	Dataset name	Dataset size used	Purpose for model training
(a)	Swiss-prot	400k protein sequences, 1000 unique UniprotKB keywords	High-fidelity sequences + metadata for training
(b)	TrEMBL	180M protein sequences, 20 unique UniprotKB keywords	Higher-quantity, lower-fidelity proteins for training
(c)	UniParc	280M protein sequences	Reference database used for full-range of proteins exposed to model
(d)	NCBI Taxonomy	100k unique taxonomy terms	Provide source organism information to the model
(e)	Pfam	56k protein sequences	Curated protein families for lysozymes
(f)	Uniref30	7k protein sequences	Chorismate mutase sequences used from an HHBlits search
(g)	NCBI nr database	13k protein sequences	Chorismate mutase sequences used from a blastp search
(h)	Interpro	17k protein sequences	Malate dehydrogenase proteins under IPR001557

Table S1: A list of publicly available datasets used to train and fine-tune ProGen. The training datasets (a-c) contain a total of 280 million unique proteins which are associated with 101,100 tags based on keywords related to biological processes and molecular function and taxonomic information (d). The fine-tuning datasets (e-h) are used to further improve ProGen's ability to generate protein sequences in local sequence neighborhoods of lysozymes, chorismate mutase, and malate dehydrogenase proteins.

Template: $\langle c_1 \rangle \langle c_2 \rangle \ldots \langle c_N \rangle\ a_1\ a_2\ a_3\ a_4\ a_5 \ldots$

Training Sample

```
<Metazoa><Chordata><Mammalia><Rodentia><Muridae><Rattus><Rattus><Norvegicus>
<NAD>
<Translocase>
<Iron><Iron-sulfur><2Fe-2S>
<Mitochondrion><Mitochondrion inner membrane>
<Transport><Electron transport><Respiratory chain>
MFSLALRARASGLTAQWGRHARNLHKTAVQNGAGGALFVHRDTPENNPDTPFDFTPENYERIEAIVRNYPEGHRAA
AVLPVLDLAQRQNGWLPISAMNKVAEVLQVPPMRVYEVATFYTMYNRKPVGKYHIQVCTTTPCMLRDSDSILETLQ
RKLGIKVGETTPDKLFTLIEVECLGACVNAPMVQINDDYYEDLTPKDIEEIIDELRAGKVPKPGPRSGRFCCEPAG
GLTSLTEPPKGPGFGVQAGL
```

Fine-tuning Sample

```
<Phage Lysozyme>
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFNQDVDAAV
RGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVI
TTFRTGTWDAYKNL
```

Figure S1: Sample sequences with associated control tags provided for model training during training and fine-tuning. The amino acid sequence (a_1, a_2, …), prepended with desired control tags ($< c_1 >$,$< c_2 >$, . . .), are formulated as tokens for ProGen to autoregressively compute the probability of the next token. Control tags are constructed to provide information regarding molecular function, biological process, cellular component, or curated protein family.

Pfam name (ID)	Description	Number of sequences in fine-tuning set	Average sequence length in fine-tuning set	Number of artificial sequences tested	Average sequence length of artificial sequences tested
Phage lysozyme (PF00959)	Glycoside hydrolase family 24 which includes lambda phage lysozyme and Escherichia coli endolysin	16488	151 (±29)	20	170 (±16)
Glyco_hydro_108 (PF05838)	Glycoside hydrolase family 108	5857	152 (±38)	45	179 (±6)
Glucosaminidase (PF01832)	Glycoside hydrolase family 74	23238	138 (±14)	8	139 (±4)
Transglycosylase (PF06737)	Resuscitation-promoting factor proteins	9824	84 (±7)	8	93 (±16)
Pesticin (PF16754)	Hydrolase enzyme secreted by Yersinia pestis and other Gammaproteobacteria to kill related bacteria occupying ecological niche.	541	167 (±28)	19	176 (±17)

Table S2: The five lysozyme families utilized for *de novo* protein sequence generation, along with details of the fine-tuning dataset. We utilize a total of 55,948 sequences from these five families obtained from Pfam, ranging from 541 to 23238 proteins per family. The proteins within each family exhibit a range of average sequence lengths, from 84 to 167 residues (one standard deviation shown in table), and ProGen is able to generate similarly diverse sequence average length statistics, from 93 to 179 residues.

	Diversity-50% [number of clusters]	Diversity-80% [number of clusters]
Natural lysozyme	1021	4115
Artificial lysozyme	1193	6206
Natural chorismate mutase	831	1324
Natural malate dehydrogenase	343	1231

Table S3: Sequence diversity of natural lysozymes, ProGen-artificial lysozymes, natural chorismate mutase, and natural malate dehydrogenase proteins. The sequence statistics indicate that lysozymes are more sequence-diverse than the other two protein systems. Also, ProGen is able to generate artificial sequences that exhibit higher diversity than its corresponding natural lysozyme library. Each sequence database is clustered by mmseqs2 with 50%/80% max identity with 80% coverage. The value of the diversity metric is the number of clusters with three or more members. A higher value represents a more diverse library.

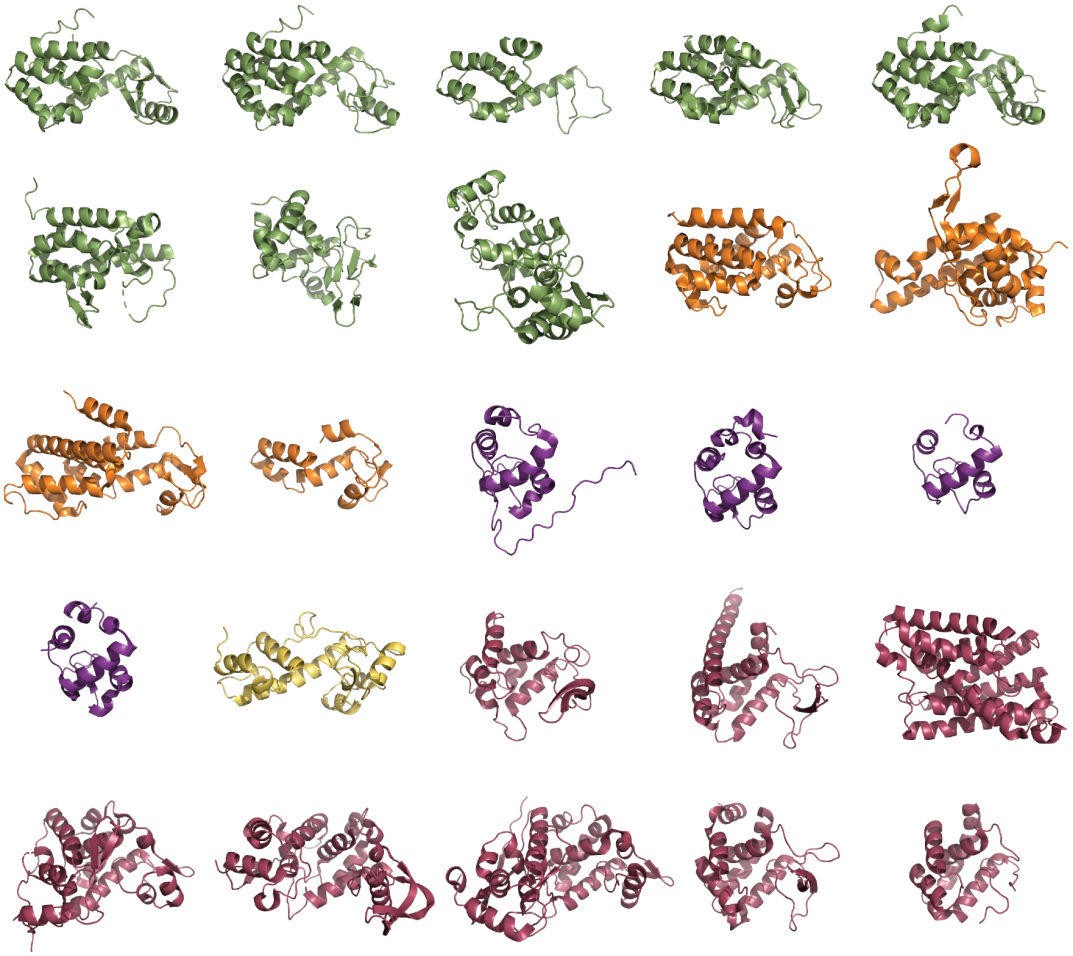


Figure S2: Crystal structures of known natural proteins chosen to exemplify the inherent diversity within the five lysozyme families selected in this study, where Phage lysozyme, (PF00959), Glyco_hydro_108 (PF05838), Transglycosylas (PF06737), Pesticin (PF16754), and Glucosaminidase (PF01832) families (with Pfam IDs) are represented as green, orange, purple, yellow, red respectively. The five families exhibit considerable structural diversity and multiple structural folds, presenting a challenging design space for a sequence-only model.

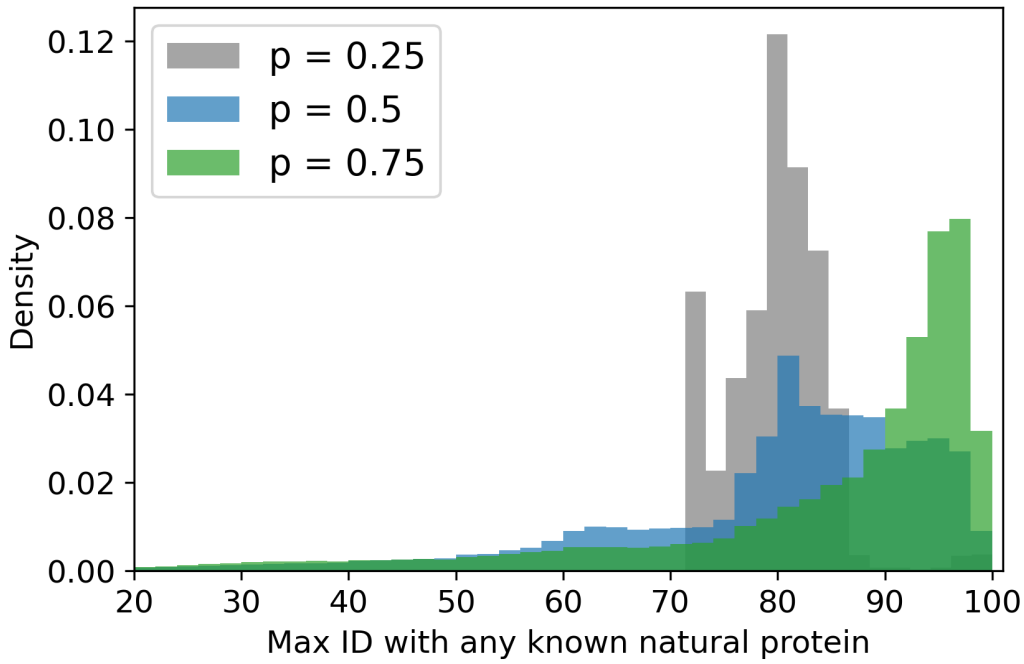


Figure S3: A density plot of max identity distribution of one million lysozyme sequences generated by ProGen at different top-p values, which is a hyperparameter used for controlling the diversity of generated sequences. A lower p setting allows the model to generate sequences that have a higher likelihood under the model but lower diversity, meaning that the mean max identity of the generated sequences may be closer to 100, but the spread is smaller. In contrast, a higher p setting results in a more diverse distribution, but at the cost of more potential mistakes. In our experiments, we generate a total of one million sequences at three different top-p settings and select a hundred from this pool for expression and characterization.

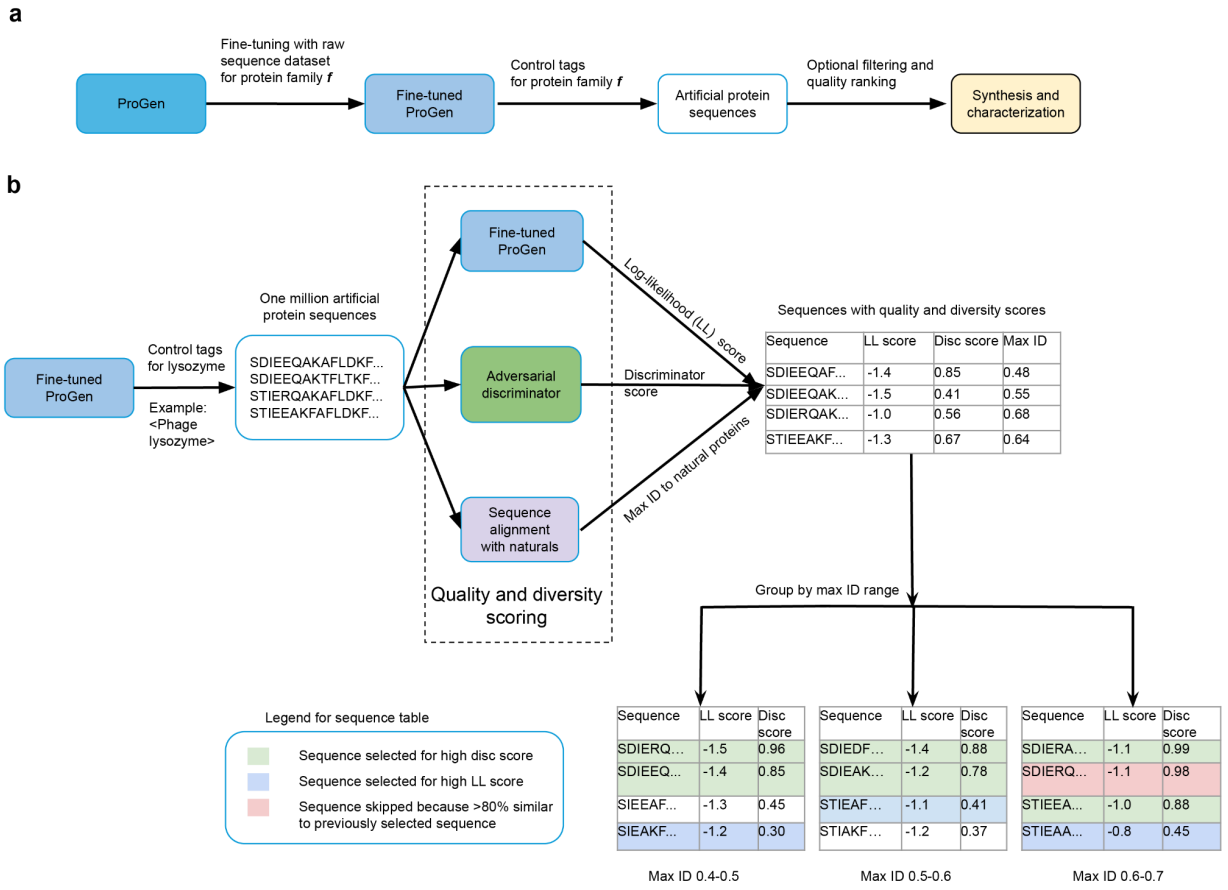


Figure S4: Generating and selecting proteins for synthesis. (a) A high-level overview of the artificial protein generation process for generic protein family ***f*** using ProGen. ProGen is fine-tuned to a raw sequence dataset of proteins within family ***f***. The resulting model conditionally can generate artificial protein sequences by conditioning on the control tag for family ***f.*** The resulting sequences can be optionally filtered using quality ranking criteria, and top ranked sequences are synthesized and characterized in a laboratory. (b) Overview of pipeline used for lysozyme generation and synthesis. ProGen was fine-tuned to a dataset of known lysozymes across the five target families. Fine-tuned ProGen was then used to generate one million artificial lysozymes conditioning on control tags for the five families. These million sequences were ranked for quality using log-likelihood scores (i.e., model likelihoods) from fine-tuned ProGen, and the probability that an adversarial discriminator predicted that the sequences were real. Since we could only test a limited number of proteins experimentally, we further shortlisted sequences across a range of sequence diversities. For this purpose, the maximum sequence identity to known natural sequences was measured for all artificial sequences. The artificial sequences were then binned into five groups of max ID range, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8 (not shown), and 0.8-0.9 (not shown). The top ranking sequences by discriminator score and log-likelihood score in each bin were selected for synthesis. Sequences were selected one by one in order of rank, and sequences that were more than 80% similar to a previously selected sequence were skipped to ensure diversity.

de novo contacts	V5-W162, V5-E166, E7-C101
de novo non-contacts	M1-L83, M1-T87, M1-I94, M1-L95, M1-D96, M1-E97, M1-R100, M1-C101, M1-V115, M1-R123, M1-L125, M1-Q126, M1-R129, M1-V135, M1-A138, M1-Q139, M1-K151, M1-S155, M1-T156, M1-F157, M1-K163, M1-E166, M1-L168, A2-G81, A2-L83, A2-N85, A2-T87, A2-I94, A2-D96, A2-E97, A2-V98, A2-R100, A2-C101, A2-V112, A2-A113, A2-V115, A2-A116, A2-N119, A2-R123, A2-M124, A2-L125, A2-Q126, A2-E127, A2-K128, A2-W130, A2-D131, A2-A134, A2-V135, A2-A138, A2-Q139, A2-K151, A2-S155, A2-T156, A2-F157, A2-W162, A2-K163, A2-E166, A2-N167, A2-L168, K3-L83, K3-T87, K3-I94, K3-D96, K3-C101, K3-R123, K3-M124, K3-Q126, K3-R129, K3-V135, K3-A138, K3-Q139, K3-S155, K3-K163, K3-E166, K3-L168, V4-I94, V4-C101, V4-N167, V4-L168, V5-I94, V5-A113, V5-M124, V5-E127, V5-K163, V5-N167, E7-K76, E7-I79, E7-Q80, E7-G81, E7-N85, E7-T87, E7-I94, E7-D96, E7-E97, E7-V98, E7-R100, E7-V112, E7-A113, E7-V115, E7-A116, E7-N119, E7-M124, E7-Q126, E7-E127, E7-K128, E7-R129, E7-W130, E7-A138, E7-Q139, E7-S155, E7-T156, E7-K158, E7-W162, E7-K163, E7-E166, E7-N167, E7-L168, V13-G73, V13-A78, V13-I79, V13-Q80, V13-G81, V13-T87, V13-I94, V13-D96, V13-R100, V13-C101, V13-A113, V13-V115, V13-Q126, V13-R129, V13-V135, V13-A138, V13-Q139, V13-R144, V13-S155, V13-T156, V13-K158, V13-K163, V13-E166, V13-L168, K18-I94, V21-C101, Q23-T87, Q23-T156, H25-G73, H25-I94, H25-C101, H25-F157, H25-K163, H25-N167, L26-I94, W29-N119, I40-K76, I40-I79, I40-Q80, I40-G81, I40-D96, I40-E97, I40-V98, I40-R100, I40-C101, I40-N119, I40-M124, I40-Q126, I40-D131, I40-V135, I40-L137, I40-Q139, I40-R144, I40-T156, I40-K158, I40-W162, I40-K163, I40-E166, I40-N167, I40-L168, K41-C101, K41-V135, K41-T156, K41-W162, D42-I94, D42-C101, A44-I94, K45-E72, K45-G73, K45-L83, K45-I94, K45-E97, K45-V98, K45-C101, K45-A113, K45-M124, K45-Q126, K45-D131, K45-A138, K45-Q139, K45-K151, K45-S155, K45-K158, K45-W162, K45-E166, K45-N167, I47-C101, Q48-G73, Q48-K76, Q48-A78, Q48-I79, Q48-Q80, Q48-G81, Q48-L83, Q48-I94, Q48-R100, Q48-C101, Q48-A113, Q48-R123, Q48-M124, Q48-Q126, Q48-E127, Q48-K128, Q48-R129, Q48-E132, Q48-V135, Q48-L137, Q48-A138, Q48-Q139, Q48-T156, Q48-K158, Q48-T159, Q48-W162, Q48-K163, Q48-E166, Q48-N167, Q48-L168, I49-I94, I49-C101, I49-M124, I49-V135, I49-E166, I49-N167, N52-L83, N52-I94, N52-D96, N52-C101, N52-A113, N52-N119, N52-M124, N52-Q126, N52-D131, N52-E132, N52-A134, N52-Q139, N52-K151, N52-S155, N52-W162, N52-N167, N52-L168, L54-I94, L54-C101, L54-V135, L54-K163, L54-E166, K57-I94, N59-I94, V61-T87, V61-I94, K65-T87,

	K65-I94, K65-A113, K65-D131, K65-A138, R68-T87, R68-I94, R68-V115, Q69-T87, Q69-I94, I70-C101, G73-N119, G73-Q126, K76-I94, Q80-T87, Q80-I94, Q80-N167, L83-I94, S84-C101, T87-N119, T87-R129, T87-K151, S89-E166, S89-N167, I91-C101, I91-A113, I91-E166, I91-N167, I91-L168, D93-I94, I94-C101, I94-N119, I94-M124, I94-E127, I94-R129, I94-E132, I94-V135, I94-S155, I94-E166, I94-N167, I94-L168, C101-N119, N119-V135, N119-T156, N119-E166

Table S4. Artificial protein L056 *de novo* pairwise interactions. After aligning L056 with a multiple sequence alignment of natural lysozymes, we report any pairwise interaction that was not present in the natural library. As we have determined the structure of artificial L056, we can further designate the *de novo* pairwise interactions that are residue contacts in the crystal structure.

Group	Number of "*de novo*" positions	Number of artificial sequences	Number of natural sequences in the MSA
Functional artificial proteins with characterized kcat/kM	120	8	21062
PF16754: Functional artificial sequences as defined by activity assay	289	10	578
PF06737: Functional artificial sequences as defined by activity assay	22	3	13466
PF00959: Functional artificial sequences as defined by activity assay	272	20	21062
PF05838: Functional artificial sequences as defined by activity assay	2821	34	7005
PF01832: Functional artificial sequences as defined by activity assay	17	5	33529

Table S5. *De novo* positions in functional artificial proteins. We evaluate if amino acids occupy *de novo* positions in our artificial proteins that are unseen in nature. We generate MSAs for each protein family, tabulate how often each amino acid at each position of the artificially-designed protein appears in the MSA, and calculate the number of amino acids that occupy positions not observed in the MSA. There are numerous *de novo* positions in ProGen-generated proteins.

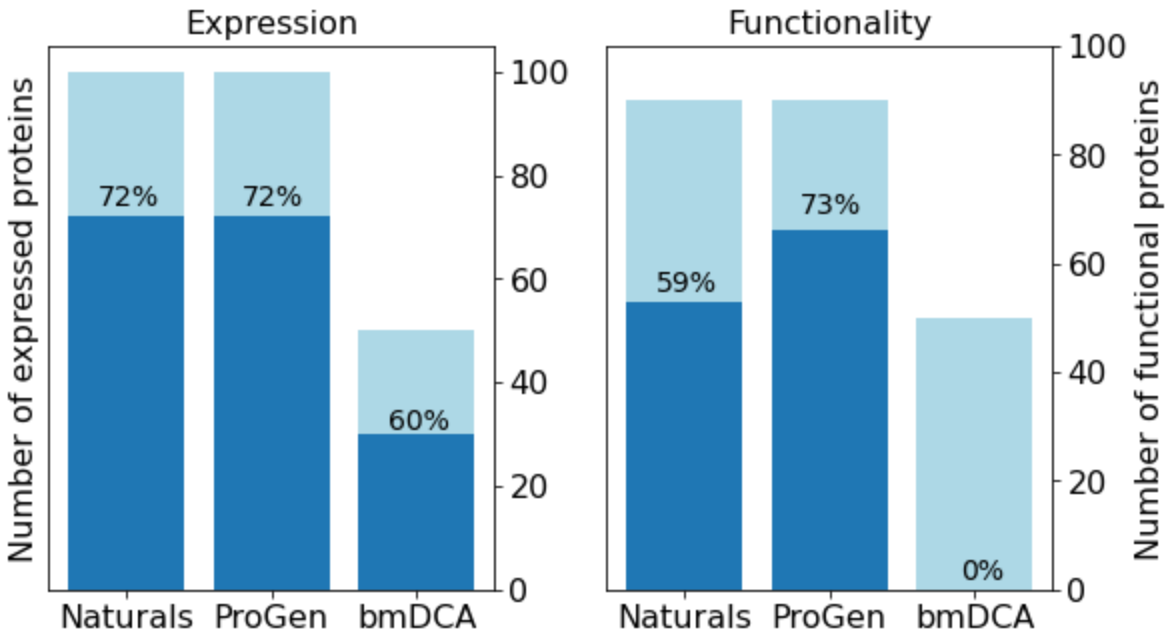


Figure S5. 72% of ProGen-generated proteins and representative natural proteins were well-expressed in cell-free protein synthesis, while the bmDCA-generated proteins showed lower expression (left). The well-expressed proteins were assayed. Among a batch of 90 random artificial sequences from the original 100, 73% of ProGen-generated proteins were found to be functional, while none of the bmDCA artificial proteins exhibited a detectable level of functionality (right).

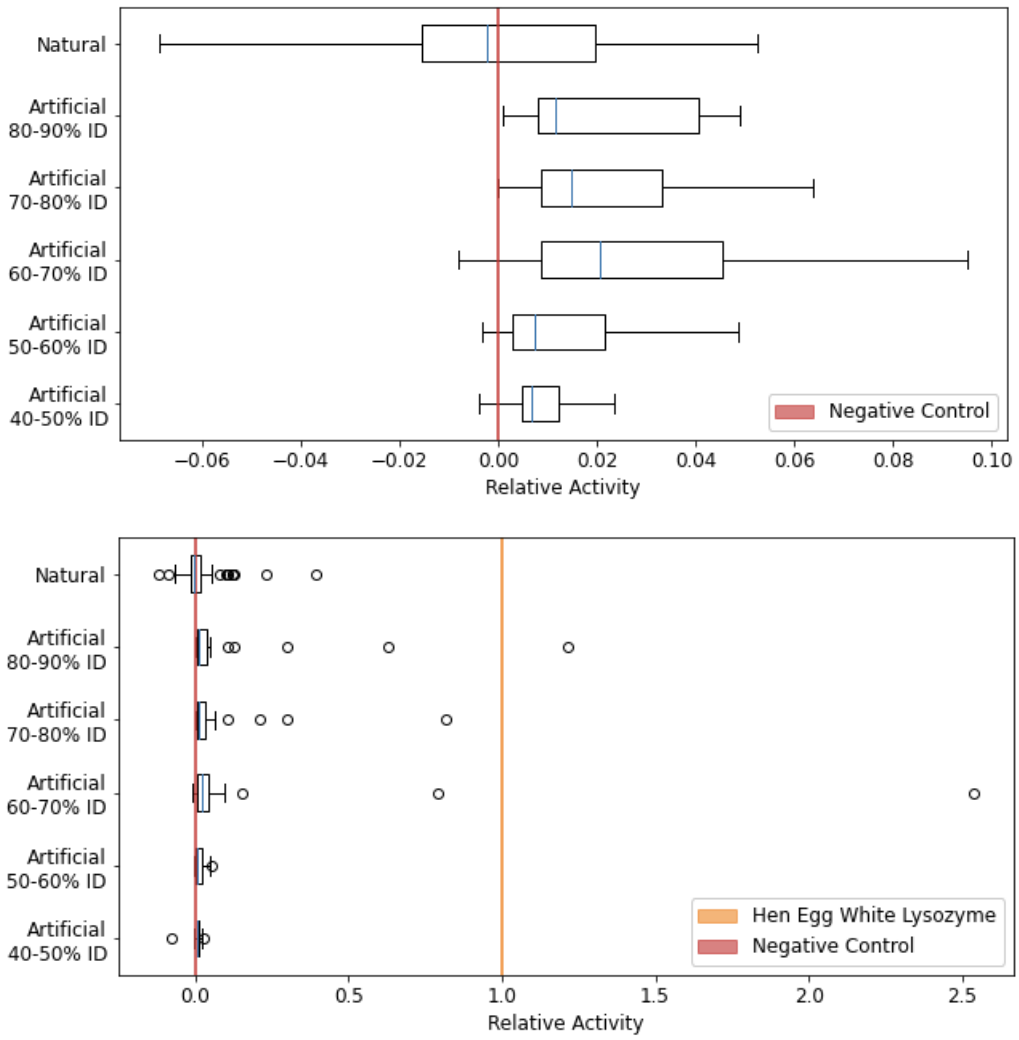
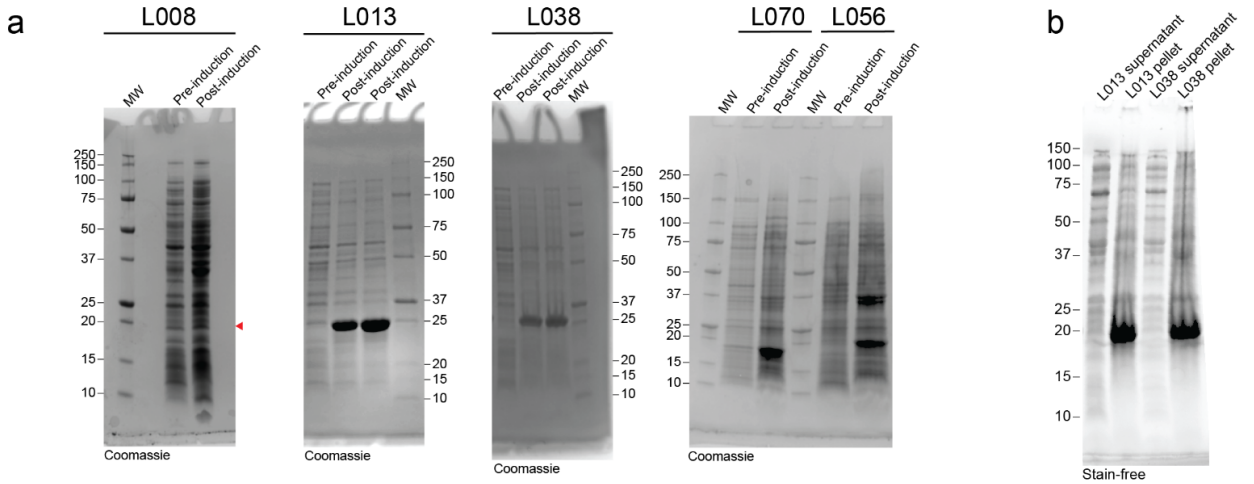
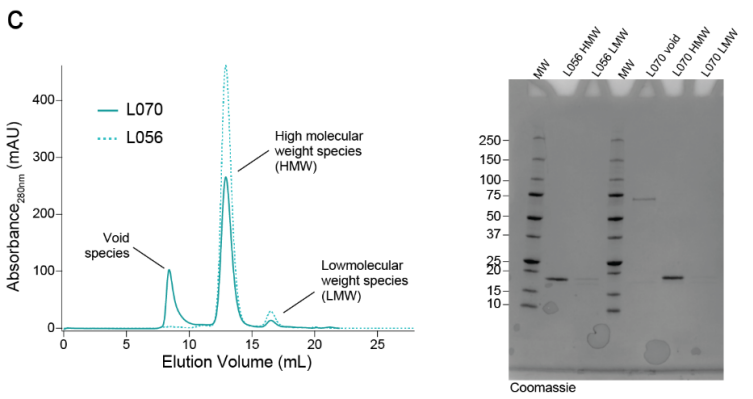
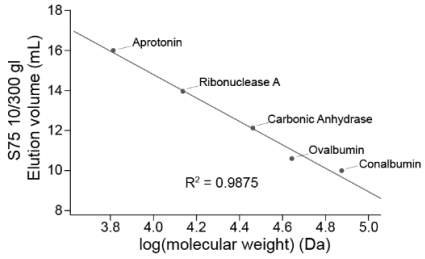


Figure S6: Relative activity of natural and artificial lysozymes as measured by the Enzchek kit with the high-throughput *in vitro* translation/transcription (HT-IVT) expression protocol. Boxplots are derived from n = 90, 23, 28, 22, 8, 9 samples for each category from top to bottom respectively. Boxes display the median, first quartile, and third quartile with whiskers which extend to 1.5x the inter-quartile range. (Top) The artificial proteins across all max identity bins have a shifted distribution toward more functional values than the natural proteins. The max identity is calculated by finding the closest sequence in any publicly available database of natural proteins. (Bottom) While the majority of natural and artificial proteins are less active than hen egg white lysozyme, there exist outliers out of the one hundred samples, with substantially higher activity.







Lysozyme	Elution volume (mL)	Measured MW (kDa)	Expected MW (kDa)
L056 HMW	12.92	20.9	19.1
L070 HMW	13.00	20.3	18.0
L056 LMW	16.927	4.3	19.1
L070 LMW	16.625	4.9	18.0

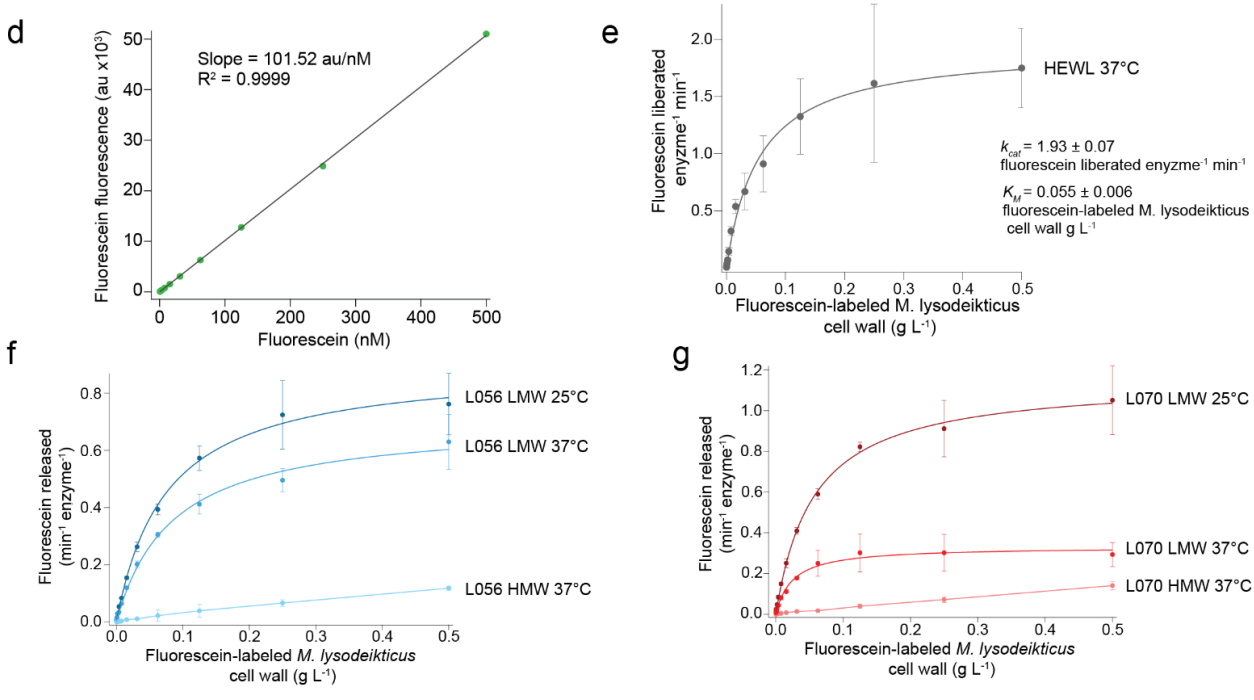


Figure S7: (a) Analysis of generated lysozyme overexpression in *E. coli* BL21(DE3) by SDS-PAGE. Red arrow indicates the expected molecular weight (MW) of L008 that was not observed. All other variants tested displayed robust overexpression. (b) Lysate clarification by centrifugation demonstrates that L013 and L038 expressed as insoluble inclusion bodies and were found entirely in the pellet fraction. Stain-free imaging was conducted as per manufacturer protocol (Bio-Rad) on 15-well Any kD mini-PROTEAN gels. (c) Representative size-exclusion chromatography elution profiles of L056 and L070 with relevant peaks labeled (left). SDS-PAGE analysis of different native molecular forms of L056 and L070 indicate that both species are a result of different oligomeric forms and not a cleavage product (middle). Standard curve for S75 10/300 gl (GE; top right) used to estimate molecular weight of the two species of L056 and L070 (bottom right). (d) Fluorescein standard curve utilized in the analysis of initial rates. (e) HEWL Michaelis-Menten kinetics using the fluorescein-labeled *Micrococcus lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit) at 37°C. Points represent the average and error bars the standard deviation of technical replicates (n = 3). Line is resultant of nonlinear curve fitting to the Michaelis Menten model (Eq. 4). (f) Low molecular weight (LMW) L056 Michaelis-Menten kinetics using the fluorescein-labeled *Micrococcus lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit) at 25°C and 37°C. Points represent the average and error bars the standard deviation of technical replicates (n = 3). Lines for LMW L056 data are resultant of nonlinear curve fitting to the Michaelis Menten model (Eq. 4). High molecular weight (HMW) L056 activity against a titration of fluorescein-labeled *Micrococcus lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit) at 37°C could not be fit to Eq. 4 and is represented here as lines connecting points. (g) Low molecular weight (LMW) L056 Michaelis-Menten kinetics using the fluorescein-labeled *Micrococcus lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit) at 25°C and 37°C. Points represent the average and error bars the standard deviation of technical replicates (n = 3). Lines for LMW L056 data are resultant of nonlinear curve fitting to the Michaelis Menten model (Eq. 4). High molecular weight (HMW) L056 activity against a titration of fluorescein-labeled *Micrococcus lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit) at 37°C could not be fit to Eq. 4 and is represented here as lines connecting points.

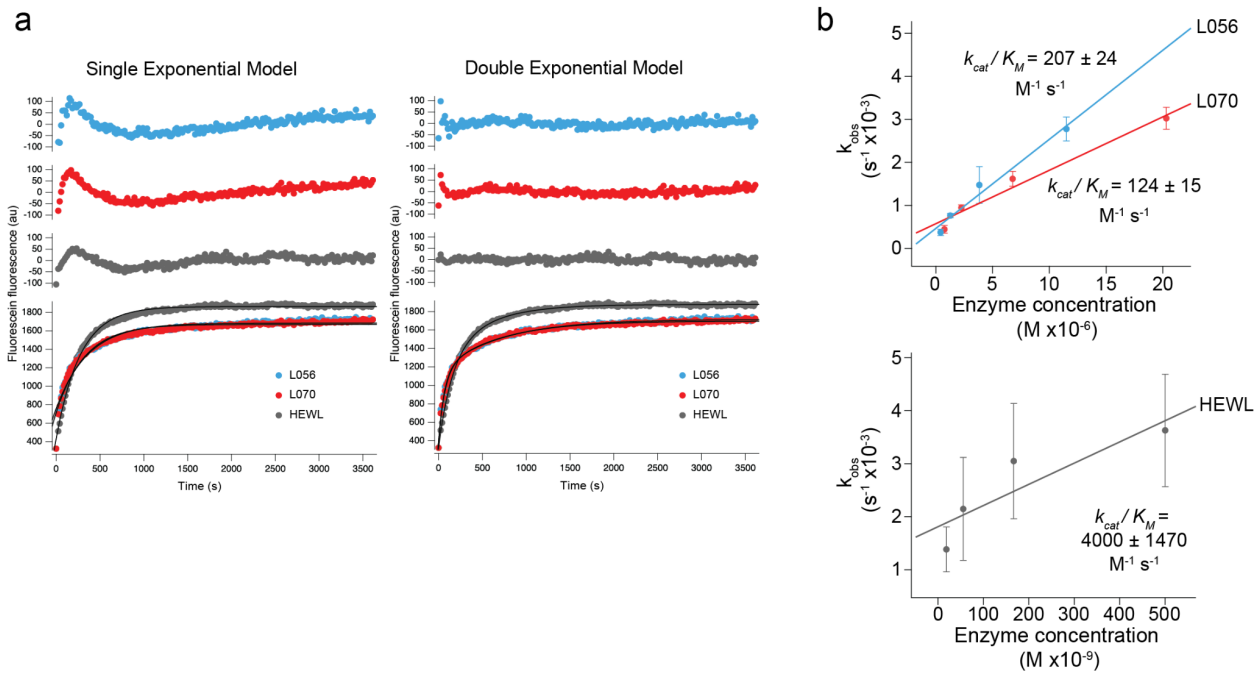


Figure S8. Pseudo-first order reaction kinetic analysis. (a) Residual analysis of data fitted by a single exponential model (left) or double exponential model (right; Eq. 6). (b) Linear fitting extrapolation of k_{cat}/K_M from L056 (blue; left), L070 (red; left) and HEWL (grey; right) from Eq. 5. Uncertainty represents standard deviation of fit. Points represent the average and error bars represent the standard deviation of technical replicates (n = 5).

	L056 (7RGR)
Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, *b*, *c* (Å)	61.15, 68.1, 95.41
α, β, γ (°)	90, 90, 90
Resolution (Å)	55.5-2.475 (2.54-2.475) *
R_{meas}	0.1372 (1.75 n/a)
$I / \sigma I$	13.72 (1.96)
Completeness (%)	99.9 (99.6)
Redundancy	12.8 (12.0)
Refinement	
Resolution (Å)	2.48
No. reflections	14740
R_{work} / R_{free}	0.2525 / 0.2915
No. atoms	
Protein	5393
Ligand/ion	14
Water	33
B-factors	62.8
Protein	69.21
Ligand/ion	94.2
Water	55.5
R.m.s. deviations	
Bond lengths (Å)	0.0027
Bond angles (°)	0.59

*Values in parentheses are for the highest-resolution shell.

Table S6: Data collection and refinement statistics (molecular replacement) for L056 crystal structure

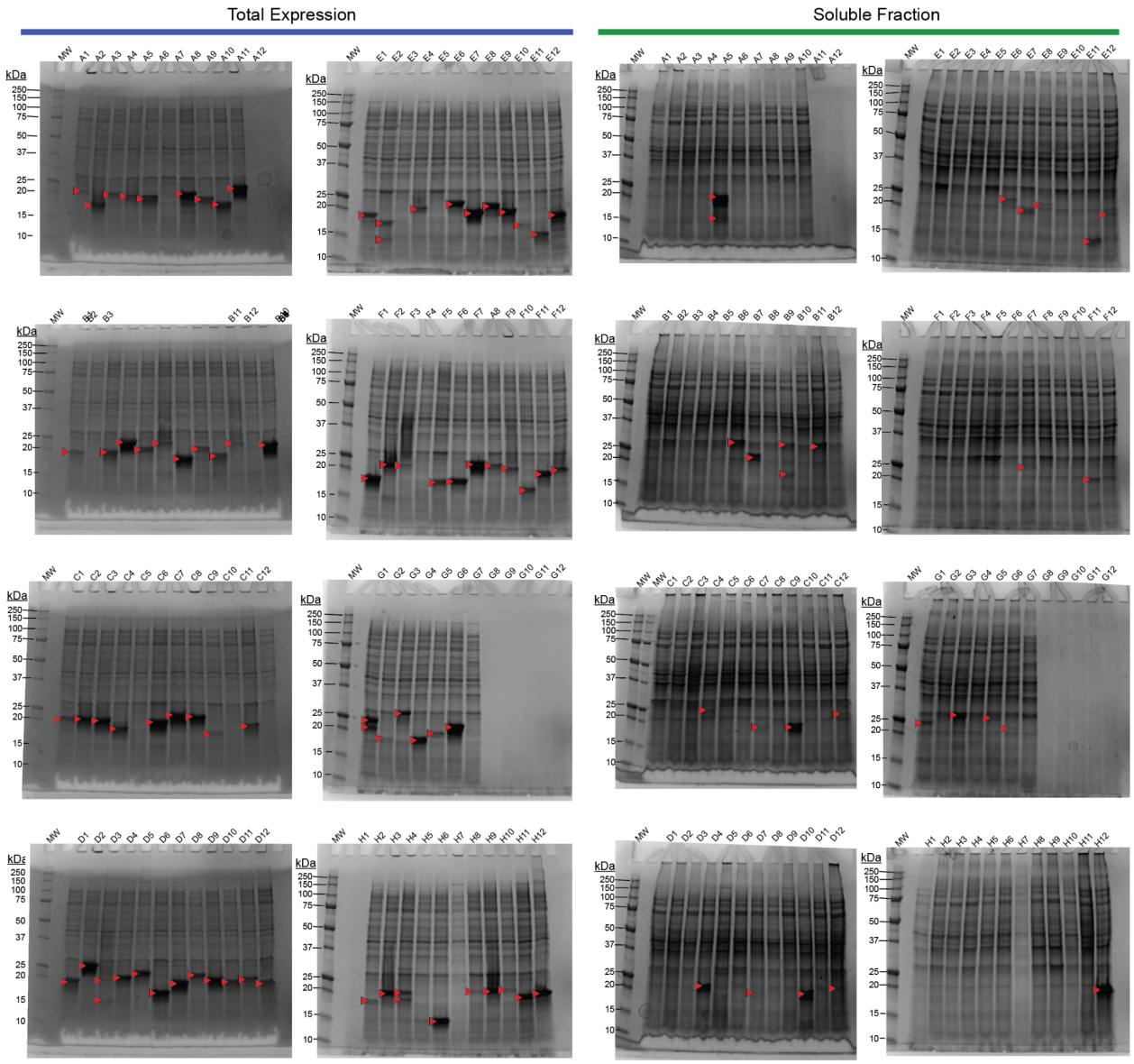


Figure S9. Expression tests of low max ID generated lysozyme variants. MW denotes lanes containing BioRad Precision Plus molecular weight markers whose values are provide to left of each gel under “kDa”. Variants were expressed in BL21(DE3) cells grown in ZYM-5052 autoinduction media[1] overnight at 37°C. High density induced cell cultures were detergent lysed (Promega), and clarified by centrifugation to assess total and soluble lysozyme expression profiles. Samples were electrophoresed on Bio-Rad anyKD TGX gels and stained with colloidal coomassie (Abcam) and shown above. Molecular weight markers are indicated in units of kDa and well ID is marked above each gel lane. Variants are labeled by well-ID corresponding to position within 96-well cell stock and red arrows denmark detectable overexpressed and/or soluble lysozymes. Well H12 corresponds to L056 that was used as a positive control. From this analysis, 89 of 96 cell stocks were viable (93%). Of viable cell stocks, total expression was

detected for 78 (88%) and of expressed variants, 24 were soluble (31%) which is similar to *E. coli* expressed protein solubility rates achieved for naturally occurring proteins[2].

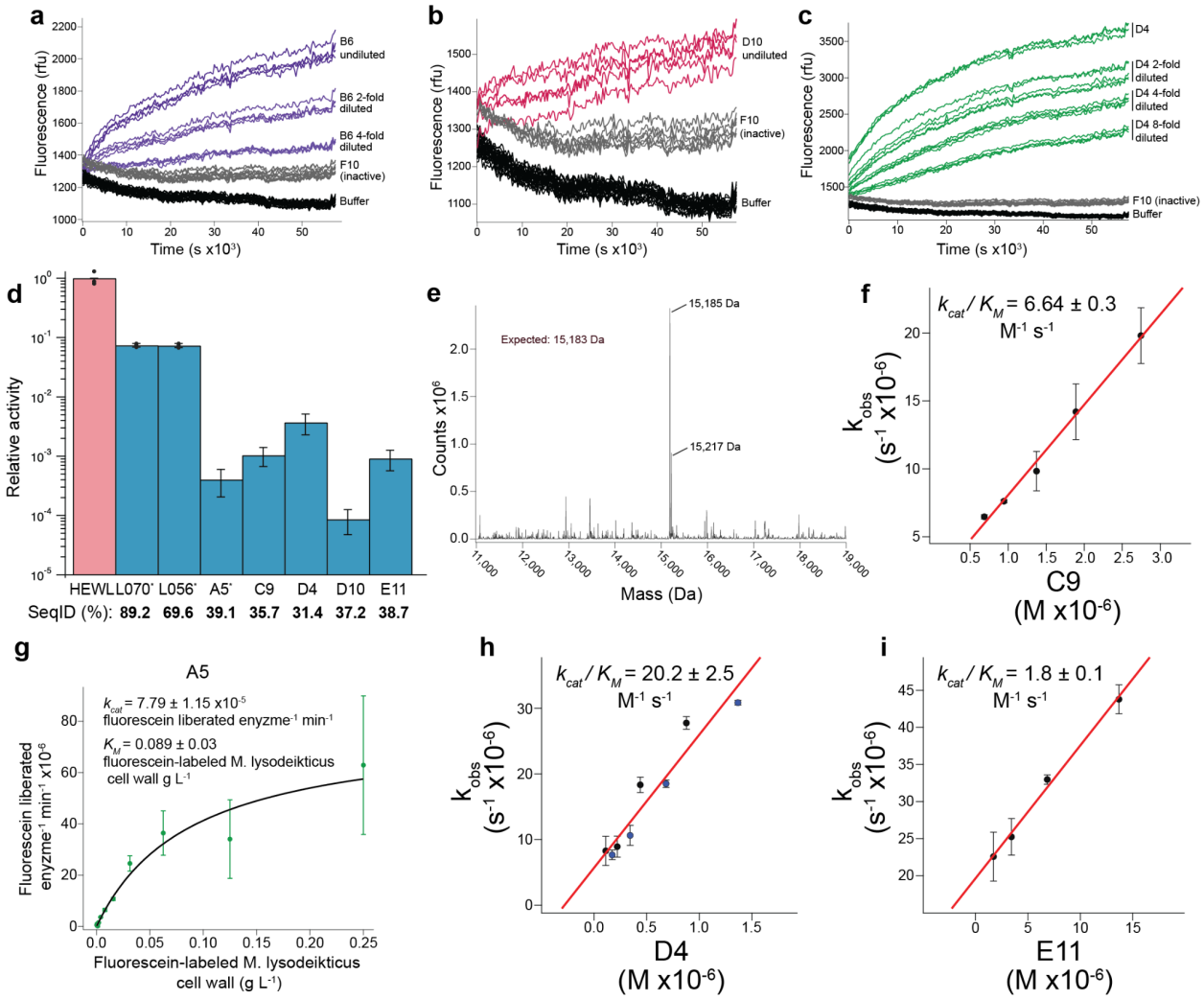


Figure S10. Kinetic characterization of low max ID lysozyme variants. (a) raw fluorescence traces from lysozyme variant B6 (max seqID 38.5%) hydrolyzing fluorescein-labeled Micrococcus lysodeikticus cell wall substrate (Molecular Probes EnzChek Lysozyme kit) with a buffer only negative control and additional verified inactive lysozyme variant (F10) negative control performed at 25°C. (b) same as (a) except displaying activity of D10 variant. D10's low activity is at least partially due to precipitation at room temperature. (c) same as (a) except displaying activity of D4 whose activity was the highest of the low max ID lysozyme variants test. (d) Mean relative activity measurements derived using Eq. 3 were error bars reflect standard deviation (HEWL, L070, and L056 n = 3 overlaid with individual measurements; A5 n = 12; C9, D4, D10, and E11 n = 16) were plotted with maximal sequence ID to any natural protein noted. (e) LC-MS spectra of C9 demonstrating correct predominant molecular weight species corresponding to C9. The +32 molecular weight adduct likely corresponds to persulfide oxidation of C9's single cysteine residue. Spectra were collected on a Waters TQ detector, ionized via electrospray after being resolved by reverse phase chromatography on a C4 column held at 40°C on a Waters Acquity LC. Sample was eluted from C4 column with a linear gradient run at 0.2 mL/min for 6.5 minutes starting at 95% solvent A (water with 0.1% formic acid)

running to 95% solvent B (acetonitrile with 0.1% formic acid) after a wash for 1.5 min in 95% solvent A. Finally, an isocratic elution consisting of 95% solvent B for 2 mins was run. (f) Linear fitting extrapolation of k_{cat}/K_M for C9 derived from Eq. 5. Uncertainty represents standard deviation derived from the fit. Data points represent the mean and error bars represent the standard deviation of technical replicates (n = 4). (g) Pseudo-first order kinetics for variant A5 could not be fit by Eq. 6 due to slow overall rate for this enzyme but initial rates could be determined and utilized to derive Michaelis-Menten parameters. Data points represent background subtracted average and error bars represent the standard deviation of technical replicates (n = 4). (h) Same as (f) except for variant D4 is presented, data represent two biological replicates whose colors are differentiated by biological replicate. Data are presented as the mean and error bars represent the standard deviation of technical replicates (n = 4) (i) same as (f) except for variant E11 is presented. Data are presented as the mean and error bars represent the standard deviation of technical replicates (n = 4)

		k_{cat} (fluorescein released enzyme^{-1} min^{-1})	K_M (*M. lysodeikticus* cell wall g L^{-1})	k_{cat} / K_M (L g^{-1} min^{-1})
HEWL	25°C	1.03 ± 0.06	0.015 ± 0.003	68 ± 14
	37°C	1.93 ± 0.07	0.055 ± 0.006	35 ± 4
L056 LMW	25°C	0.90 ± 0.02	0.075 ± 0.005	12.0 ± 0.8
	37°C	0.70 ± 0.02	0.084 ± 0.007	8.3 ± 0.7
L070 LMW	25°C	1.15 ± 0.02	0.056 ± 0.003	21 ± 1.2
	37°C	0.33 ± 0.01	0.025 ± 0.003	13 ± 1.6
A5	25°C	7.79 ± 1.15 (x 10^{-5})	0.089 ± 0.03	8.8 ± 3.3 (x 10^{-4})

	k_{cat} / K_M at 37°C (M^{-1} s^{-1})
HEWL	4000 ± 1470
L056 HMW	207 ± 24
L070 HMW	124 ± 15

	k_{cat} / K_M at 25°C (M^{-1} s^{-1})
C9	6.64 ± 0.3
D4	20.2 ± 2.5
E11	1.8 ± 0.1
A5*	0.10 ± 0.05
L056 LMW*	1371 ± 536
L070 LMW*	2347 ± 913

Table S7. Derived Michaelis-Menten Constants. HEWL represents hen egg white lysozyme, LMW represents low-molecular weight and HMW represents high-molecular weight. * denotes k_{cat} / K_M values that were converted from (L g^{-1} min^{-1}) to (M^{-1} s^{-1}) using HEWL data collected at 37°C as a standard conversion ratio. Uncertainty represents standard deviations that were propagated through all conversion steps with n ≥ 3 technical replicates (precise replicates are illustrated in associated figure legends). Michaelis-Menten parameters could not be derived for the B6 artificial enzyme due to enzyme concentration being below a detectable limit nor for the D10 artificial enzyme due to low activity but relative activity and raw EnzChek cell wall substrate fluorescence traces are presented in Figure S9.

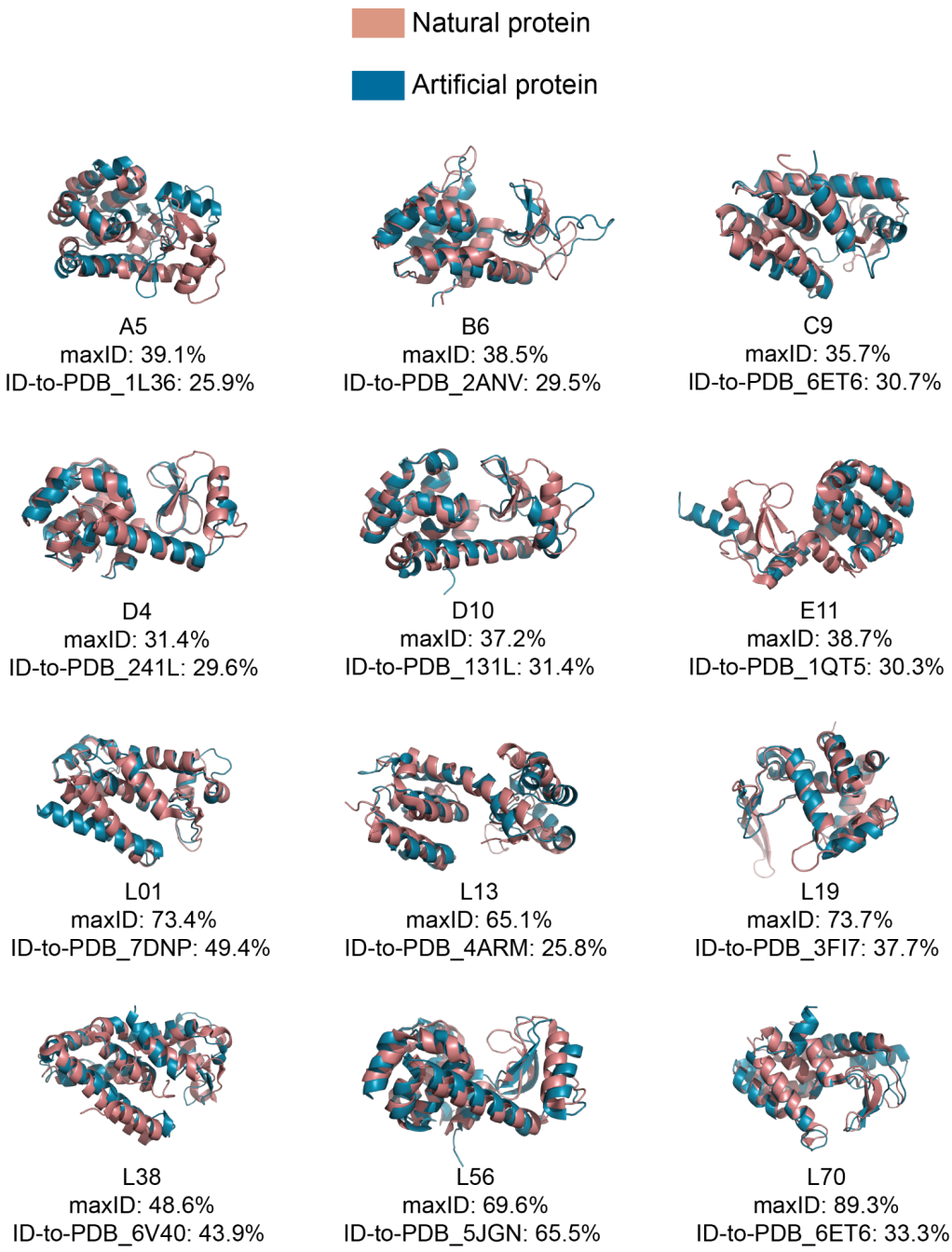


Figure S11. Although generated functional proteins are dissimilar from nature in sequence space, the predicted structures of generated proteins are similar to natural structures. To note, structural information was not provided in model training. Shown are predicted natural and generated structures with AlphaFold2. The closest known PDB was found with FoldSeek[3] and overlaid. The maxID corresponds to the maximum identity to any known natural protein and the ID-to-PDB corresponds to the identity to the most similar structure in PDB.

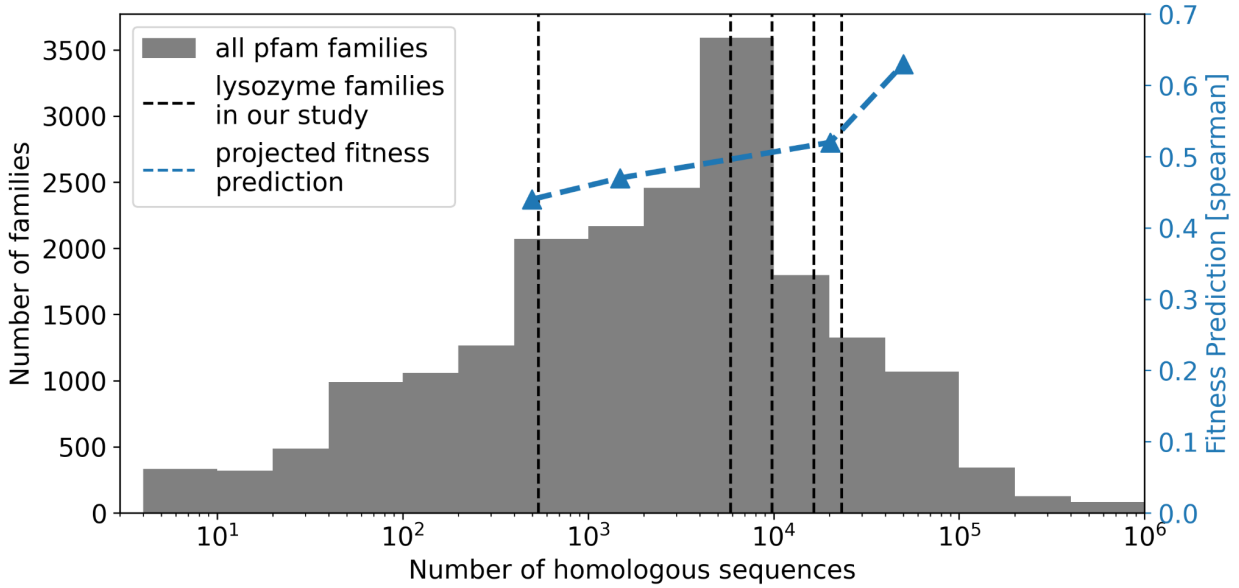


Figure S12. Distribution of available sequences across different protein families in context to our model's capability. The histogram of homologous sequences, available for fine-tuning, for all protein families curated by pfam is shown in gray. The lysozyme families in our study are denoted by vertical lines. Function prediction performance by correlating the likelihood of our model (fine-tuned to GFP[4], AAV[5], CM[6], and PABP[7] protein sequences respectively) to assay-labeled fitness data.

Supplementary References

1. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
2. Bhandari, B. K., Lim, C. S. & Gardner, P. P. TISIGNER.com: web services for improving recombinant protein production. *Nucleic Acids Res.* **49**, W654–W661 (2021).
3. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
4. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
5. Bryant, D. H. *et al.* Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
6. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
7. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).