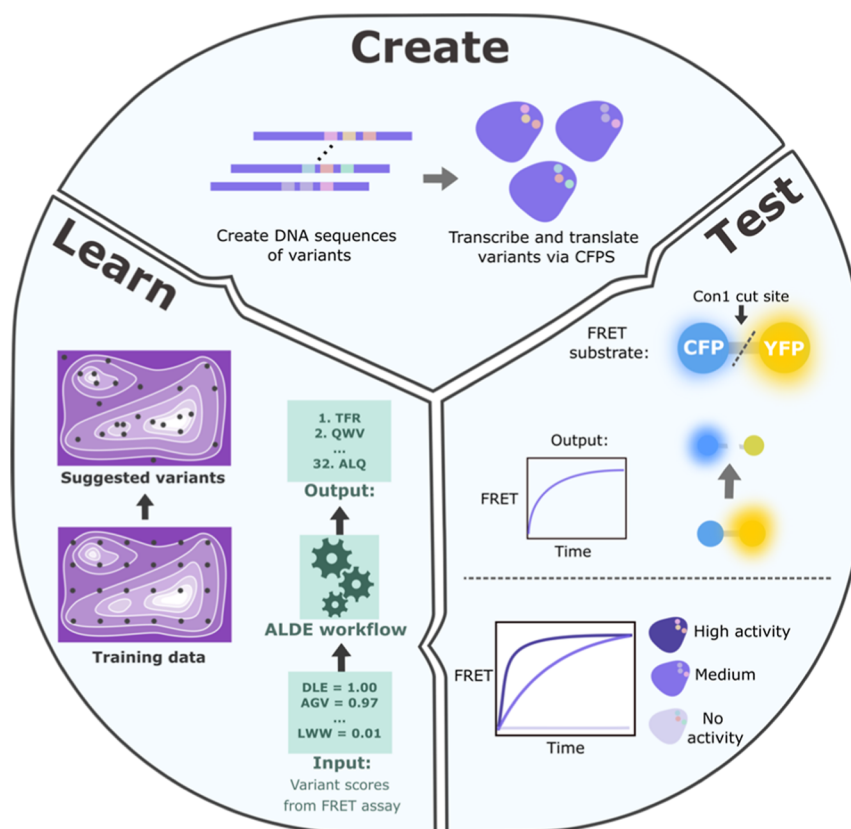

Figure 3. Overview of workflow. The process is split into three phases: (1) create, (2) test, (3) learn. First, 48 random variants were created per region within the protease (A or B, as specified in Figure 2). Linear DNA encoding for these variants was combined with CFPS reagents and incubated at 37 °C to produce the protease variants in vitro. Next, protease variants were tested for activity by addition of FRET substrate and monitoring an increase in FRET over time. The catalytic "speed" of each variant was recorded and translated to a fitness score normalized to WT. This sequence-fitness data was provided to the Active Learning-assisted Directed Evolution' (ALDE) ML workflow[19] as training data. The workflow then provides suggested variants to best explore the fitness landscape. Linear DNA encoding 32 of these suggested variants was then combined with CFPS reagents for in vitro expression, and the cycle of create−test−learn continues.

with improved rates of catalysis. Sanchez and Ting use a directed evolution approach coupled to a high-throughput yeast FACS (Fluorescence-Activated Cell Sorting) workflow to identify uTEV, which has an improved catalytic rate.[9] Another system[10] uses yeast endoplasmic sequestration screening to test variants of TEV produced by error-prone PCR, finding eTEV, which has an improved rate of catalysis derived from an increase in turnover rate. Sumida et al. express and purify de novo TEV designs at a small scale, identifying several designs with improved stability and functionality compared to the starting TEV.[11] These examples demonstrate different approaches to

protease design and screening, and all methodologies are successful in identifying new functional proteases, each with their own distinct advantages. Here, we detail a new protease-screening workflow combining the use of ML and cell-free protein synthesis (CFPS).

CFPS is a powerful tool for protein production and functionality screening.[12−14] By removing the machinery from the cell required for transcription and translation, it can be uncoupled from cellular survival and the scope for applications is significantly widened.[15,16] For example, it enables the production of proteins incompatible with cell survival or the
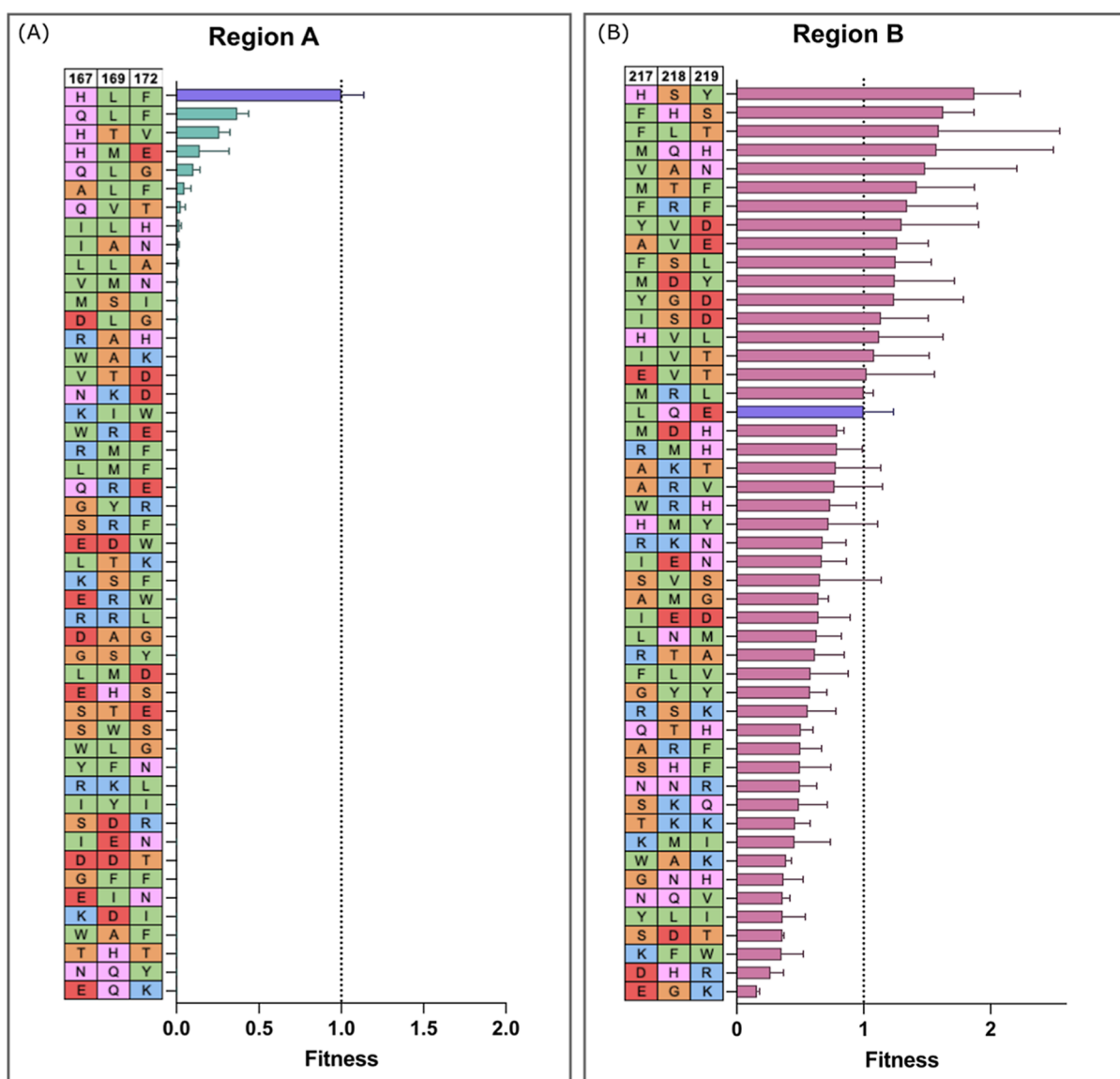
Figure 4. Sparse sampling of mutational landscape for Con1 protease. 48 random variants were tested for both region A and region B. (A,B) Based on the speed of substrate cleavage, variants were scored, and values normalized to the original Con1 (shown in purple) to give a final fitness score for each. The identity of each variant at the residues to be mutated is indicated in the table, with amino acids colored based on type: green = hydrophobic, pink = polar, red = negatively charged, orange = small non-polar, blue = positively charged. Error bars represent S.D. from mean with 3 replicates.

use of substrates impermeable to the cell wall.[17] Because CFPS is an open system, the reaction conditions are more easily programmable compared to living cells, ionic strength, pH, and redox potential can be controlled for production of specific proteins.[18]

We demonstrate CFPS's role in the protein ML workflow by rapidly assessing 100's of protease variants for functionality by combining CFPS with an assay for protease activity. Providing this sequence-fitness data to an ML algorithm[19] to find optimal variants, we identify several active variants of our test protease and improve the overall fitness of the protease by 4-fold. The throughput of the system is demonstrated by the ability to produce and screen an initial round of 48 variants within 6 h.

RESULTS

Here we present the results of combining CFPS with ML to identify variants of a protease with enhanced activity. The protein of interest for this campaign is Con1,[20] a designed

protease created for the removal of tags from proteins after purification. Con1 exhibits high solubility and desirable cleavage specificity in comparison to TEV. Our goal was to identify variants of the Con1 protease that display enhanced fitness, defined in this instance as the increase in initial rate of substrate cleavage at a single fixed substrate concentration.

We first used AlphaFold3[21] to model the Con1 protease with substrate bound (Figure 2). We then used Robetta alanine scan[22] to identify residues predicted to contribute significantly to substrate binding. Six were identified: H167, L169, F172, L217, Q218 and E219, which are predicted to fall on opposite sides of the substrate. We therefore split our mutagenesis scheme to target each of these regions independently, with region A consisting of H167, L169, and F172, and region B consisting of L217, Q218, and E219 (Figure 2).

We created 48 protein variants per region (A or B), where each of the three residues was randomly mutated to any other amino acid simultaneously, rather than one residue at a time
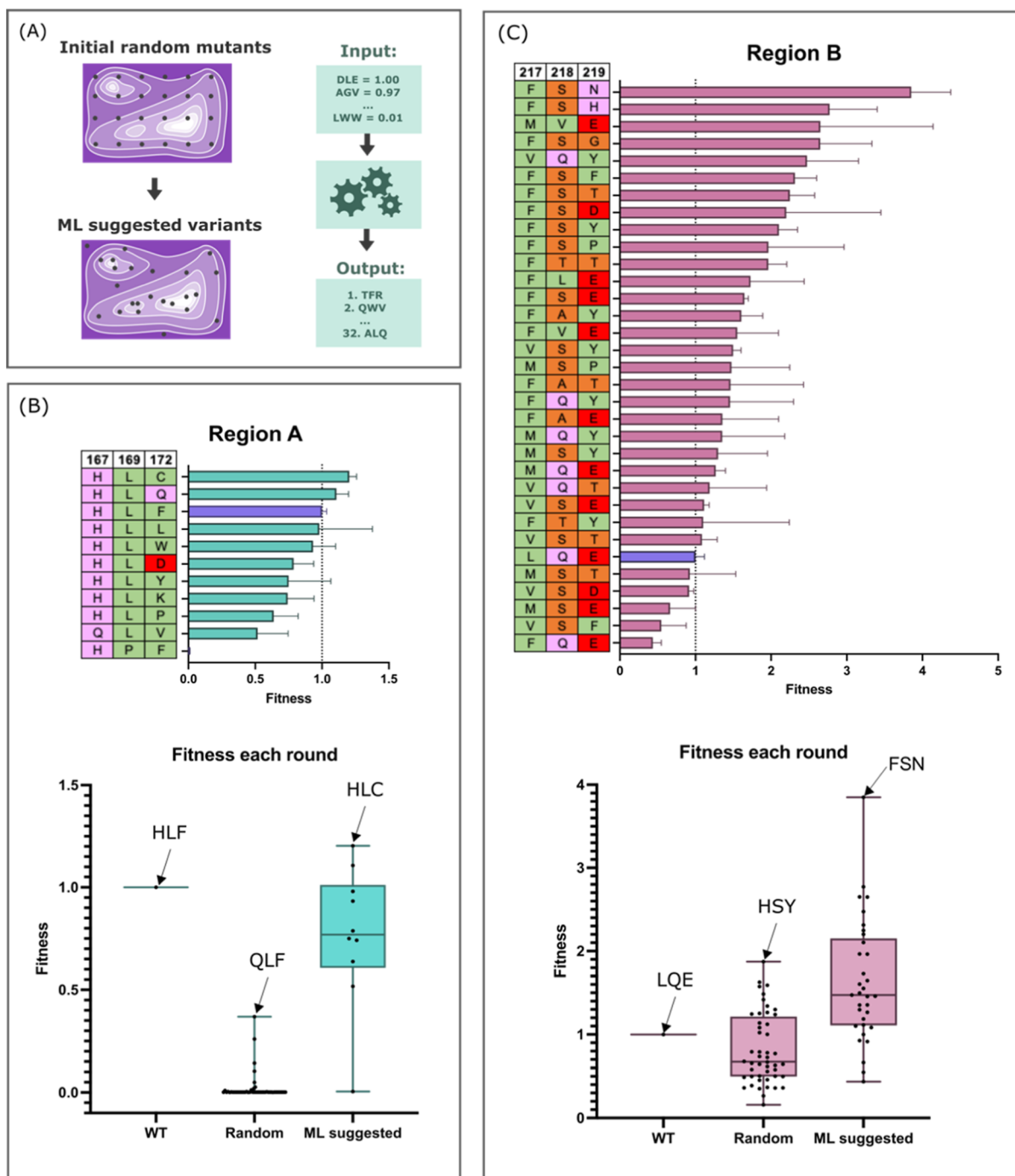
Figure 5. Suggested variants from ML workflow improve fitness of protease successfully. (A) Training data as shown in Figure 4 was input to the ML workflow, which then predicted the best next variants to test in the fitness landscape. The top ML-predicted variants were made via CFPS and screened via FRET substrate, with 10 tested for region A (B) and 32 for region B (C), with fitness shown in ranked order for each mutant with corresponding identity. Error bars represent S.D. from mean with 3 replicates. Overall improvement of variant fitness is shown in each boxplot, where both the mean fitness and maximum fitness improve for both regions. The average fitness for each variant is shown as a single data point, with top variants at each step annotated with their identity.

(Figure 3). DNA templates encoding each of these variants were combined with cell-free reagents (lysate, amino acids, additives), and the proteases were synthesized in vitro.

The activity of each Con1 protease variant was assessed by using a fluorescence resonance energy transfer (FRET)-based assay. The substrate for this assay is a purified fusion protein, in which CFP (donor) and YFP (acceptor) are linked by the specific Con1 protease substrate cut site. By monitoring fluorescence of both the donor and acceptor proteins, a FRET ratio can be calculated and tracked over time, providing a measure of the substrate cleavage rate for each variant (Figure 3).

Sparse Sampling of the Fitness Landscape Provides an Insight into Crucial Residues Required for Function. Screening of the 48 variants for regions A and B revealed striking differences between the tolerance of each region to mutation.
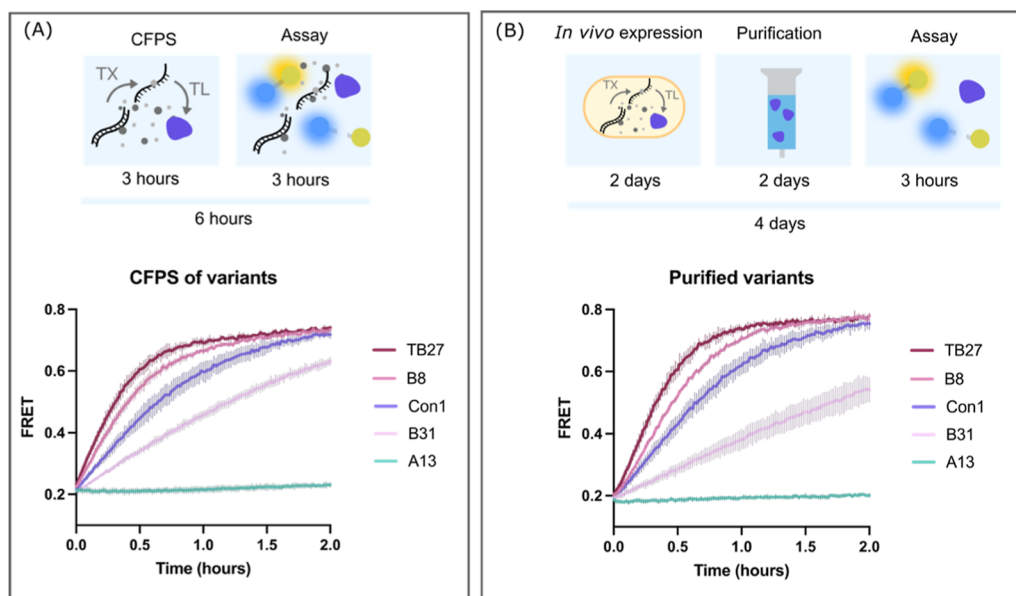
Figure 6. Comparison of protease activity assayed in CFPS with purified protein. (A) Four protease variants with varying apparent activity were selected from the CFPS screen and tested alongside the original Con1 protease. TB27 is the best performing variant from the ML-suggested batch for region B variants, with a sequence identity of FSN at positions 217, 218, 219. B8 is the best performing variant from the initial random variants for region B, with identity HSY at 217, 218, 219. B31 is a low performing variant from the initial random variants for region B, with identity TKK at 217, 218, 219. A13 is an inactive variant from the initial random variants for region A, with an identity V, M, N at 167, 169, 172. Shown here is the FRET output over time for each variant after production in cell-free. The time taken to obtain this data was 6 h total. (B) The same four variants were transformed into BL21 *E. coli* cells and expressed, purified, quantified, normalized to 0.1 μM, and assayed. This took one working week. The output FRET over time is shown, and as in (A), the variants show similar trends of activity. Shown in both line plots is the mean signal from $n = 3$, with S.D. shown as error bars.

Figure 4 shows the variant sequences tested and their activity relative to that of the original Con1 protease.

For region A, mutations are not well tolerated, as seen by the overwhelming number of inactive variants (Figure 4A). Only a few out of the 48 tested show any activity, and these variants are very similar to starting identity, for example, the best variant, A10, is QLF, which is very similar to the original sequence in these positions of HLF. In region A, none of the tested variants were more active than the original Con1.

For region B variants, we observed a much greater range of activity, with all variants exhibiting some activity against the substrate and several exhibiting higher activities than that of the starting Con1 protease against the substrate (Figure 4B). The mutations in regions A and B and their differing range of activities provide two contrasting sets of data to train the ML workflow on.

ML-Suggested Variants Improve Mean and Maximum Fitness. The sequence and corresponding fitness score for each variant served as training data for the 'Active Learning-assisted Directed Evolution' (ALDE) ML workflow.[19] ALDE trains an ensemble of deep neural networks (DNNs) to map the sequence to fitness using the variant data provided. The algorithm then proposes new variants to test by balancing exploitation (variants predicted by the model with high fitness) with exploration (variants for which the model has high uncertainty) (Figure 5A). When applied iteratively, this should lead to the reliable identification of variants with high fitness. Amino acid identity was recorded as a one-hot encoding, with no indication of the amino acid's chemical nature. Each variant's fitness score was specified by the initial rate of substrate cleavage. The normalized fitness scores of these ML-suggested variants alongside their identity are shown in Figure 5B,C.

Because random variants in region A were mostly non-functional, we hypothesized these residues were essential for protease function and intolerant to mutation. Since the likelihood of finding a fast variant seemed low, we only selected 10 ML-predicted variants to synthesize and screen. As shown in Figure 5B, the overall fitness was much improved, with both the mean fitness and maximum fitness improving from the random 48 variants. However, the top variants display fitness scores similar to the original Con1, and looking at the original identity (HLF) compared to the top variant in round 1 (QLF) with round 2 (HLL), there is a strong preference for original-like residues at these positions. Although region A variants did not achieve high fitness scores, the results demonstrate a successful application of ALDE and identification of important residues for protease functionality, where the activity of suggested variants (90%) is far greater than the activity of initial random variants (~20%).

Region B variants were much more fruitful, in terms of improved activity (Figure 5C). The top 32 predicted variants were synthesized and screened, with many faster than original Con1. Both the mean and maximum fitness improved compared to the random variants. The best performing variant (TB27) displayed fitness 4-fold higher than the original Con1 protease.

Protease Activity in CFPS Reflects Activity from Purified and Normalized Proteases. To compare the activity of different enzyme variants fairly, they should be assayed at the same concentration. Because each CFPS reaction was provided with the same concentration of variant DNA, and the variants differed by only a maximum of three amino acids, we did not expect significant differences in expression levels. Nevertheless, it was still important to confirm that observed

changes in protease activity were not due to differences in the protease concentration.

We therefore selected four protease variants with varying levels of activity measured during CFPS screening (Figure 6A). These variants were expressed in *E. coli*, purified, normalized to the same concentration, and assayed using the FRET substrate (Figure 6B). As shown in Figure 6, comparable trends of activity are observed when the variants are screened either directly in CFPS or as purified proteins at equal concentrations, therefore indicating that trends in variant activity in CFPS are reflecting genuine differences in protease activity.

■ DISCUSSION

Here, we present CFPS as a straightforward, fast, and accessible tool for screening enzyme variants. Applied in microplate format as a medium-throughput technique (~10−100's), it can be strategically combined with ML tools to effectively navigate the fitness landscape and improve desired activity. This powerful workflow is demonstrated through the rapid discovery of an enzyme variant 4✕ faster than the starting sequence.

A key benefit of this workflow is the rapid speed at which an advantageous variant can be discovered. ML enables the efficient identification of high fitness variants within a complex fitness landscape by assessing combinations of mutations simultaneously, rather than in a stepwise "greedy" fashion.[7] By combining this powerful analysis method with a fast and efficient data collection approach, CFPS coupled with protease activity screen, the cycle of create−test−learn can rapidly traverse the fitness landscape with minimal lab contact time required. The experimental side of the workflow could be sped up even further by use of a faster commercial CFPS system or by use of robotics to automate the pipetting.[23,24] However, demonstrated here is an approach accessible to most laboratories to set up in-house for a relatively low cost, to find better variants of a protein of interest.

Perhaps the biggest obstacle to overcome in the process of beginning a protein design campaign is identifying an assay that can score your protein function accurately, and to this point, we emphasize the unique benefits of using CFPS compared to cell-based assays: the inherently open system of CFPS allows for addition of a standard concentration of substrate that may be incompatible with living cells. Purification of the product from CFPS for further screening is also possible by chromatography or by surface capture.[25,26]

We show that the activity observed directly from CFPS is comparable to the activity observed with purified proteins. However, not every design campaign may be this straightforward, and normalizing activity to variant expression levels is a desirable additional tool to this workflow. For example, in a recent study,[28] ten enzyme homologs were expressed in CFPS and screened for activity, combined with a strategy to measure expression level to normalize activity levels and optimize expression. We would suggest the adaptation of one of the recently published protein tags for protein quantification in CFPS[27−30] as an addition to the C-terminus of the designed protein to improve accuracy of activity scoring.

Enzyme design is not straightforward because there are many different features that contribute to enzymatic activity, not all of which have been identified for many classes of enzyme.[31] In our design campaign, we note that an exploratory approach to variant creation worked well, revealing dramatic differences between the tolerance of regions A and B for mutagenesis. These differences were not predictable at the computational analysis

stage of the variant design. Our work illustrates how such screening can provide information beyond that which can be obtained from the analysis of sequence conservation. Evolution acts to identify the best sequence for a particular function in the context of a living organism. When a single enzyme is taken out of that context, the optimal sequence for the desired function is not necessarily the same. For example, position 167 of region A is conserved between similar proteases,[20] but positions 169 and 172 are not. Nevertheless, positions 167, 169, and 172 all display the same intolerance to mutation.

We demonstrate through CFPS paired with ML a 4-fold improvement of kinetic properties to our test protease. In total, 138 variants were rapidly screened in CFPS, each protein variant collection taking only 6 h total to produce and screen simultaneously in the same plate. This work outlines an accessible and fast approach to finding higher fitness variants of your protein of interest, and we envision that this workflow can be adapted to many diverse applications of protein design.

■ METHODS

General *E. coli* Methods. *Standard Overnight Cell Growth.* *E. coli* cells were picked from a single colony on an LB agar plate with the appropriate antibiotic into 5 mL of LB with the appropriate antibiotic and grown overnight at 37 °C with shaking. For DNA preparation, TOP10 cells were used (F-*mcrA* Δ(*mrr-hsd*RMS-*mcr*BC) *φ80lacZ*ΔM15 Δ*lacX74 nup*G *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*E15 *gal*K16 *rps*L(StrR) *end*A1 *λ-*). For protein expression, BL21 Gold(DE3) cells were used (F-*omp*T *hsdSB*(*rB−mB−*) *dcm* (TetR) *gal* *λ*(DE3) *endA* The). For CFPS lysate production, Rosetta-gami2 cells were used Δ(ara-leu)7697 ΔlacX74 ΔphoA PvuII phoR araD139 ahpC galE galK rpsL F′[lac+ lacIq pro] gor522::Tn10 trxB pRARE2 (CamR, StrR, TetR). Plasmids were transformed into competent *E. coli* cells following standard protocols.[32]

DNA Purification and Quantification. Plasmids were purified from *E. coli* following protocols described by the manufacturer using the QIAprep Spin Miniprep Kit (Qiagen). Linear DNA was purified from PCR mixture or agarose gels using the Promega Wizard SV Gel and PCR Clean-up System following protocols described by the manufacturer. Purified DNA solutions were quantified by A260 and stored at −20 °C.

DNA Sequencing. Plasmid sequences were verified by DNA sequencing performed either by DNA Sequencing & Services (www.dnaseq.co.uk) or Source Bioscience (www.sourcebioscience.com), using primers provided by the company.

Design and Cloning of Variants. Residues to mutate within the protease were selected by computational protein design tools. First, the amino acid sequence of Con1 and the peptide substrate sequence were provided to AlphaFold3, to create five different predicted structures of the protease interacting with the substrate peptide. These predicted structures were then put through Robetta Alanine scan, an online server developed by the Baker lab. Non-catalytic residues were considered to be mutated if they scored higher than 1 kcal/mol. This information, alongside a visual assessment of the residue locations on the predicted structure of Con1, led to the final plan for variant creation.

The identity of the random protease variants was decided by a python script that assembled 48 random variants with 3 identities per variant, which could be any of the amino acids. Variants for each region were checked for complete amino acid representation (Figure S3). All designs were then created on

Benchling and primers designed to change the codons for 3 amino acids specified in each variant.

All variants were synthesized by AQUA cloning,[33] using the original Con1 protease sequence in pMAL plasmid as the template and mutating the DNA sequence via primer overhangs. All sequences and primers used are listed in the Supporting Information. Designed primer pairs for each Con1 variant were combined with the original Con1 plasmid for PCR by Phusion DNA polymerase, according to manufacturer's instructions. After validation of successful DNA amplification by gel electrophoresis, DNA of each variant was purified from the PCR mixture by a Zymo DNA cleanup kit. Pure DNA solution of each variant was then Dpn1 treated to remove any residual methylated plasmid. 0.5 μL of each treated DNA (150 ng total), 0.5 μL of FD buffer, 0.5 μL of Dpn1, and 3.5 μL of H_2O were combined in PCR tubes and incubated for 15 min at 37 °C, followed by 5 min at 80 °C to inactivate the enzyme. 2.5 μL of this treated mixture was combined with 25 μL of chemically competent TOP10 *E. coli* cells and incubated on ice for 15 min before heat shock at 42 °C for 45 s. After a brief incubation on ice, 200 μL of LB was added to each and incubated at 37 °C for 20 min with shaking. As ampicillin was used as the selective antibiotic, a short recovery of the cells was sufficient for growth. The total transformation mixture was plated on ampicillin (100 μg/mL final concentration) LB agar plates and incubated overnight at 37 °C. Generally, this protocol resulted in 10−40 colonies that displayed 95% successful identity (validated by sequencing).

Cell-Free Protein Synthesis. *Lysate Production.* This protocol is adapted from Kwon and Jewett (2015). Typically, addition of 10 nM linear DNA to our homemade CFPS system yields 1 μM protease in a 5 μL reaction. 2xYTPG (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl, 7 g/L KH_2PO_4, 3 g/L K_2HPO_4, and 18 g/L glucose) was inoculated with 1/200 dilution of overnight cultures of Rosetta-gami2 *E. coli* cells. Cultures were grown for 2 h at 37 °C with shaking, then induced with 0.4 mM IPTG and grown for a further 2 h in the same conditions, before growth arrest by placing on ice. Cells were harvested by centrifugation at 10,000g for 10 min at 4 °C, the supernatant was discarded, and cell pellets were resuspended with 80 mL of buffer A (10 mM tris acetate (pH 8.2), 14 mM magnesium glutamate, 60 mM potassium glutamate) per 400 mL cells harvested. Cells were collected by centrifugation at 4500 rpm for 10 min at 4 °C. The washing and cell harvesting process was repeated twice more, and cell pellets were stored at −70 °C for future downstream processing. Cell pellets were resuspended with 1 mL of buffer A per 1 g of wet cell mass and homogenized by vortexing. 1.5 mL aliquots were sonicated (Fisher120 W sonicator with probe for 0.5−15 mL) with pulses of 10 s on and 10 s off until a total energy output of 556 J was achieved while incubated on ice. Lysate was clarified by centrifugation at 12,000g at 4 °C for 10 min. Supernatant was removed and centrifuged again to remove any residual insoluble material. Clarified supernatant was placed in a clean 1.5 mL tube and incubated at 37 °C for 1.5 h with shaking (220 rpm) in a "run-off" reaction, intended to allow the completion of translation of any mRNA associated with ribosomes. Samples were then centrifuged at 12,000g at 4 °C for 10 min. Supernatant was removed and aliquoted into 25 μL samples, which were stored at −70 °C until they were required for use.

Energy Solution Production. Energy solution was optimized for protease production via a Design of Experiments (DoE) approach. We used a central composite design with axial points to assess the best concentrations of Mg-glutamate, K-glutamate, and PEG-8000. Further details about the DoE process and results are outlined in the Supporting Information (Figure S1). Energy solution was assembled from stock solutions of all constituents, making a solution with a stock concentration 4× the working concentration of each component. In this 4× stock solution are the following components. Amino acid stock solution contained 50 mM each of the following amino acids: alanine, arginine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, and valine. Tyrosine was prepared separately, in an acidic solution (pH ∼5.2) also at a final concentration of 50 mM. Stock batches of energy solution were prepared in volumes of 3 mL with the following recipe: HEPES (pH 8) 200 mM, ATP 6 mM, GTP 6 mM, CTP 3.6 mM, UTP 3.6 mM, tRNA 0.8 mM, CoA 1.04 mM, NAD 1.32 mM, cAMP 3 mM, folinic acid 0.27 mM, spermidine 4 mM, 3-PGA 120 mM, amino acids 6 mM, tyrosine 3 mM, PEG-8000 8%, Mg-glutamate 31 mM, K-glutamate 880 mM, and DTT 1 mM. Aliquots of 25 μL were then stored at −70 °C until required.

Preparation of DNA for Cell-Free Reactions. DNA PCR product was used for CFPS reactions, with expression from a pTac promoter upstream of the protein coding DNA sequence (full sequences available in Supporting Information, Table S1). DNA was purified by DNA Clean & Concentrator Kit (Zymo Research) following described protocols. Final DNA concentration was measured by absorbance at 260 nm on a Nanodrop (DeNovix DS-11), and DNA was stored at −20 °C until required.

Assembly of the Cell-Free Reaction. Cell free reactions were prepared with a final volume of 5 μL in wells of a low volume 384 microplate (Greiner, 788096). Master mix (MM) solutions were prepared for triplicates of each sample, with a final volume of 17 μL. Each MM contained 6.13 μL of lysate, 4.38 μL of energy solution, 4.38 μL of DNA (to give a final working concentration of 10 nM), 1.75 μL of buffer A, and 0.875 μL of GamS nuclease inhibitor protein (NEB, P0774S). The MM was thoroughly mixed before 5 μL was deposited into each well using a multichannel pipette to dispense 12 sample MM at a time. The plate was sealed with an aluminum seal (Thermo, Z721557), which we found drastically reduces evaporation compared with plastic seals and removes the need for wax. The plate was then incubated on a benchtop plate shaker at 37 °C with 1000 rpm shaking for 3.5 h.

FRET Activity Assay. Protease variants were assayed directly in CFPS reagents or in purified form in buffer (100 mM Tris−HCl (pH 8.0), 25 mM NaCl, 5% glycerol) in a low volume 384 microplate (Greiner, 788096). 1 μL of purified FRET substrate was added to each 5 μL solution containing protease to give a final concentration of 5 μM. Substrate was added quickly to all wells using a multichannel pipette to reduce the variability in reaction start time. The plate was sealed with an aluminum seal and quickly placed into the plate reader (POLARstar OMEGA) to measure fluorescence every 1 min at (excitation/emission): (430/480), (430/520), both with gain set to 1000.

Protein Expression and Purification. *Cell Growth and Protein Expression.* Overnight cultures were diluted 100-fold into LB containing the appropriate antibiotic and grown at 37 °C with shaking until OD600 reached 0.6−0.8. Protein expression was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and growth continued for a further 20 h at 20 °C with

shaking. Cells were collected by centrifugation at 6000*g* for 10 min and pellets stored at −20 °C until needed.

Lysis and Clarification. Cells were resuspended in lysis buffer containing cOmplete Protease Inhibitor Cocktail (Sigma-Aldrich) according to the manufacturer's instructions at a ratio of 1:50 volume of buffer to original cell culture volume. Resuspended cells were sonicated (Soniprep 150, MSE) on ice for 30 s, followed by a 30 s rest period. This sonication-rest cycle was repeated until cell lysis was achieved. Clarified cell lysates were prepared by centrifugation at 10,000*g* for 30 min at 4 °C.

Affinity Purification. Proteins were purified via hexahistidine tag using Ni-NTA agarose (Qiagen) according to the manufacturer's instructions. Purification was monitored using SDS-PAGE. Fractions containing protein at approximately 90% or greater purity were pooled and dialyzed against the desired storage buffer. Representative SDS-PAGE gel showing affinity purification of mutants is shown in Figure S8.

Size Exclusion Chromatography. Further purification of proteins by size exclusion chromatography (SEC) with a Superdex-75 or Superdex-200 column was used when affinity chromatography did not provide sufficient purity. Purification by SEC is shown in Figure S9.

Protein Quantification. Protein concentration was determined by measuring absorbance at 280 nm using the extinction coefficient of each protein calculated from the amino acid sequence using the tools available on Benchling [Biology Software].

Protein Concentration. Buffer exchange and simultaneous concentration of protein solutions was performed using Amicon Ultra Centrifugal Filters (Merck).

SDS-PAGE. Protein expression and purification were monitored using SDS-PAGE of samples alongside Precision Plus Protein Dual Xtra Prestained Protein Standards as a molecular weight marker. Protein bands were visualized by staining using InstantBlue Coomassie Protein Stain according to manufacturer's instructions and were subsequently imaged using a Bio-Rad Gel Doc XR+ system.

Data Analysis and Scoring of Activity. FRET ratio was calculated by dividing the blank (water) subtracted fluorescent signal at 480 nm by the blank subtracted fluorescent signal at 520 nm, when solution was excited at 430 nm. This FRET ratio was plotted against time (GraphPad prism). In this project, we looked to increase the initial rate of substrate cleavage. To score the variants for this metric, the FRET data observed over time was truncated at FRET = 0.55, as this provided the best indication of initial rates of cleavage (Figures S4−S6). The data was filtered using a python script. The filtered data were then plotted in GraphPad, and a linear line of best fit to each replicate was made. The slope values for these linear lines indicated the speed of the initial reaction. Each replicate was normalized to the original Con1 protease by division of the average slope value.

ML Implementation. Best candidates for the second round of activity screening were generated by batch Bayesian optimization using the ALDE package.[19] The training dataset consisted of the starting Con1 sequence and associated fitness score (1), plus 48 Con1 variants, randomly mutated at positions *X*, *Y*, and *Z* with associated FRET activity used as a measure of fitness. Amino acids were represented using one-hot encoding. The surrogate model used to evaluate the fitness landscape over all possible amino acid combinations was an ensemble of 5 DNNs with bootstrapping, where each DNN was supplied with random 90% of the available training data and treated as samples from a Bayesian posterior distribution. At each iteration of

approximate Bayesian optimization, the proposal that maximized the Thomson sampling acquisition function from a DNN randomly sampled from the ensemble, which was then used as the input for the next iteration. After completion, the 32 top suggested variants were selected as candidates for a second round of FRET activity measurements. Computation was performed using the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.5c00062.

Full data for Figure 4 dataset A mutants (XLSX)

Full data for Figure 4 dataset B mutants (XLSX)

ML-suggested variant dataset for Figure 5 (XLSX)

Comparison of CFPS and purified variants for Figure 6 (XLSX)

DoE for protease expression; calibration of FRET substrate against known concentrations of protease; identity of all original random mutants; time-course FRET data for 48 B mutants; dataset of FRET data truncated at 0.55; validation of fitness scoring against FRET time-course data; full list of output variants from ALDE; affinity purification SDS-PAGE; size exclusion chromatography SDS-PAGE; purified variants stocked at equimolar concentrations SDS-PAGE; Con1 annotated DNA and protein sequence; FRET substrate annotated DNA and protein sequence; and oligonucleotides used in this study (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ella Lucille Thornton − *Centre for Engineering Biology, Institute of Quantitative Biology, Biochemistry and Biotechnology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, Scotland;* ⓞ orcid.org/0000-0001-7599-0561; Email: ella.thornton@ed.ac.uk

Lynne Regan − *Centre for Engineering Biology, Institute of Quantitative Biology, Biochemistry and Biotechnology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, Scotland;* Email: lynne.regan@ed.ac.uk

Authors

Jeremy T. Boyle − *Centre for Engineering Biology, Institute of Quantitative Biology, Biochemistry and Biotechnology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, Scotland*

Nadanai Laohakunakorn − *Centre for Engineering Biology, Institute of Quantitative Biology, Biochemistry and Biotechnology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, Scotland;* ⓞ orcid.org/0000-0002-6446-482X

Complete contact information is available at:
https://pubs.acs.org/10.1021/acssynbio.5c00062

Author Contributions

E.L.T.: Conceptualization, investigation, analysis, writing—original draft, writing—editing. J.T.B.: Analysis, writing—editing. N.L.: Conceptualization, supervision, writing—editing.

■ REFERENCES

(1) Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(2) Notin, P.; Rollins, N.; Gal, Y.; Sander, C.; Marks, D. Machine learning for functional protein design. *Nat. Biotechnol.* **2024**, *42* (2), 216−228.

(3) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **2020**, *10* (2), 1210−1223.

(4) Freschlin, C. R.; Fahlberg, S. A.; Romero, P. A. *Machine Learning to Navigate Fitness Landscapes for Protein Engineering*; Elsevier Ltd., 2022..

(5) Yang, J.; Li, F. Z.; Arnold, F. H. Opportunities and challenges for machine learning-assisted enzyme engineering. *Am. Chem. Soc.* **2024**, *10*, 226.

(6) Wittmann, B. J.; Yue, Y.; Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **2021**, *12* (11), 1026−1045e7.

(7) Marques, A. D. Machine learning for protein engineering. *Handbook of Machine Learning Applications for Genomics*; *Studies in Big Data*; Springer, 2022, Vol. *103*, pp. 19−29..

(8) Vanella, R.; Kovacevic, G.; Doffini, V.; Fernández De Santaella, J.; Nash, M. A. High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. *Chem. Commun.* **2022**, *58* (15), 2455−2467.

(9) Sanchez, M. I.; Ting, A. Y. Directed evolution improves the catalytic efficiency of TEV protease. *Nat. Methods* **2020**, *17* (2), 167−174.

(10) Denard, C. A.; et al. YESS 2.0, a tunable platform for enzyme evolution, yields highly active TEV protease variants. *ACS Synth. Biol.* **2021**, *10* (1), 63−71.

(11) Sumida, K. H.; et al. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* **2024**, *146* (3), 2054−2061.

(12) Laohakunakorn, N. Cell-free systems: a proving ground for rational biodesign. *Front. Bioeng. Biotechnol.* **2020**, *8*, 1−8.

(13) Tinafar, A.; Jaenes, K.; Pardee, K. Synthetic biology goes cell-free. *BMC Biol.* **2019**, *17* (1), 1−14.

(14) Thornton, E. L.; Paterson, S. M.; Stam, M. J.; Wood, C. W.; Laohakunakorn, N.; Regan, L. *Applications of Cell Free Protein Synthesis in Protein Design*; John Wiley and Sons Inc., 2024..

(15) Karim, A. S.; Jewett, M. C. Cell-free synthetic biology for pathway prototyping. *Methods Enzymol.* **2018**, *608*, 31−57.

(16) Gregorio, N. E.; Levine, M. Z.; Oza, J. P. A user's guide to cell-free protein synthesis. *Methods Protoc.* **2019**, *2*, 24.

(17) De Los Santos, E. L. C.; Meyerowitz, J. T.; Mayo, S. L.; Murray, R. M. Engineering transcriptional regulator effector specificity using computational design and in vitro rapid prototyping: developing a vanillin sensor. *ACS Synth. Biol.* **2016**, *5* (4), 287−295.

(18) Hunt, A. C.; Vögeli, B.; Hassan, A. O.; Guerrero, L.; Kightlinger, W.; Yoesep, D. J.; Krüger, A.; DeWinter, M.; Diamond, M. S.; Karim, A. S.; et al. A rapid cell-free expression and screening platform for antibody discovery. *Nat. Commun.* **2023**, *14* (1), 3897.

(19) Yang, J.; et al. Active learning-assisted directed evolution. *Nat. Commun.* **2025**, *16* (1), 714.

(20) Shams, R.; Regan, L. A new highly specific and soluble protease for precise removal of N-terminal purification tags. *bioRxiv* **2024**.

(21) Abramson, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* **2024**, *630* (8016), 493−500.

(22) Kortemme, T.; Kim, D. E.; Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci. SKTE* **2004**, *2004*, pl2.

(23) Rapp, J. T.; Bremer, B. J.; Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **2024**, *1* (1), 97−107.

(24) Ekas, H. M.; Wang, B.; Silverman, A. D.; Lucks, J. B.; Karim, A. S.; Jewett, M. C. An automated cell-free workflow for transcription factor engineering. *ACS Synth. Biol.* **2024**, *13*, 3389.

(25) Madani, A.; et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **2023**, *41* (8), 1099−1106.

(26) Thornton, E. L.; Paterson, S. M.; Gidden, Z.; Horrocks, M. H.; Laohakunakorn, N.; Regan, L. Self-assembling protein surfaces for in situ capture of cell-free-synthesized proteins. *Front. Bioeng. Biotechnol.* **2022**, *10* (July), 1−9.

(27) Copeland, C. E.; Heitmeier, C. J.; Doan, K. D.; Lee, S. C.; Porche, K. B.; Kwon, Y.-C. Expanding the cell-free reporter protein toolbox by employing a split mneongreen system to reduce protein synthesis workload. *ACS Synth. Biol.* **2024**, *13* (6), 1663−1668.

(28) Yuan, Q.; Wu, M.; Liao, Y.; Liang, S.; Lu, Y.; Lin, Y. Rapid prototyping enzyme homologs to improve titer of nicotinamide mononucleotide using a strategy combining cell-free protein synthesis with split GFP. *Biotechnol. Bioeng.* **2023**, *120* (4), 1133−1146.

(29) Pham, T. D.; Poletti, C.; Tientcheu, T. M. N.; Cuccioloni, M.; Spurio, R.; Fabbretti, A.; Milon, P.; Giuliodori, A. M. FAST, a method based on split-GFP for the detection in solution of proteins synthesized in cell-free expression systems. *Sci. Rep.* **2024**, *14* (1), 8042.

(30) Lévrier, A.; et al. Split reporters facilitate monitoring of gene expression and peptide production in linear cell-free transcription-translation systems. *ACS Synth. Biol.* **2024**, *13*, 3119.

(31) Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **2010**, *19*, 1817−1819.

(32) Green, M.; Sambrook, J. *Molecular Cloning: A Laboratory Manual*, 4th ed.; Cold Spring Harbor Press, 2012.

(33) Beyer, H. M.; Gonschorek, P.; Samodelov, S. L.; Meier, M.; Weber, W.; Zurbriggen, M. D. AQUA cloning: a versatile and simple enzyme-free cloning approach. *PLoS One* **2015**, *10* (9), No. e0137652.