Computational scoring and experimental evaluation of enzymes generated by neural networks

Sean R. Johnson [1,6], Xiaozhi Fu [2,6], Sandra Viknander[2], Clara Goldin [2], Sarah Monaco [7], Aleksej Zelezniak [2,3,4] ✉ & Kevin K. Yang [5] ✉

In recent years, generative protein sequence models have been developed to sample novel sequences. However, predicting whether generated proteins will fold and function remains challenging. We evaluate a set of 20 diverse computational metrics to assess the quality of enzyme sequences produced by three contrasting generative models: ancestral sequence reconstruction, a generative adversarial network and a protein language model. Focusing on two enzyme families, we expressed and purified over 500 natural and generated sequences with 70–90% identity to the most similar natural sequences to benchmark computational metrics for predicting in vitro enzyme activity. Over three rounds of experiments, we developed a computational filter that improved the rate of experimental success by 50–150%. The proposed metrics and models will drive protein engineering research by serving as a benchmark for generative protein sequence models and helping to select active variants for experimental testing.

Nature provides a wealth of proteins that can be used as biocatalysts to produce valuable products ranging from commodity chemicals to lifesaving pharmaceuticals[1]. Advances in DNA synthesis and recombinant DNA techniques have made it possible to clone genes that encode proteins into industrial organisms such as *Escherichia coli*[2]. Therefore, the use of recombinant proteins for industrial and therapeutic purposes has been highly successful[3,4]; however, the requirements associated with human applications are often not satisfied by natural proteins and engineering is needed to adapt them to human needs[5].

A conventional method for moving beyond natural sequence and function space is to use directed evolution by starting from a natural protein and iteratively screening mutations until the protein acquires the desired properties[5,6]. Many mutations result in nonfunctional proteins[7,8] and up to 70% of random single-amino acid substitutions result in decreased activity[9–12]. Computational models enable the generation of new and diverse sequences from a protein family, thereby uncovering

previously untapped functional sequence diversity and reducing the number of nonfunctional sequences that need to be tested[13]. Typically, these generative models are trained either on large collections of protein sequences, for example the entire UniProt database of millions of sequences[14–16], or on a set of proteins from a specific family[17,18], with the goal of learning the training distribution to sample novel sequences with desired properties. The underlying assumption of these models is that natural proteins are under evolutionary pressure to be functional; therefore, novel sequences drawn from the same distribution will also be functional[19]. Many generative protein models have been proposed, including models based on deep neural networks, such as generative adversarial networks (GANs)[17], variational autoencoders (VAEs)[18,20], language models[15,16,21–24] and other neural networks[25,26], as well as statistical methods such as ancestral sequence reconstruction (ASR)[27,28] and direct coupling analysis (DCA)[29–31]. However, comparing the ability of these methods to generate functional proteins remains

[1]New England Biolabs, Ipswich, MA, USA. [2]Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden. [3]Institute of Biotechnology, Life Sciences Centre, Vilnius University, Vilnius, Lithuania. [4]Randall Centre for Cell & Molecular Biophysics, King's College London, Guy's Campus, London, UK. [5]Microsoft Research, Cambridge, MA, USA. [6]These authors contributed equally: Sean R. Johnson, Xiaozhi Fu. [7]Unaffiliated: Sarah Monaco. ✉e-mail: aleksej.zelezniak@chalmers.se; yang.kevin@microsoft.com

a challenge because of limited experimental work evaluating model performance; likewise, there is no experimental validation supporting common computational metrics.

Typically, protein generative models are evaluated by comparing the distribution of generated sequences to natural controls using alignment-derived scores, for example, identity to the closest natural sequence[15,17]. The few reported results from biological assays[15,20,23,29,32,33] used different experimental systems, making comparisons difficult because many factors can contribute to poor expression and activity (Supplementary Table 1), ranging from mutations disrupting protein folding and stability[34] to codon usage hindering expression[35,36]. Thus, computational metrics for predicting the activity of generated sequences should account for as many factors as possible. For example, alignment-based metrics such as sequence identity or BLOSUM62 scores[37] rely on homology to natural sequences and are good at detecting general sequence properties. However, they do not account for epistatic interactions and give equal weight to all positions[38]. In contrast, alignment-free methods do not require homology searches, are fast to compute and can potentially identify all sequence defects based on the likelihoods computed by protein language models[39]. Protein language models are sensitive to pathogenic missense mutations[40], predict evolutionary velocity[41] and capture viral immune-escape mutations[42]. Structure-supported metrics, including Rosetta-based scores[43], AlphaFold2 (ref. [44]) residue confidence scores and likelihoods computed by neural network inverse folding models[45–47], use atomic coordinates to capture protein function; however, they can be expensive to use, especially when evaluating thousands of sequences. Although it is important to rationally choose metrics for computationally evaluating sequences, it is crucial to experimentally validate the ability of the metrics to predict function.

In this study, we focused on assessing computational metrics to predict the functionality of computer-generated protein sequences. We experimentally evaluated in silico metrics for the ability to predict in vitro enzyme activity using sequences produced by three generative models trained on two enzyme families. Over three rounds of experiments (Fig. 1a) we developed and experimentally validated composite metrics for protein sequence selection (COMPSS), a framework that allows the selection of up to 100% of phylogenetically diverse functional sequences. COMPSS is generalizable to any protein family, and we provide examples as Google Colab notebooks. Our study demonstrates a composite computational metric for evaluating generated sequences that predicts experimental success. In addition to selecting active sequences for experimental validation, the proposed metrics are a first step toward establishing a standard for evaluating the performance of current and future protein generative models and will hopefully be a catalyst for driving progress in protein engineering.

Results

In this work, a protein is considered experimentally successful if it can be expressed and folded in *E. coli* and has activity above background in an in vitro assay (Methods). We tested three protein generative models: (1) the transformer-based multiple-sequence alignment (MSA) language model ESM-MSA[48]; (2) a convolutional neural network with attention trained as a GAN (ProteinGAN)[17]; and (3) a phylogeny-based statistical model for ASR[28]. Although ESM-MSA is not trained as a generative model, it can be used to generate new sequences via iterative masking and sampling[49,50]. ASR is also not a truly generative model as it is constrained within a phylogeny to traverse backward in evolution without the ability to navigate sequence space in a new direction. However, it has successfully resurrected ancient sequences[51] and increased enzyme thermotolerance[52]. To combine the strengths of the different methodologies, we considered alignment-based, alignment-free and structure-based metrics (Fig. 1b).

We experimentally evaluated the metrics on two enzyme families, malate dehydrogenase (MDH) and copper superoxide dismutase (CuSOD). Both MDH and CuSOD have substantial sequence diversity, have numerous members in the Protein Data Bank (PDB) and are physiologically significant[53,54]. They are also relatively small (300–350 residues for MDH and 150–250 residues for CuSOD) but are complex proteins active as multimers, and their activity can be assayed by spectrophotometric readout.

Round 1: Naive generation results in mostly inactive sequences
For round 1, we constructed training sets to train the generative models of CuSOD and MDH. We collected 6,003 CuSOD and 4,765 MDH sequences from UniProt (Supplementary Table 2) with the Pfam domains for each protein family. CuSOD sequences had only a single Sod_Cu domain, whereas MDH sequences had an Ldh_1_N domain followed by an Ldh_1_C domain and no other Pfam domains. Nontypical domain architectures occurred in 6.3% and 1.7% of sequences in CuSOD and MDH, respectively. Sequences were truncated around the annotated domains to remove possible signal peptides, transmembrane domains and extraneous unannotated domains. Signal peptides are N-terminal leader sequences that facilitate secretion and are present in many proteins[55]. Signal peptides are frequently cleaved after secretion and are not present in the mature protein. In heterologous expression systems, signal peptides may not efficiently direct secretion or be cleaved, thereby interfering with protein expression[56]. Proteins with transmembrane domains are difficult to express and purify in heterologous systems[57]. We generated >30,000 sequences from the ASR, ProteinGAN and ESM-MSA models (Supplementary Table 3) and selected 144 sequences for experimental validation: 18 for each model and a set of natural test sequences. All generated and natural test sequences were selected to have 70–80% identity to the closest natural training sequence (Supplementary Table 4).

Of all experimentally tested sequences, including natural sequences, 19% were active (Extended Data Table 1 and Supplementary Figs. 1–6). None of the CuSOD ESM-MSA or test sequences and only two of the CuSOD GAN sequences were active. None of the MDH GAN or ESM-MSA sequences were active, but six of the MDH test sequences were active. In contrast, ASR generated 9 of 18 and 10 of 18 active enzymes for CuSOD and MDH, respectively.

We investigated the potential reasons for poor performance. We observed that natural test sequences with predicted signal peptides or transmembrane domains in the pretruncation sequences (Methods) were significantly overrepresented in the nonactive set (one-tailed Fisher test, $P = 0.046$). For CuSOD, a literature search[58,59] combined with examination of the assayed sequences and the available CuSOD crystal structure (PDB: 4B3E)[60] showed that CuSOD is active as a homodimer (or sometimes a tetramer) and that the truncations we made to the natural sequences often removed residues at the dimer interface, likely interfering with expression and activity. Thus, we made equivalent truncations to our positive-control enzymes, human SOD1 (ref. [61]) (hSOD, GenBank: NP_000445.1), *Potentilla atrosanguinea* CuSOD[62] (paSOD, GenBank: AFN42318.1) and *E. coli* SOD[63] (E.SOD, GenBank: NP_416173.1), and confirmed loss of activity for hSOD and paSOD (Supplementary Figs. 7 and 8). Overtruncation also affected ASR sequences, yet many were still active, possibly due to the widely reported stabilizing effect of ASR[27,64,65].

To further test the hypothesis that overtruncation led to a lack of activity in the round 1 natural CuSOD test sequences, we assayed an additional 14 natural CuSOD proteins and 2 proteins from the evolutionarily distinct FeSOD family (pretest group). In nature, eukaryotic CuSOD proteins are typically cytosolic and lack a signal peptide, whereas bacterial CuSOD proteins are typically secreted via a signal peptide[58]. CuSOD sequences were selected on the basis of kingdom (eukaryotic, viral or bacterial), and the presence of a signal peptide was predicted using Phobius[66]. Sequences with predicted signal peptides were truncated at the predicted cleavage site. Both of the chosen bacterial FeSOD proteins lacked a predicted signal peptide, as does *E. coli*
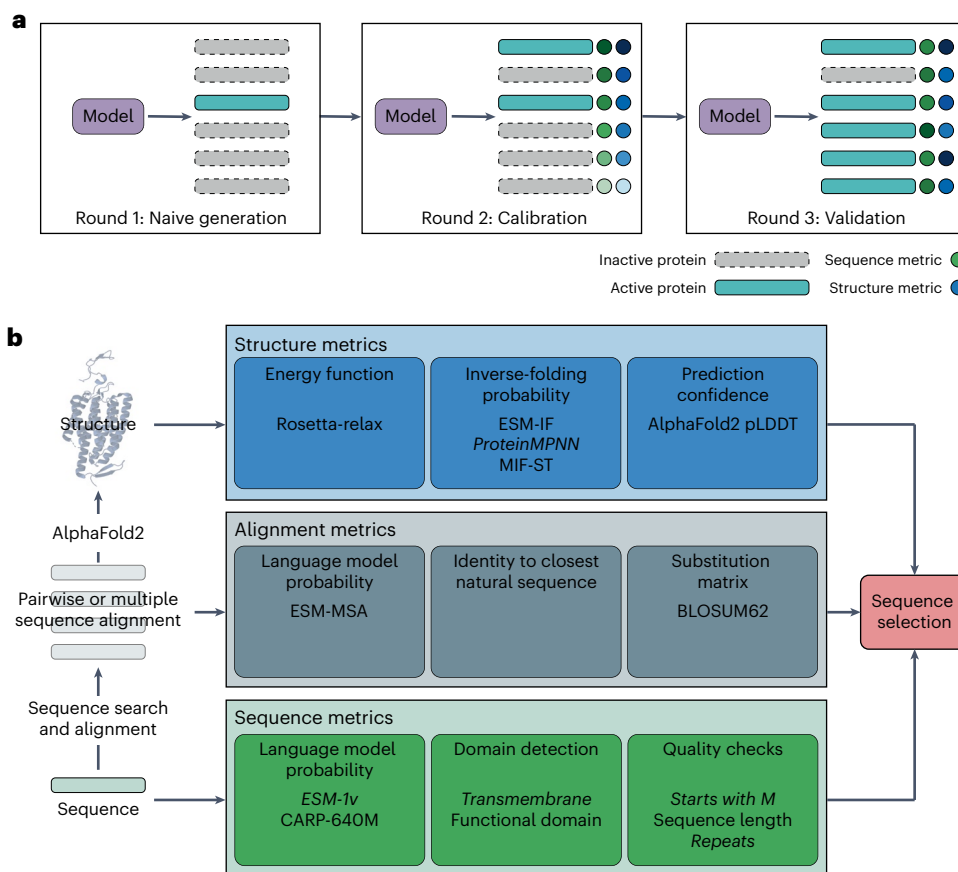
Fig. 1 | Study design. a, COMPSS was developed and tested over three rounds of experiments: (1) naive generation; (2) calibration; and (3) validation. **b**, COMPSS selects sequences using metrics calculated from single sequences, alignments and predicted structures. Metrics in italics were used in the final COMPSS filter.

FeSOD[63] (SodB, GenBank: NP_416173.1, positive control). Activity was noted in 8 of the 14 CuSOD sequences, including 3 of the 4 eukaryotic enzymes and a single viral enzyme, all of which lacked a predicted signal peptide and were expressed in their full-length form (Supplementary Figs. 9–11). Three of the seven signal peptide-clipped bacterial CuSOD enzymes also had activity. For MDH, overtruncation was less problematic and 6 of 17 natural test sequences were active (Extended Data Table 1 and Supplementary Figs. 1a, 2a and 3a).

Round 2: Calibration data for COMPSS
Consolidating the lessons learned in round 1, we retrained the models and tested additional sequences to calibrate the computational metrics. Specifically, for the training set and natural test sequences, we used only full-length natural sequences, removing sequences with predicted transmembrane domains and signal peptides. We also increased the identity band, choosing generated sequences with 80–90% identity to the closest training sequence. For CuSOD, we selected only eukaryotic or viral proteins (Supplementary Table 2). For sequence generation, we used the same method as in round 1 for generating ASR and GAN sequences but modified the ESM-MSA sampling procedure to improve the quality of generated sequences (Supplementary Table 3). ESM-MSA sampling for round 1 used MSAs composed of randomly selected training sequences masked and sampled across the entire MSA. For round 2, only one training sequence from the MSA was masked and sampled at a time based on an MSA composed of the training sequences most similar to the resampled sequence. Sequences generated with the revised ESM-MSA sampling method had higher metric scores, including ESM-1v and identity to the closest training sequence (Supplementary Fig. 12d). We selected 18 sequences each from ASR, GAN and ESM-MSA. Only 13 natural test sequences were selected, because we had already

screened 5 similar natural sequences in the remediation for round 1. Natural test sequences were selected using the same criteria as used for model-generated sequences. The number of expressed enzymes with activity above the background was substantially higher than that in round 1, with 66% of natural controls showing activity when expressed in *E. coli*, and at least 50% of generated sequences were active for every model–enzyme family combination except for GAN–MDH, where only 2 of 18 sequences were active (Extended Data Table 1 and Supplementary Figs. 13–16).

To calibrate the metrics against enzymatic activity, we computed alignment-based (identity, BLOSUM62 (ref. 37), PFASUM15 (ref. 67), phmmer top 30 average score, ESM-MSA) and sequence-only alignment-free (CARP-640M[68], ESM-v1 (ref. 39), net charge[23], abs(net charge), charged fraction) metrics. We also predicted AlphaFold2 structures[44,69] and used the predicted structures to calculate structure-based Rosetta energies[43], solvent-accessible surface area (SASA)[23,70], ProteinMPNN[45], ESM-IF[46] and MIF-ST[47]. The experimentally tested sequences were selected to span the entire range of scores for each metric (Supplementary Table 4). To identify metrics capable of detecting failure modes undetectable by an expert human, we manually excluded candidates with large insertions, deletions or long repeats and added an N-terminal methionine to seven generated sequences. Apart from the alignment-based ESM-MSA metric, none of the metrics strongly correlated with sequence identity, suggesting that our chosen metrics are orthogonal to sequence identity (Supplementary Fig. 17). In contrast, structure-based metrics were substantially correlated, with the highest correlations between inverse folding neural network scores (Fig. 2b and Supplementary Fig. 18).

Area under the curve receiver operating characteristic values (AUC-ROCs) between activity and each metric (Fig. 2a, Table 1 and
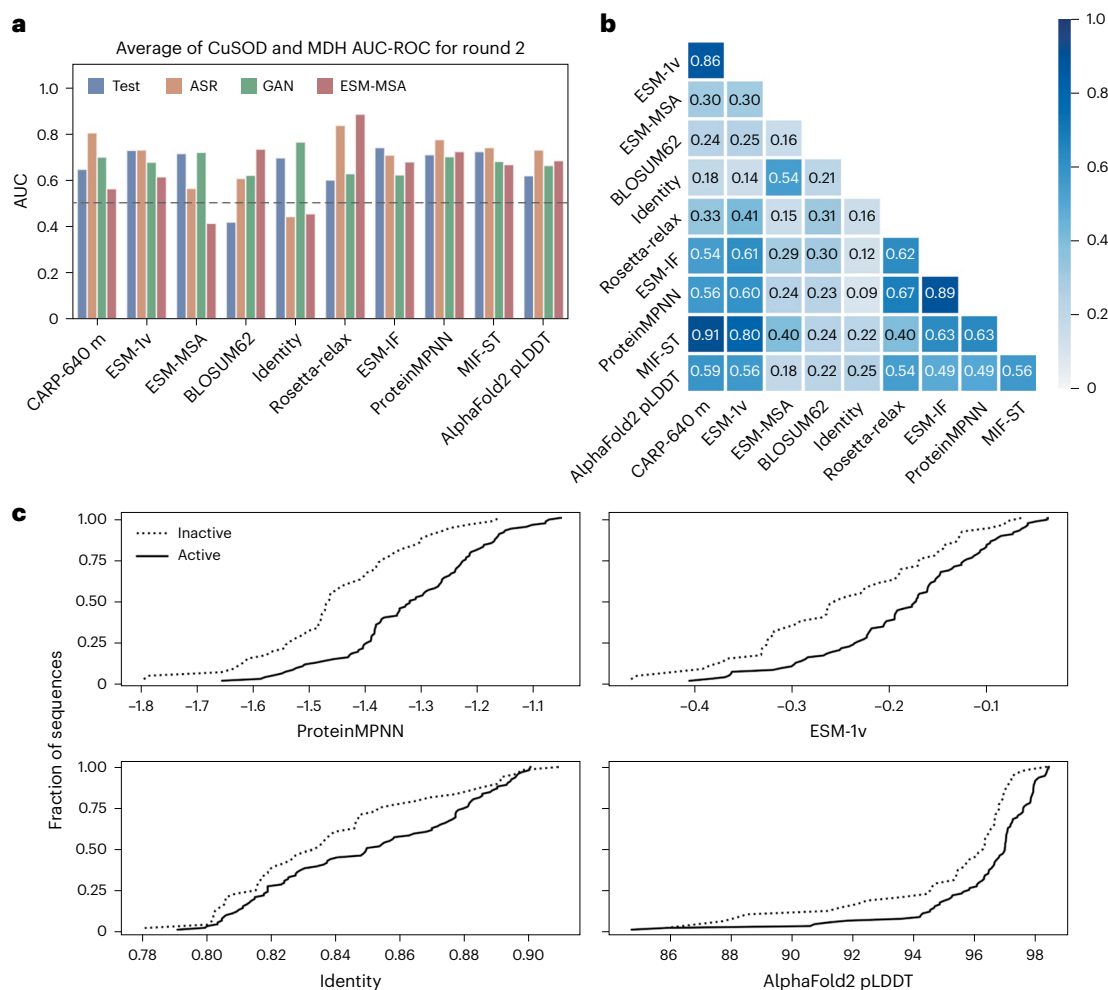
Fig. 2 | Computational metrics of sequences experimentally tested in round 2. a, AUC-ROC scores of activity versus metrics. The average of CuSOD and MDH. **b**, Spearman correlations between metrics. The average of CuSOD and MDH.

c, Empirical cumulative distribution functions of selected metrics for active (solid lines) and inactive (dashed lines) sequences. Curves represent the pooled results of all generative models and both enzyme families.

Supplementary Figs. 18–20) indicate that the metrics are predictive of activity, but none stands out as superior to the others. Inverse folding metrics, on average, best predicted enzymatic activity, showing an AUC-ROC of 0.72 when combining all models and families. AlphaFold2 residue confidence pLDDT scores were significantly predictive for CuSOD (Fig. 2c; Wilcoxon rank sum $P = 5 \times 10^{-4}$), but not for MDH (Table 1). Sequence identity did not predict activity (Table 1, Fig. 2c and Supplementary Fig. 18).

Round 3: Validation: COMPSS enriches active protein generation

Next, we devised an in silico filter to virtually screen large numbers of generated sequences to identify probable active sequences with <80% identity to the closest natural sequence. Based on round 2 (Fig. 2, and Table 1), no single metric was sufficiently general against multiple failure modes (Supplementary Table 1); therefore, we tested a filter composed of a combination of the ESM-1v and ProteinMPNN metrics (Fig. 3a and Supplementary Fig. 21). This combination was attractive because ESM-1v is sequence based, ProteinMPNN considers structural information and neither metric is strongly correlated with sequence identity (Fig. 2a). Although most inverse folding or energy function-based metrics performed similarly (Table 1), ProteinMPNN was the most computationally efficient. Rosetta-relax performed best on MDH sequences but is more computationally expensive than other structure-based metrics. The fast-to-compute ESM-1v protein language

sequence model showed the best performance for alignment-free metrics, with an average AUC-ROC of 0.68. Furthermore, the two metrics were only moderately correlated, with Spearman's $\rho = 0.60$ (Fig. 2a), suggesting that they capture distinct features.

To select a threshold for ESM-1v scores, we analyzed the results for the natural test sequences from round 2. We found that the highest enrichment of active sequences occurred at approximately the 20th percentile of the ESM-1v scores (the top 38% and 17% for CuSOD and MDH, respectively) (Supplementary Fig. 22). For prospective validation of our sequence prioritization strategy, in round 3, we used the 10th percentile of the natural sequences, making the threshold more stringent because, in practice, the score should be derived from untested natural sequences. To validate our strategy, we focused on the GAN and ESM-MSA models because ASR sequences performed consistently well in rounds 1 and 2 and in the literature. The filter begins with automated quality checks for sequences starting with methionine and lacking long repeats and transmembrane domains, intended to approximate human intuition. Next, we randomly selected 200 ESM-1v threshold-passing sequences with 50–80% identity to natural sequences for each model and enzyme family, predicted their structure using AlphaFold2 and randomly selected 18 of the top 40 sequences based on ProteinMPNN scores for each model and enzyme family combination. For each sequence selected for experimental validation, a negative control was randomly chosen from the sequences failing the ESM-1v filter with an identity to the closest training sequence within

Table 1 | AUC-ROCs of each metric versus experimentally measured activity in round 2

Input	Metric type	Metric	CuSOD					MDH					Average
			Test	ASR	GAN	ESM-MSA	All	Test	ASR	GAN	ESM-MSA	All	All
Single sequence	Residue counting	Net charge	0.73	0.24	0.28	0.28	0.30	0.57	0.50	0.21	0.51	0.48	0.39
		Abs(net charge)	0.27	0.76	0.72	0.75	0.70	0.18	0.17	0.75	0.49	0.39	0.55
		Charged fraction	0.44	1.00	0.42	0.69	0.57	0.36	0.47	0.33	0.49	0.43	0.50
	Language model	CARP-640M	0.73	0.94	0.74	0.70	0.76	0.57	0.67	0.66	0.43	0.60	0.68
		ESM-1v	**0.85**	**0.88**	**0.69**	**0.75**	**0.76**	**0.61**	**0.58**	**0.66**	**0.48**	**0.60**	**0.68**
		ESM-1v mask6	0.83	0.94	0.71	0.76	0.78	0.55	0.53	0.50	0.48	0.53	0.66
		ESM-MSA	0.63	0.53	0.65	0.37	0.53	0.80	0.60	0.79	0.45	0.70	0.61
Sequence alignment	Substitution matrix	Avg(phmmer top 30)	0.60	0.41	0.60	0.79	0.68	0.61	0.40	0.71	0.60	0.50	0.59
		BLOSUM62	0.38	0.62	0.67	0.71	0.61	0.46	0.60	0.57	0.75	0.63	0.62
		PFAMSUM15	0.35	0.18	0.63	0.68	0.61	0.46	0.60	0.54	0.75	0.62	0.62
	Identity	Identity	0.59	0.35	0.74	0.65	0.62	0.80	0.53	0.79	0.26	0.60	0.61
Structure	Energy function	Rosetta-relax	0.65	0.94	0.67	0.89	0.78	0.55	0.73	0.59	0.88	0.75	0.76
	Inverse folding	ESM-IF	0.75	0.88	0.64	0.85	0.76	0.73	0.53	0.61	0.51	0.65	0.70
		ProteinMPNN	**0.78**	**0.88**	**0.65**	**0.90**	**0.80**	**0.64**	**0.67**	**0.75**	**0.55**	**0.70**	**0.75**
		MIF-ST	0.75	0.88	0.68	0.79	0.77	0.70	0.60	0.68	0.55	0.68	0.72
	Surface area	SASA	0.28	0.18	0.43	0.24	0.30	0.59	0.18	0.32	0.58	0.47	0.39
		Polar SASA	0.28	0.53	0.56	0.31	0.40	0.38	0.24	0.46	0.62	0.50	0.45
		Apolar SASA	0.30	0.12	0.35	0.23	0.27	0.71	0.27	0.25	0.56	0.47	0.37
		Percent polar SASA	0.40	0.88	0.75	0.49	0.60	0.29	0.42	0.52	0.57	0.52	0.56
	Prediction confidence	AlphaFold2 pLDDT	0.77	0.88	0.61	0.88	0.77	0.46	0.58	0.71	0.49	0.55	0.66

The ESM-1v and ProteinMPNN metrics shown in bold were selected to be part of the filter for round 3.

1% of that for the passing sequence. The stringency of the ESM-1v filter led to phylogenetic bias, particularly for MDH; nevertheless, the set of screened enzymes covered approximately the same space in round 3 as in round 2 (Fig. 3f,g and Supplementary Figs. 23 and 24).

A total of 144 selected and control sequences were expressed in *E. coli*, purified and assayed for activity (Supplementary Figs. 25–30). The selected enzymes had an identity to the closest natural sequence of >69%. Most of the selected sequences showed in vitro activity, including 94% (17 of 18) and 100% of ESM-MSA CuSOD and MDH enzymes, respectively (Fig. 3b–e and Extended Data Table 1). Furthermore, when combining sequences from both models and enzyme families, 74% were active, which is a 77% higher success rate (two-tailed Fisher test, $P = 0.00018$) than for the sequence-filter-failing control sequences (Fig. 3e and Supplementary Fig. 31). Furthermore, 83% (44 of 53) of active generated sequences selected by COMPSS had activity levels within an order of magnitude of those of the wild-type controls (Fig. 3d), suggesting that COMPSS enriches sequences of sufficient activity and diversity to be possible starting points for engineering with directed evolution.

We also used ProGen[15], a 1.2-billion-parameter protein language model, to generate lysozyme sequences from the glucosaminidase and transglycosylase families. We selected 18 passing and 18 identity-matched controls from each family, as well as 12 previously reported active lysozymes, four natural and eight generated by ProGen. In our study, 14 of 84 sequences expressed and could be purified from *E. coli* (Supplementary Fig. 32), but only the previously reported L056 and L070, generated by ProGen from the phage lysozyme (PF00959) family, showed activity (Supplementary Table 6). The ProGen study used multiple language models, including TAPE-BERT discriminators fine-tuned on lysozymes, to select sequences, suggesting that generating and selecting active enzymes from ProGen requires fine-tuning both the generative and discriminative models for each family.

To further validate COMPSS, we tested it against previously published datasets of experimentally characterized sequences from six additional families generated by models with different architectures from those trained in this work (Supplementary Table 5), including enzymes from five evolutionarily distinct lysozyme families generated by ProGen[15] and chorismate mutases generated by bmDCA[29]. When applying COMPSS to these sequences, we omitted the identity and 'starts with M' filters as they would eliminate most sequences in these datasets. For five of the six families, there was a higher fraction of functional enzymes among those passing the COMPSS sequence filter than among those failing the filter, and ProteinMPNN AUC-ROCs ranged from 0.6 to 1.0 (Fig. 4).

The critical importance of training data curation is evident from the improved success rates between round 1, round 2 and control round 3 sequences. Furthermore, none of the 42 natural and 61 synthetic chorismate mutases with predicted signal peptides were active. Simple sequence quality checks also contributed to the success rate. Applying the round 3 sequence quality checks to round 2 sequences showed that 13 of 24 (54%) check-failing sequences were active, compared with 78 of 115 (68%) check-passing sequences. In round 3, 10 of 21 (48%) check-failing negative-control sequences were active, compared with 20 of 51 (39%) check-passing negative-control sequences, and 53 of 72 (74%) of sequences selected by the combined COMPSS filter, including quality checks, ESM-1v scores and ProteinMPNN scores.

To deconvolve the relative contributions of ESM-1v and ProteinMPNN to the success of the round 3 selections, we considered sequences generated and assayed in rounds 2 and 3 (which used the same training sets and generative models), as well as the chorismate mutase and lysozyme literature datasets. We divided the CuSOD, MDH and chorismate mutase datasets into quadrants based on median ProteinMPNN and ESM-1v scores and calculated the metric AUC-ROCs, Spearman correlations and percentage of active enzymes for the whole
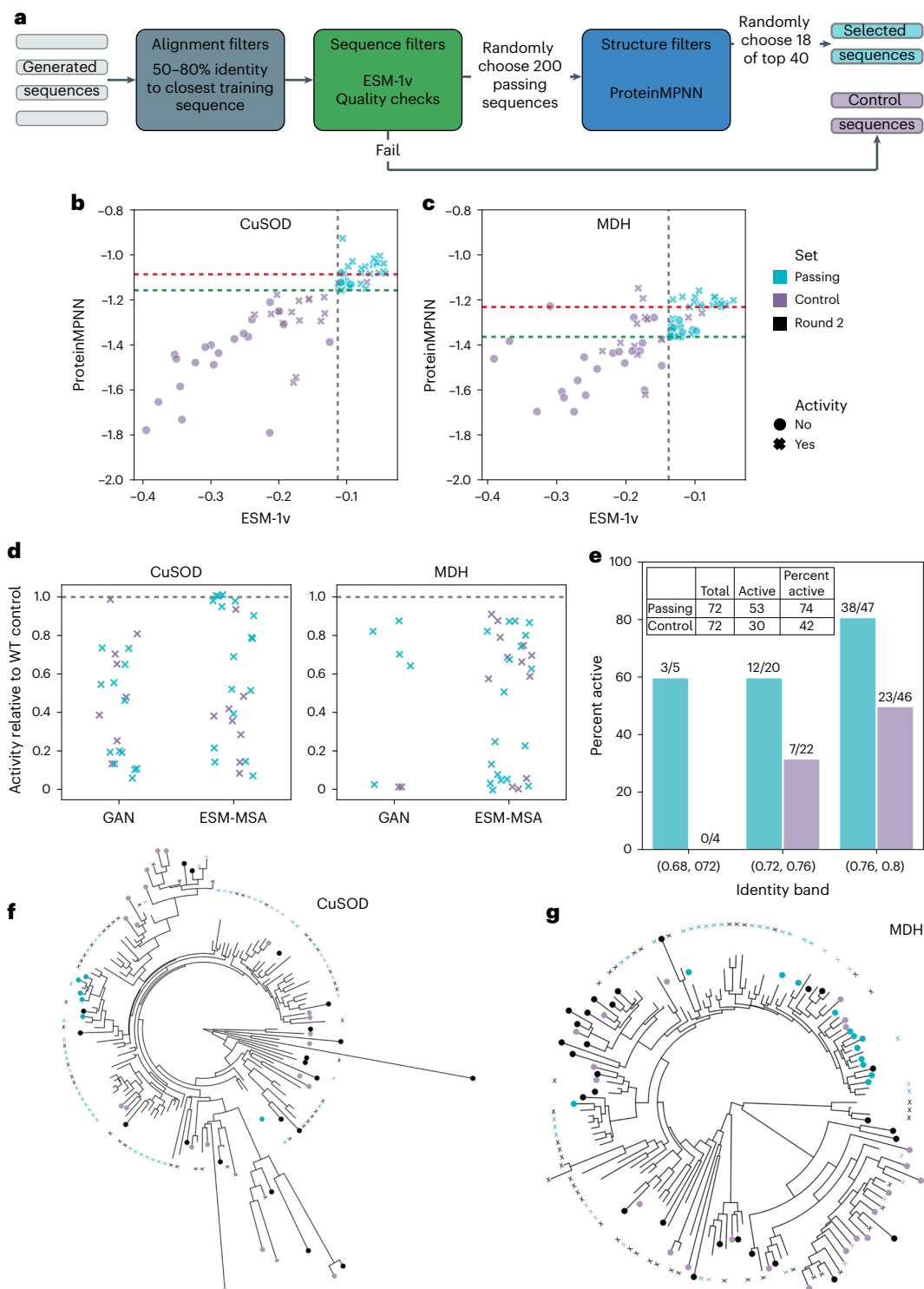
Fig. 3 | Round 3 selection and experimental results. a, Computational filter for selecting sequences to screen in round 3. **b,c**, CuSOD (**b**) and MDH (**c**) ESM-1v and ProteinMPNN scores for round 3 selected passing (teal) and control (violet) sequences. The vertical dashed gray line indicates the top 10th percentile cutoff for ESM-1v score calculated from the test sequences. Horizontal dashed lines are the ProteinMPNN scores of the 40th-ranked sequences among the 200 selected sequences sequences for which the scores were calculated for each model–family

combination, including ProteinGAN (lower, green) and ESM-MSA (upper, red). Control sequences to the right of the gray line are possible if they failed one of the quality checks. **d**, Comparison of the specific activities of active generated sequences versus wild-type (WT) controls (Methods). **e**, Active enzymes by identity band. **f**, Phylogenetic tree of CuSOD enzymes screened in round 2 (black) and round 3 (passing, teal; control, violet). An X indicates an active sequence and a filled circle indicates an inactive sequence. **g**, Phylogenetic tree of MDH sequences.

datasets as well as for each quadrant (Supplementary Fig. 33). The ProteinMPNN to ESM-1v Spearman correlation was higher in the whole dataset than in any individual quadrant. Furthermore, the lower-left quadrant (lowest scores on both metrics) had the lowest success rate.

For chorismate mutase and CuSOD, the upper-right quadrant (highest scores on both metrics) had the highest success rates, whereas for MDH, the upper-left and upper-right quadrants (high ProteinMPNN scores) had similar success rates. Thus, it is clear that sequences with
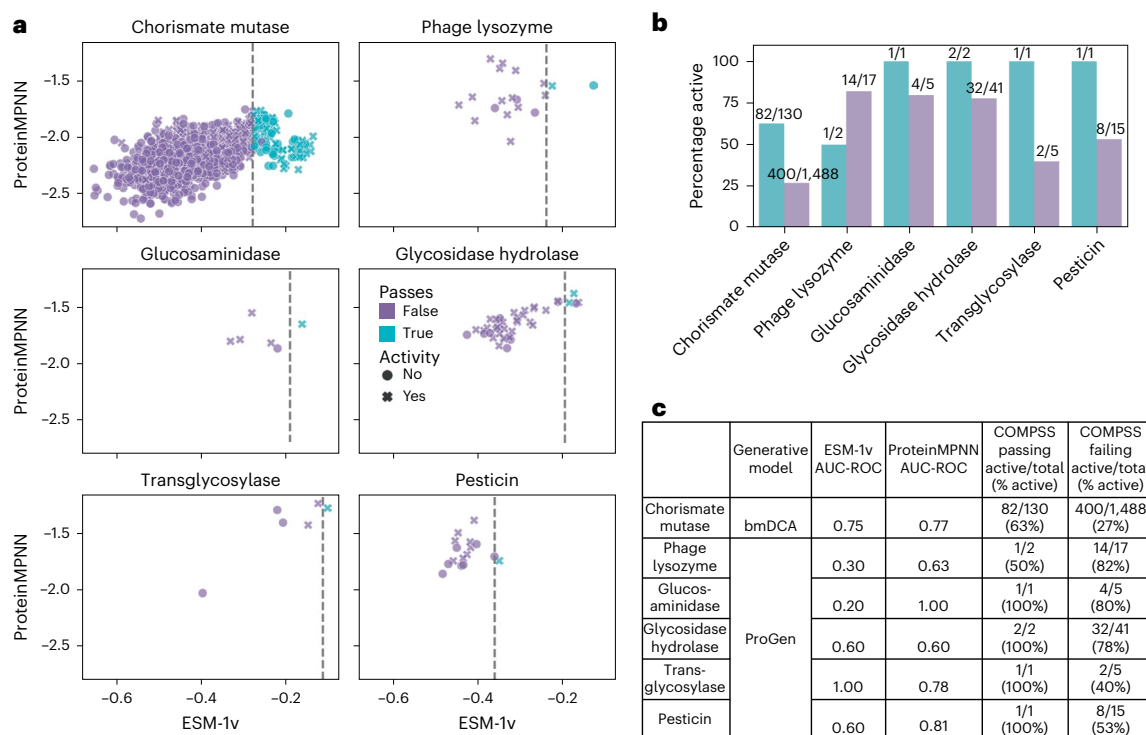
Fig. 4 | Validation of COMPSS against enzymes generated from six previously unseen protein families and two new models. a, ESM-1v and ProteinMPNN scores of generated enzymes from the six families. The vertical dashed gray line indicates the top 10th percentile cutoff for ESM-1v score calculated from natural sequences from the same families. **b**, Proportion of active enzymes stratified by family and whether they passed the COMPSS sequence filter. **c**, COMPSS performance of the six families from publicly available data.

low ESM-1v scores are likely to also have low ProteinMPNN scores and be inactive, confirming that only computing ProteinMPNN scores for sequences with high ESM-1v scores and selecting candidates from the top right quadrant efficiently improves success rates while limiting the number of expensive structure predictions.

The tested lysozyme sequences all passed a stringent filter based on TAPE-BERT and ProGen: there were no negative controls[15]. Thus, we treated them as though they had already passed through a language model filter and were comparable to the two right-hand quadrants (ESM-1v score above the median) of the other datasets and compared the statistics of sequences with ProteinMPNN scores above and below the median (Supplementary Fig. 33). We observed trends similar to those in the half with high ESM-1v scores of the other datasets. Combining the five lysozyme families, sequences with a ProteinMPNN score above the median had a 38% higher success rate than those below the median (39 of 46 versus 27 of 44). Therefore, Madani et al.[15] may have improved their success rate by including a filter based on an inverse folding model. ProteinMPNN provides orthogonal information, increasing the success rate over language model-based selection alone.

We also used the published datasets to assess the effectiveness of an ESM-1v- and ProteinMPNN-based filter on sequences having less than 70% identity to the closest training sequence. Dividing the sequences with less than 70% identity into quadrants, we observed the same trends as in the full dataset (Supplementary Fig. 34), with AUC-ROCs of 0.91 and 0.92 for ProteinMPNN and ESM-1v for the chorismate mutase data, respectively, and 0.67, 0.67 and 0.67 for ProteinMPNN on the three lysozyme datasets with more than one sequence in the bin with less than 70% identity.

Discussion

Discovering new enzymes is difficult because it requires an understanding of the molecular mechanisms, expression and folding while operating within biological and physical constraints. Generative

protein sequence models mimic these constraints by learning to sample from natural sequence distributions. Deep neural networks have led to advances in generative protein design[71], enabling the design of de novo proteins with specified folds[45,72]. Because of the emergent complexity underlying catalysis (Supplementary Table 1), predicting which enzyme sequences will express and fold in soluble and active forms remains challenging, limiting the discovery of novel enzymes. To facilitate enzyme discovery from non-natural sequence spaces, we experimentally evaluated a diverse set of in silico metrics to determine their efficacy in predicting sequence activity. We selected a subset of the metrics to form COMPSS. We validated the filter's effectiveness by prospectively testing synthetic proteins, which resulted in up to a twofold increase in the number of active protein sequences (Extended Data Table 1). Similar results were obtained by independently validating COMPSS on previously published datasets[15,29] of six enzyme families generated by models not trained in this study (Fig. 4).

Our experimentally validated end-to-end framework for generating and selecting new active enzyme variants (Tables 1 and 2) consists of three steps. The first step is curating sequences to obtain a high-quality training dataset. Machine learning is a data-centric practice as much as it is algorithmic; therefore, dataset curation was crucial. Neither ESM-MSA nor the local and much smaller ProteinGAN model performed well in the naive round 1 setting (Extended Data Table 1). In round 2, we observed that removing sequences containing transmembrane domains and signal peptides enriched the dataset for soluble proteins and allowed both models to generate active enzymes at rates above 60%. In round 2, we also selected sequences with a broad range of scores on the metrics and observed which metrics predicted activity (Fig. 2). For the round 3 validation measurements, we combined sequence-based quality checks, the sequence-based ESM-1v metric and the structure-based ProteinMPNN metric to select generated sequences, resulting in enrichment of active sequences (Fig. 3a and Supplementary Fig. 21). While AlphaFold2 (ref. 44) accurately predicts

Table 2 | Overview of the steps and considerations in a typical workflow for generating new active enzyme variants

Step	Description	Examples and considerations
Curate training data	Gather a list of natural sequences likely to have the target activity and express in the target system.	In addition to UniProt and/or NCBI nr, search expanded databases, such as Mgnify for prokaryotic enzymes or NCBI TSA for eukaryotic enzymes.
		Pay attention to the domain content. Unusual domain content indicates neofunctionalization. In some cases, the domain with the activity of interest will retain its function in the absence of the other domains; therefore, it may be safe to remove extraneous domains.
		Pay attention to the presence of localization tags and transmembrane domains. In many cases, these interfere with expression. In some cases, they can be removed without impacting enzyme function.
		Filter out unusually short or long sequences, or sequences with other indications that they may be pseudogenes, fragments or derived from poor gene calling.
		Use hmm-profile or structure searches in addition to sequence searches to find a broader diversity of training sequences.
		Use a clustering algorithm, such as CD-HIT, to reduce the overrepresentation of enzymes from highly sequenced phyla.
Generate new sequences	Use generative models to generate additional members of the enzyme family. Most of these generative models rely on training or fine-tuning of natural sequences curated in the first step of the workflow.	ASR
		Generative adversarial networks (ProteinGAN)
		Language models (such as ProtGPT2, ProGen or ESM-MSA
		VAE
		DCA-based methods
		Inverse folding models (ProteinMPNN, ESM-IF)
Select sequences	Select a subset of natural and generated sequences for experimental evaluation. In campaigns where all sequences are natural or ancestral reconstructions, random selection of candidates may be effective, particularly if care is taken in training data curation. For generative models that produce a lower proportion of active sequences, additional filtering may be required.	Randomly select candidate sequences.
		Select sequences with high similarity to the best candidates from previous screening rounds or the literature (phylogeny-based selection).
		Select sequences with mutations known to be associated with the target phenotype.
		The same curation criteria for natural sequences are also applicable to generated sequences.
		Additional criteria may be used to address failure modes common to the generative models used. For example, models may tend to produce overly short or repetitive sequences.
		Sequences can be scored and ranked based on various metrics. Reasonable scores for these metrics can be estimated from natural sequences. Alternatively, candidates can be selected from the highest-scoring sequences.
		In this study, we settled on a filter composed of six criteria.

protein structures from MSAs[73], the residue confidence pLDDT score does not consistently predict enzyme activity within sets of homologous synthetic sequences (Table 1 and Supplementary Fig. 18). AlphaFold2 predicts high-confidence structures even for inactive sequences (Fig. 2c and Supplementary Fig. 17). Likewise, for sequences with 70–90% identity to natural sequences, neither identity nor sequence similarity captures functional differences (Table 1 and Fig. 2a,c). In contrast, likelihoods from protein language models and inverse folding models moderately predicted in vitro enzyme activity and were weakly correlated with each other (Table 1 and Fig. 2a). Therefore, we used one language model (ESM-1v) and one inverse folding model (ProteinMPNN) in COMPSS.

Our study does not benchmark generative models against each other but instead evaluates metrics to identify those widely applicable across models and enzyme families. Nevertheless, we found that ASR outperforms neural network models in naive generation, indicating room for improvement in protein deep generative models. We observed a wide range of AUC-ROC scores for different combinations of the generative model, enzyme family and metric. Different models and enzyme families may have different failure modes captured by different metrics. Some metrics use underlying models similar to the generative models, which may lead to overfitting. Despite evaluating over 2,200 enzyme variants from eight families, given the sheer size of the protein sequence space, a more extensive dataset could help tease apart the complex interplay between generative models, metrics and protein families and explain that interplay in biologically meaningful terms.

The core idea of COMPSS is to select sequences by prefiltering with fast, biologically motivated quality checks and protein language model scores, followed by a slower step of structure prediction and inverse folding model scoring. Many variations on this core idea are possible, and we do not recommend blindly applying COMPSS to new protein families without considering their biological complexities. In addition to adjusting the ESM-1v score cutoff using natural sequences, biologically motivated quality filters should be chosen on a per-family basis. For example, long repeats or transmembrane domains are required for function in some protein families. The 'starts with M' filter was a good marker of full-length sequences for our dataset because all training sequences started with methionine. In cases where the training sequences have N-terminal truncations, a minimum length filter would similarly eliminate fragments.

We showed that, by carefully curating training data for sequence generation and prioritizing sequences for experimental testing using a multipart filter, as high a proportion as 100% of enzymes with in vitro activity can be achieved, with sequence identities between 70% and 80% to the closest naturally occurring enzymes. We provide Google Colab notebooks for generating new sequences using ESM-MSA and calculating metrics for any user-supplied sequences or structures. Our dataset of more than 500 experimentally tested enzymes and metrics

can serve as reference benchmarks for predicting the function of the generated sequences. The presented end-to-end workflow provides a powerful and flexible framework for generating diverse libraries of active enzymes, enabling deeper explorations of the functional sequence space.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-024-02214-2.

References

1. Bornscheuer, U. T. et al. Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
2. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).
3. Rosa, S. S., Prazeres, D. M. F., Azevedo, A. M. & Marques, M. P. C. mRNA vaccines manufacturing: challenges and bottlenecks. *Vaccine* **39**, 2190–2200 (2021).
4. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Biocatalysis: enzymatic synthesis for industrial applications. *Angew. Chem. Int. Ed. Engl.* **60**, 88–119 (2021).
5. Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed. Engl.* **57**, 4143–4148 (2018).
6. Jäckel, C., Kast, P. & Hilvert, D. Protein design by directed evolution. *Annu. Rev. Biophys.* **37**, 153–173 (2008).
7. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).
8. Orr, H. A. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J. Theor. Biol.* **238**, 279–285 (2006).
9. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210 (2004).
10. Axe, D. D., Foster, N. W. & Fersht, A. R. A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry* **37**, 7157–7166 (1998).
11. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput. Biol.* **11**, e1004421 (2015).
12. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
13. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
14. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
15. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
16. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
17. Repecka, D. et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
18. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
19. Wu, Z., Johnston, K. E., Arnold, F. H. & Yang, K. K. Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
20. Lian, X. et al. Deep learning-enabled design of synthetic orthologs of a signaling protein. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521443 (2022).
21. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
22. Sgarbossa, D., Lupo, U. & Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *eLife* **12**, e79854 (2023).
23. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521521 (2022).
24. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978 (2023).
25. Li, A. J. et al. Neural network-derived Potts models for structure-based protein design using backbone atomic coordinates and tertiary motifs. *Protein Sci.* **32**, e4554 (2023).
26. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
27. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).
28. Foley, G. et al. Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *PLoS Comput. Biol.* **18**, e1010633 (2022).
29. Russ, W. P. et al. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
30. Trinquier, J., Uguzzoni, G., Pagnani, A., Zamponi, F. & Weigt, M. Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.* **12**, 5800 (2021).
31. Tian, P., Louis, J. M., Baber, J. L., Aniana, A. & Best, R. B. Co-evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed. Engl.* **57**, 5674–5678 (2018).
32. Tian, P. et al. Design of a protein with improved thermal stability by an evolution-based generative model. *Angew. Chem. Int. Ed. Engl.* **61**, e202202711 (2022).
33. Schmitt, L. T., Paszkowski-Rogacz, M., Jug, F. & Buchholz, F. Prediction of designer-recombinases for DNA editing with generative deep learning. *Nat. Commun.* **13**, 7966 (2022).
34. Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A. & Clark, P. L. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl Acad. Sci. USA* **117**, 3528–3534 (2020).
35. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).
36. Zrimec, J. et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).
37. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
38. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **40**, 1114–1122 (2022).
39. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems* (eds. Beygelzimer, A., Dauphin, Y., Liang, P. & Wortman Vaughan, J.) **34** (NeurIPS, 2021).
40. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
41. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**, 274–285 (2022).

42. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

43. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS ONE* **8**, e59004 (2013).

44. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

45. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

46. Hsu, C. et al. Learning inverse folding from millions of predicted structures. In *Proc. 39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) Vol. 162, 8946–8970 (PMLR, 2022).

47. Yang, K. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Eng. Des. Sel.* **36**, gzad015 (2023).

48. Rao, R. M. et al. MSA transformer. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) Vol. 139, 8844–8856 (PMLR, 2021).

49. Johnson, S. R., Monaco, S., Massie, K. & Syed, Z. Generating novel protein sequences using Gibbs sampling of masked language models. Preprint at *bioRxiv* https://doi.org/10.1101/2021.01.26.428322 (2021).

50. Wang, A. & Cho, K. BERT has a mouth, and it must speak: BERT as a Markov random field Language model. https://doi.org/10.48550/arXiv.1902.04094 (2019).

51. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* **397**, 1–21 (2016).

52. Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. *Sci. Rep.* **10**, 15493 (2020).

53. Ahn, J. H. et al. Enhanced succinic acid production by *Mannheimia* employing optimal malate dehydrogenase. *Nat. Commun.* **11**, 1970 (2020).

54. Younus, H. Therapeutic potentials of superoxide dismutase. *Int. J. Health Sci.* **12**, 88–93 (2018).

55. Freudl, R. Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.* **17**, 52 (2018).

56. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).

57. Miroux, B. & Walker, J. E. Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **260**, 289–298 (1996).

58. Miller, A.-F. Superoxide dismutases: ancient enzymes and new insights. *FEBS Lett.* **586**, 585–595 (2012).

59. Potter, S. Z. et al. Binding of a single zinc ion to one subunit of copper-zinc superoxide dismutase apoprotein substantially influences the structure and stability of the entire homodimeric protein. *J. Am. Chem. Soc.* **129**, 4575–4583 (2007).

60. Strange, R. W., Hough, M. A., Antonyuk, S. V. & Hasnain, S. S. Structural evidence for a copper-bound carbonate intermediate in the peroxidase and dismutase activities of superoxide dismutase. *PLoS ONE* **7**, e44811 (2012).

61. Kajihara, J., Enomoto, M., Nishijima, K., Yabuuchi, M. & Katoh, K. Comparison of properties between human recombinant and placental copper-zinc SOD. *J. Biochem.* **104**, 851–854 (1988).

62. Kumar, A., Dutt, S., Bagler, G., Ahuja, P. S. & Kumar, S. Engineering a thermo-stable superoxide dismutase functional at sub-zero to >50 °C, which also tolerates autoclaving. *Sci. Rep.* **2**, 387 (2012).

63. Carlioz, A. et al. Iron superoxide dismutase. Nucleotide sequence of the gene from *Escherichia coli* K12 and correlations with crystal structures. *J. Biol. Chem.* **263**, 1555–1562 (1988).

64. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).

65. Wheeler, L. C., Lim, S. A., Marqusee, S. & Harms, M. J. The thermostability and specificity of ancient proteins. *Curr. Opin. Struct. Biol.* **38**, 37–43 (2016).

66. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

67. Keul, F., Hess, M., Goesele, M. & Hamacher, K. PFASUM: a substitution matrix from Pfam structural alignments. *BMC Bioinf.* **18**, 293 (2017).

68. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst.* **15**, 286–294.e2 (2024).

69. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

70. Mitternacht, S. FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res.* **5**, 189 (2016).

71. Ferruz, N. et al. From sequence to function through structure: deep learning for protein design. *Comput. Struct. Biotechnol. J.* **21**, 238–250 (2023).

72. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

73. Hu, M. et al. Exploring evolution-aware &-free protein language models as protein function predictors. In *Advances in Neural Information Processing Systems* (eds. Koyejo, S et al.) **35** (NeurIPS, 2022).

Methods

See the Supplementary Methods for additional details throughout.

Data curation

Round 1 CuSOD. UniProt[74] sequences containing exactly one Sod_Cu Pfam[75] domain were downloaded. Hmmsearch (Hmmer, http://hmmer. org/; ref. [76]) identified the Sod_Cu domain envelopes. Sequences were truncated to remove extraneous sequences beyond the bounds of the Sod_Cu match. Additional quality filtering was performed. Sequence duplicates were removed using CD-HIT[77] at an identity threshold of 80%, and 80% and 20% were randomly sorted into a 'training' and a 'test' set, respectively. A training MSA was generated by an iterative process using MUSCLE (v3.8)[78].

Round 1 MDH. All UniProt sequences containing an Ldh_1_N Pfam domain followed by an Ldh_1_C domain were downloaded. LDH and MDH enzymes, based on enzyme commission number[79], 1.1.1.27 for LDH and 1.1.1.37 for MDH, were downloaded from SwissProt. MUS-CLE and hmmbuild were used to build profile hidden Markov models of both sets. Hmmsearch was used to score each UniProt Ldh_1_ N/Ldh_1_C sequence against the MDH and LDH profiles and sequences that had a stronger match to the MDH profile were retained. Additional processing was performed exactly as with the round 1 CuSOD data curation.

Quantification of domain architectures. See the Supplementary Methods.

Round 2 CuSOD pretest. UniProt CuSOD proteins were obtained as described above (round 1 CuSOD). The kingdom of origin for each sequence was obtained from the UniProt annotation. Transmembrane domains and signal peptides were predicted using Phobius[66]. Sequences with transmembrane domains were discarded. Signal peptides were removed from sequences predicted to contain them. A set of 14 representative CuSOD and 2 FeSOD proteins were manually selected for experimental screening, including eukaryotic, viral and bacterial proteins predicted to not contain signal peptides, and bacterial proteins with predicted signal peptides removed.

Rounds 2 and 3 CuSOD. All eukaryotic transcriptomes available from the NCBI Transcriptome Shotgun Assembly (TSA) sequence database[80] were downloaded. Transdecoder (https://github.com/TransDecoder/ TransDecoder) was used to extract the protein sequences from transcriptomes. Hmmsearch[76] was used to identify proteins with exactly one Sod_Cu domain. This set of proteins was combined with the list of eukaryotic and viral CuSOD proteins from UniProt. Additional quality filtering was performed. Sequences that were more than 85% identical (based on usearch[81] search_global) to a sequence screened in a previous round were discarded. The remaining sequences were deduplicated at 90% using CD-HIT and then split 90% and 10% into training and test groups, respectively. A training MSA was generated.

Rounds 2 and 3 MDH. Hmmsearch[76] was used to search Mgnify[82] for sequences containing exactly one Ldh_1_C and one Ldh_1_N domain. The list of Mgnify proteins was added to the list of UniProt (curation described above). Additional quality filtering was performed. Sequences were deduplicated at 90% using CD-HIT. Sequences with identity greater than 85%, based on usearch search_global, to a sequence experimentally screened in round 1 were discarded. The remaining sequences were split into training (90%) and test (10%) sets. A training MSA was generated.

Phylogenetic trees. Trees were constructed using FastTree from MSAs generated by MAFFT[83]. Trees were rooted and the midpoint was rendered using ETE3 (ref. [84]).

Chorismate mutase and lysozymes. See the Supplementary Methods.

Generative models

ESM-MSA-1b sampling. Sequences were generated by iterative masking and sampling using the ESM-MSA-1b model[48]. ESM-MSA-1b is a neural network model trained to fill in the wild-type amino acids in masked positions of a protein MSA. The model can be used to generate new sequences by running MSA masking and prediction iteratively, each time replacing the wild-type amino acids at the masked positions with an amino acid drawn from the probability distribution returned by the model. The use of masked language models to generate new sequences was first proposed by Wang and Cho[50], and the strategy has been applied to protein sequences in at least three prior works[22,49,85].

See the Supplementary Methods for more detail on the parameters used.

ProteinGAN. Generative adversarial models were trained using the training sets for CuSOD and MDH. Then, for each family, sequences were generated by sampling vectors from the latent space using a truncated normal distribution. For rounds 1 and 2, 10,048 sequences were generated for each family. For round 3, 560,016 and 160,064 sequences were generated for CuSOD and MDH, respectively.

Ancestral sequence reconstruction. Maximum-likelihood trees were generated from the training set reference MSAs using FastTree[86]. Ancestral sequence reconstructions were generated from the trees using the joint reconstruction function of the GRASP[28] command line tool. Metrics were calculated, and candidates were selected from the entire set of reconstructed sequences.

ProGen. See the Supplementary Methods.

Computational metrics

AlphaFold2. AlphaFold2 (ref. [44]) was used to predict the structures of test sequences and all generated sequences that passed the first filtering step.

Phobius. The jphobius[66] (https://phobius.sbc.su.se/data.html) executable was used to predict the presence of signal peptides or transmembrane domains.

ESM-1v and CARP-640M. Scores calculated from the ESM-1v[39] and CARP-640M[68] models were the average of the log probabilities of the amino acid in each position. Without masking, this calculation can be done with a single forward pass over each sequence. With partial masking, it can be done in a number of passes equal to one per masked_fraction.

ESM-MSA. Scores from the ESM-MSA-1b[48] model were calculated in a manner similar to that for ESM-1v scores, using the average log probability across the whole sequence. The metric was calculated using phmmer[76] to find the 31 closest training sequences to each query, align the 32 sequences with MAFFT and calculate the average log probabilities from six passes with a masking interval of six.

ProteinMPNN, ESM-IF and MIF-ST. The proteinMPNN[45] and ESM-IF[46] scores are the average log likelihood of the query residues using the AlphaFold2-predicted structure. The MIF-ST[47] score was calculated using the extract_mif.py script from the protein sequence models repository (https://github.com/microsoft/protein-sequence-models).

Rosetta-relax. The Rosetta (v2020.08.61146)[43] relax program was used to relax the AlphaFold2 structures.

Distance to the closest training sequence. The most similar training sequence was found using ggsearch36 from the FASTA package[87], the BLOSUM62 scoring matrix and a gap open penalty of 10 and gap extend penalty of 2. The Hamming distance was then calculated from the gapped alignment between the query and the top hit sequences. Identity was calculated as $1 − Hamming\_distance$.

BLOSUM62 and PFASUM15 mutant position mean. The closest training sequence was found using ggsearch36 as described above. From the alignment to the closest training sequence, the mean BLOSUM62 score[37] across all mismatched positions was calculated, ignoring positions where either the query or the reference had a gap. We also calculated the alignments and scores using an alternative matrix, the PFASUM15 matrix[67].

Longest repeat. Scores were calculated for the longest single-amino acid repeat and the longest 2-mer, 3-mer and 4-mer repeat in each sequence. The scores were calculated as $−1 ×$ the number of repeat units. Therefore, the sequence AAAAAA would have a single-amino acid repeat score of $−6$, a 2-mer score of $−3$, a 3-mer score of $−2$ and a 4-mer score of $−1$. The sequence LALALALA would have a 1-mer score of $−1$, a 2-mer score of $−4$, a 3-mer score of $−1$ and a 4-mer score of $−2$.

SASA. SASA, polar SASA and apolar SASA were calculated from the AlphaFold2-predicted structures using the freesasa package (https://freesasa.github.io/). The percentage of polar SASA was calculated using the formula $100 ×$ polar SASA/SASA.

Net charge, Abs(net charge) and charged fraction. Charges were calculated by summing the numbers of glutamate and aspartate residues and lysine and arginine residues for negative and positive charges, respectively.

Avg(phmmer top 30). The phmmer top 30 average score was calculated by running a phmmer search of the experimentally tested sequences against the training sequences and averaging the scores of the top 30 hits.

Selection of sequences for in vitro assays
Round 1. The selected sequences had 70% and 80% identity to the closest training sequence and diverse scores on the ESM-1v metric.

Round 2 pretest. CuSOD sequences were selected on the basis of the kingdom of origin (eukaryotic, viral or bacterial) and the presence of Phobius-predicted signal peptides. Sequences with predicted signal peptides were truncated at the predicted signal peptide cleavage site. Two bacterial FeSOD proteins, both lacking a predicted signal peptide, and the previously characterized[63] *E. coli* FeSOD (as a positive control) were also assayed.

Round 2. The selected sequences had between 80% and 90% identity to the closest training set sequence and diverse scores on the ESM-1v and ESM-MSA metrics. Sequences were also filtered by manual inspection to remove those with large insertions or deletions compared to the closest reference sequences or long repeats, and a methionine was added to the start of a few of the sequences.

Round 3. Sequences were selected on the basis of a series of filters. The first filter removed sequences having (1) less than 50% or greater than 80% identity to the closest training sequence; (2) an ESM-1v score below the top 10th percentile threshold compared to the test sequences; (3) no starting methionine; (4) a predicted transmembrane domain; and (5) a single-amino acid repeat longer than three amino acids or an amino acid pair repeat longer than four amino acids, as repeats were more common in ESM-MSA-generated sequences

than in natural sequences (Supplementary Fig. 35). For each enzyme family, 200 ESM-MSA-generated sequences and 200 GAN-generated sequences were randomly selected from the sequences that passed the first filter, and their structures were predicted with AlphaFold2. ProteinMPNN scores were calculated for each structure, and the 40 sequences with the highest scores from each model–enzyme combination were retained. Of the top 40 sequences, 18 were randomly selected for expression and functional characterization. For each passing sequence that was selected for functional characterization, a corresponding control sequence was selected from the list of sequences that failed the sequence filter. Control sequences were identical to the closest training sequence within 1% of the passing sequence.

Newly generated ProGen lysozyme sequences. See the Supplementary Methods.

Experimental assays
Bacterial strains, plasmids and growth conditions. *E. coli* BL21(DE3) was used as the host strain for MDH and SOD expression in this study. Cells were grown on LB medium at 37 °C and supplemented with 100 µg ml⁻¹ ampicillin (cat. no.171254, Merck).

Sequences were optimized based on *E. coli*-preferred codons using the Twist Bioscience web interface (www.twistbioscience.com). A 30-bp sequence (TTTGTTTAACTTTAAGAAGGAGATATACAT) composed of ribosomal binding site sequences and a spacer were added at the 5′ terminus of all genes. Genes were ordered from Twist Bioscience as clones in pET-21(+) between the EcoRI and NotI sites.

The pET21b plasmid harboring the MDH4 gene from a previous study[17] was used as a positive control for MDH enzymes. Human SOD1 (ref. 61) (hSOD, GenBank: NP_000445.1), *Potentilla atrosanguinea* CuSOD[62] (paSOD, GenBank: AFN42318.1) and *E. coli* SOD[63] (E.SOD, GenBank: NP_416173.1) were codon optimized, synthesized as described above and used as positive controls for SOD enzymes. Blank plasmid pET21b was used as a negative control for both MDH and SOD enzymes.

Plasmid construction for truncated control sequences. See the Supplementary Methods and Supplementary Table 7.

Competent cell preparation and plasmid transformation. Competent cells of *E. coli* BL21(DE3) were prepared using the calcium chloride method[88].

See the Supplementary Methods for details.

Protein expression and purification. Protein expression was achieved by diluting the overnight cultures 1:30 into 2.5 ml autoinduction Terrific Broth (TB) medium including trace elements (cat. no. AIMTB0210, Formedium) and supplemented with 100 µg ml⁻¹ ampicillin in a 24-well format. All cells were cultivated in 24-well plates in an Eppendorf ThermoMixer C. For MDH expression, cells were grown for 4 h at 37 °C, followed by overnight growth at 16 °C while shaking at 200 rpm. For SOD expression, cells were grown for 4 h at 37 °C, followed by another 3 h at 25 °C with shaking at 200 rpm.

Cells were collected by centrifugation at 3,000*g* for 10 min. Cell pellets were suspended in 200 µl BugBuster reagent (cat. no. 70584, Merck) supplemented with 1 µl 2,000 U ml⁻¹ DNase I (cat. no. 79254, Qiagen) and incubated at 37 °C with shaking at 200 rpm for 30 min. After incubation, 10-µl mixtures were aliquoted and kept in −20 °C as the total protein (T) sample for gel electrophoresis. The mixture was centrifuged at maximum speed for 10 min and the pellets were discarded. Then, 10 µl of the supernatant was aliquoted and kept at −20 °C as the soluble protein (S) sample for gel electrophoresis. The supernatants were used for protein purification using the following procedures.

Talon resins (cat. no. 635653, Takara Bio) were washed twice with a binding buffer (50 mM NaH₂PO₄, 300 mM NaCl, 10 mM imidazole,

pH 7.4) and then suspended in the same volume of binding buffer as the resin bed amount. Talon resin (50 µl) was loaded into Pierce microspin columns (cat. no. 89879, ThermoFisher). Each supernatant sample was added to the loaded column and incubated at 4 °C for 30 min in a thermomixer.

The columns were then centrifuged at 20g for 30 s and the flow waste was discarded. Resins were washed with 600 µl of wash buffer three times (50 mM NaH$_2$PO$_4$, 300 mM NaCl, 20 mM imidazole, pH 7.4) and centrifuged at 20g for 30 s each time. Finally, the resins were incubated with 100 µl of elution buffer at 4 °C for 30 min in a thermomixer and proteins were then eluted with centrifugation at 20g for 1 min. Another 100 µl of elution buffer was added to repeat the elution steps, and the two portions of elutions were individually mixed. The two eluate fractions were then combined and transferred to a 96-well desalting plate (cat. no. 89807, Thermo Scientific), which was pre-equilibrated with the sample buffer (50 mM NaH$_2$PO$_4$, 300 mM NaCl, pH 7.4). Protein samples were kept at −80 °C after adding 1× protein-stabilizing cocktail (cat. no. 89806, Thermo Scientific). Then, 10 µl of the proteins was aliquoted and kept at −20 °C as the purified protein (P) sample for gel electrophoresis.

For enzymes from round 2 and round 3 and the truncated enzymes from the round 2 pretest, protein concentrations were measured by Qubit Protein Assay (cat. no. Q33211, Thermo Scientific).

Gel electrophoresis. Total, soluble and purified proteins of each sample were mixed with 1× loading buffer (4× loading buffer recipe: 0.2 M Tris-HCl, 0.4 M DTT, 277 mM SDS, 6 mM bromophenol blue, 4.3 M glycerol) and then heated at 85 °C for 5 min in a PCR cycler. Denatured proteins were analyzed by SDS–PAGE with precast gels (cat. no. WG1403A, Thermo Scientific), followed by Coomassie staining with InstantBlue (cat. no. ISB1L-53, Kem-en-tec). Spectra multicolor broad-range protein ladder (cat. no. 26634, Thermo Scientific) was also loaded to analyze the protein sizes.

Enzymatic assay. To test for MDH activity, 2 µl or 100 µg ml^{-1} of purified protein in round 1 was added to a reaction mixture containing approximately 1.5 mM NADH (cat. no. 10128023001, Merck), 2.0 mM oxaloacetic acid (cat. no. O4126, Sigma) and 20 mM HEPES buffer (pH 7.4). Assays were performed in triplicate in a 96-well format. All components were added using multichannel pipettes to avoid the reaction time lag of each well. The final reaction volume was 100 µl, and the reaction was carried out at room temperature in a transparent 96-well microplate (cat. no. 0020821, Sarstedt). MDH activity was measured in triplicate by following NADH oxidation to NAD$^+$, with an absorbance reading at 340 nm performed in kinetics mode for 15 min in a BMG Labtech SPECTROstar nano spectrophotometer. Unspecific oxidation of NADH was monitored in the no-substrate controls, and these values were subtracted from the other samples. Conversion from absorption values to NADH concentration was carried out using Beer–Lambert law $c = A/(d × ε)$, in which the extinction coefficient $ε$ value is 6.22 mM^{-1} cm^{-1}, and the path length for 100 µl in a 96-well plate (d) is 0.29 cm. For samples that did not show any catalytic activities, a tenfold volume, which is 20 µl of purified proteins, was used to perform the assay for a second time.

For MDH in round 2 and round 3, 20 µg ml^{-1} enzymes together with the positive-control MDH4 were used in the assay as described above for quantitative comparison of catalytic activities, excerpt for samples 1564 and 1546 from round 2, for which the concentration of 0.2 µg ml^{-1} was used due to low protein yields.

SOD activity was measured with a SOD assay kit (cat. no. 19160, Sigma) in a 96-well format, and all components were added using multichannel pipettes to avoid the reaction time lag of each well. For SOD from round 1, an aliquot (2 µl) of purified protein was added to each well containing 98 µl working solution. Assays of each sample were performed in triplicate and in one 'No XO' well. xanthine oxidase

working solution (10 µl) was added to each well at the end, except for the 'No XO' wells. 'No SOD' and 'blank' assays were also performed in triplicate. 'No SOD' wells contained 10 µl dilution buffer, 80 µl working solution and 10 µl xanthine oxidase working solution, while 'blank' wells contained 20 µl dilution buffer and 80 µl working solution. Plates were incubated in the plate reader, which was preset at 37 °C. Absorbance at 450 nm was measured in the kinetics mode for 30 min. For proteins that did not show any catalytic activity, a tenfold volume of 20 µl of purified proteins was used to perform the assay a second time.

For SOD from round 2 and round 3, 5 µg ml^{-1} of enzymes were used in the assay as described above for quantitative comparison of catalytic activity.

To assay the truncated proteins, 85 µg ml^{-1} of all samples were used in the enzymatic assay.

For details on the lysozyme assays see the Supplementary Methods.

Data analysis
For MDH, the absorbance value was plotted over time. The absorbance values of all samples at the endpoint of the assay were compared to the negative control by t-test analysis. Samples were considered active if the end absorbance value was significantly lower than that of the negative control, $P ≤ 0.05$.

For SOD, enzyme activity was measured as the percentage inhibition of the rate of WST-1 formazan formation and calculated using the following equation with absorbance value at 20 min. The inhibition rate was compared to the negative control by the t-test, and those with activity significantly higher than the negative control were considered active with $P ≤ 0.05$.

SOD activity (inhibition rate %) = $((A − B) − (C − D))/(A − B) × 100$, where A is the absorbance value of the 'no SOD' control, B is the absorbance value of the blank, C is the absorbance value of the sample and D is the absorbance value of the 'no XO'.

Assay data were analyzed using GraphPad Prism v8.0.0 for Windows, GraphPad Software (www.graphpad.com).

Semiquantitative comparisons of enzyme activities. Data from round 3 enzyme assays using 20 µg ml^{-1} MDH or 5 µg ml^{-1} SOD, as described above, were used for semiquantitative comparisons of enzyme-specific activity (Fig. 3d).

For MDH, MDH4 was used as a wild-type positive control, and for SOD, hSOD, paSOD and E.SOD were used as wild-type positive controls.

For MDH, absorbance at 340 nm was converted to NADH concentration and the average difference in the concentration between the 0 and 90 s time points of the assay was used as a measure of enzyme activity. Some enzymes, including the MDH4 control, converted substrate very quickly, such that most of the substrate was converted before the first time point. Therefore, we replaced any values below 275 µM at time 0 with the mean value from the negative control. Values were averaged over three technical replicates and divided by the average of the MDH4 samples.

For SOD, the inhibition rate (%), calculated as described above, was used as a measure of enzyme activity. Values were averaged over three technical replicates and divided by the average of the hSOD, paSOD and E.SOD samples.

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

Data availability
Training sequences were curated from sequences available from UniProt (release 2022_05) (https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2022_05/knowledgebase/), the NCBI

Transcriptome Shotgun Assembly database (https://www.ncbi.nlm.nih.gov/genbank/tsa/) or Mgnify (2022_05) (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2022_05/). All generated sequences, curated natural sequences, train/test splits, predicted structures, metrics scores, phylogenetic trees and tabulations of experimental results are available on Zenodo (https://doi.org/10.5281/zenodo.7688667)[89].

Code availability
Code for regenerating figures and links to Colab notebooks for calculating metrics and generating sequences using ESM-MSA are available on Github (https://github.com/seanrjohnson/protein_scoring)[90]. Locally executable code for generating sequences from ESM-MSA is also available on GitHub (https://github.com/seanrjohnson/protein_gibbs_sampler)[91].

References
74. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
75. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
76. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
77. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
78. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
79. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
80. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
81. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
82. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
83. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
84. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
85. Hawkins-Hooker, A. & Jones, D. T. MSA-conditioned generative protein language models for fitness landscape modelling and design. In *Machine Learning for Structural Biology Workshop* (NeurIPS, 2021).
86. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
87. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
88. Sambrook, J. & Russell, D. W. Preparation and transformation of competent *E. coli* using calcium chloride. *CSH Protoc.* **2006**, pdb.prot3932 (2006).
89. Johnson, S. R. et al. Computational scoring and experimental evaluation of enzymes generated by neural networks. *Zenodo* https://doi.org/10.5281/zenodo.7688667 (2024).
90. Johnson, S. R., Monaco, S. & Yang, K. K. Protein scoring. *GitHub* https://github.com/seanrjohnson/protein_scoring (2024).
91. Johnson, S. R., Monaco, S., Massie, K. & Syed, Z. Protein Gibbs sampler. *GitHub* https://github.com/seanrjohnson/protein_gibbs_sampler (2024).

Author contributions
A.Z., K.K.Y., X.F. and S.R.J. conceptualized the ideas and experiments. X.F. and C.G. performed the laboratory experiments. S.V., S.M. and S.R.J. performed in silico experiments and wrote the related software. X.F. and S.R.J. analyzed the data. K.K.Y., A.Z., X.F. and S.R.J. wrote and edited the manuscript. All authors read and approved the final manuscript.

Competing interests
Sarah Monaco is currently an employee and shareholder of Invitae. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41587-024-02214-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02214-2.

Correspondence and requests for materials should be addressed to Aleksej Zelezniak or Kevin K. Yang.

Peer review information *Nature Biotechnology* thanks James Fraser and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Extended Data Table 1 | CuSOD and MDH experimental results from all rounds

	Family	Model	Total	Expressed	Soluble	Active	Percent active
Round 1	**CuSOD**	test	17	5	3	0	0
		ASR	18	12	9	9	50
		GAN	18	10	5	2	11
		ESM-MSA	18	12	4	0	0
	MDH	test	17	12	6	6	35
		ASR	18	18	14	10	56
		GAN	18	14	0	0	0
		ESM-MSA	17	13	0	0	0
Round 2 pre-test	**CuSOD**	test	14	12	11	8	57
	FeSOD	test	2	1	1	2	100
Round 2	**CuSOD**	test	13	11	11	7	54
		ASR	18	18	18	15	83
		GAN	18	18	18	12	67
		ESM-MSA	18	18	18	12	67
	MDH	test	18	18	15	14	78
		ASR	18	16	16	13	72
		GAN	18	15	3	2	11
		ESM-MSA	18	17	12	9	50
Round 3	**CuSOD**	GAN-passing	18	16	16	13	72
		GAN-control	18	11	9	8	44
		ESM-MSA-passing	18	18	18	17	94
		ESM-MSA-control	18	11	11	8	44
	MDH	GAN-passing	18	4	4	5	28
		GAN-control	18	0	0	2	11
		ESM-MSA-passing	18	18	18	18	100
		ESM-MSA-control	18	14	12	12	67

Expressed: new visible band on SDS-PAGE gel of total protein compared with empty vector control. Soluble: new visible band on SDS-PAGE gel of soluble protein compared with empty vector control. Active: Measured activity significantly different from empty vector control. In some cases, no bands were visible, but activity was detected. For some rows, the total is less than 18, for the following reasons: in Round 1, three of the genes could not be synthesized by Twist; in Round 2, we selected only 13 CuSOD test sequences because we had already tested some similar natural sequences in Round 2 pre-test.

Corresponding author(s): Aleksej Zelezniak, Kevin K Yang

Last updated by author(s): Mar 12, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequences used in the analyses were obtained from UniProt (release 2022_05), NCBI Transcriptome Shotgun Assembly database, or Mgnify (2022_05).
Data analysis	Assay data was analyzed using GraphPad Prism 8.0.0 Additional custom python scripts for data analysis are available at: https://github.com/seanrjohnson/protein_gibbs_sampler https://github.com/seanrjohnson/protein_scoring

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Training sequences were curated from sequences available from UniProt (release 2022_05) (https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not relevant
Population characteristics	not relevant
Recruitment	not relevant
Ethics oversight	not relevant

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All samples were processed in one batch for protein expression and purification. For enzymatic assays, the sample size was chosen considering practicalities when measuring assays in 96-well plates.
Data exclusions	All data points are included in the study.
Replication	All measurements were done using at least three biological replicates. Biological replicates were performed for each sample for protein expression and purification twice. Successfully purified proteins were continued for the following SDS gel and enzymatic assays. All enzymatic assays were performed with technical triplicates for each sample. All attempts at replication were successful.
Randomization	Not relevant, all measurements within each round were done on the same plate.
Blinding	We did not performed blinding as all samples were treated within the same protocol in 96 well plates without group bias. Experiments of expression, purification and enzymatic assay were blinded until data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology and archaeology
☒	Animals and other organisms
☒	Clinical data
☒	Dual use research of concern

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging