nature biotechnology

Large language models generate functional protein sequences across diverse families

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser & Nikhil Naik

Shuxian Zou

2023-07-20

Outline

- Rational protein design
- Method
- Dataset
- Training
- Results
- Conclusion

Rational protein design

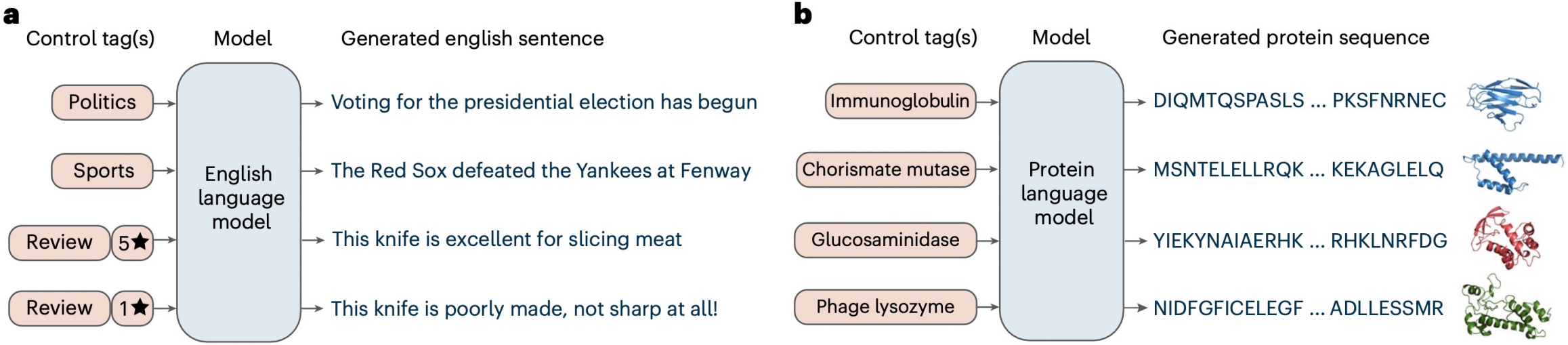
- **Structural-based de novo design method**
  - employ simulations grounded in biophysical principles
  - rely on limited structural data and intractable biophysical simulations

- **Coevolutionary method**
  - build statistical models from evolutionary sequence data to specify novel sequences with desired function or stability
  - tailored to specific protein families, rely on MSA

- **Deep-learning language models**
  - strong representation learning ability
  - can generate artificial protein sequences

ProGen: aim for controllable generation

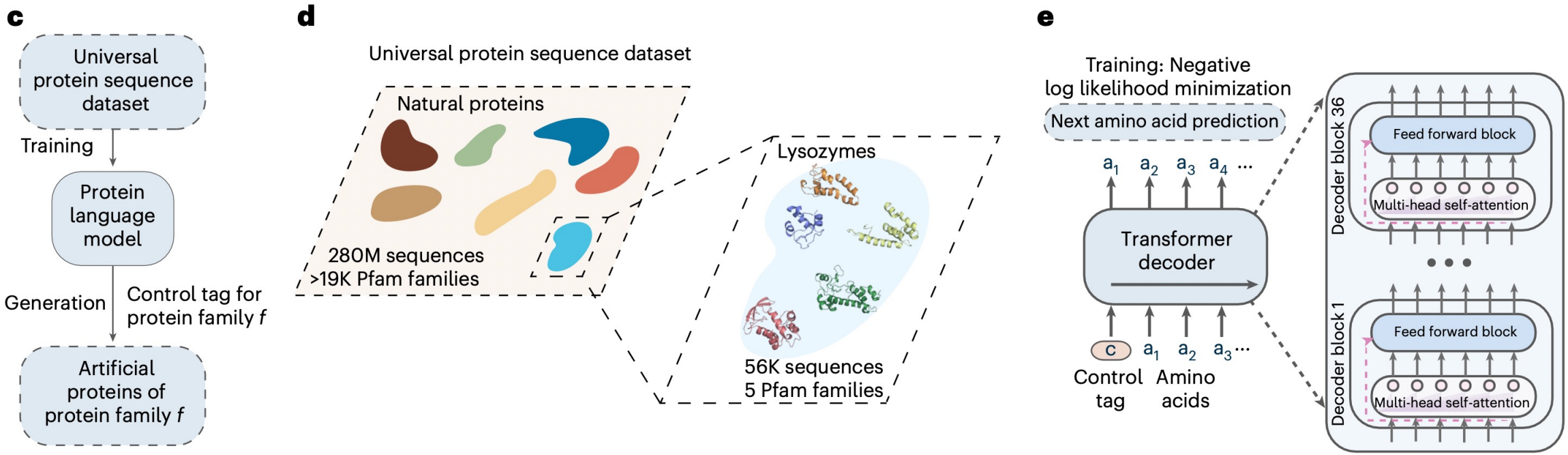
- Conditional protein language model
- Control tags: protein family, biological process, molecular function, etc
- ProGen can be prompted to generate full-length protein sequences for any protein family from scratch
- It can be fine-tuned in a specific protein family to enhance performance



[1] CTRL: a conditional transformer language model for controllable generation. arXiv 2019. 822 citations.

ProGen: 1.2B GPT-like PLM

- Trained on 280 million protein sequences from >19000 families
- Autoregressive language modeling objective
- Transformer model with 36 layers, 8 heads, and 1.2B parameters
- Training was performed across 256 Google Cloud TPU v3 cores for 2 weeks



Datasets

Training 280M seqs

tags

Fine-tuning

	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

Datasets

- **Control tags:**
  - **Keyword tags**: include 1,100 terms ranging from cellular component, biological process, and molecular function terms from UniProtKB
  - **Taxonomic tags**: include 100,000 terms from the NCBI taxonomy across the eight standard taxonomic ranks (Domain, Kingdom, …, Genus, Species)

- **Train/test split:**
  - Training: 280M
  - IID-test: 1M
  - OOD-test: 100K from 20 protein families

Sample inputs

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
```

ProGen training setup

- Include each seq and its reverse
- Prepend each seq with a corresponding subset of control tags
  - average token length of control tags during pretraining was eight
  - for a given seq, there can be multiple versions across databases, each with its own associated control tags
- Include a sample with the seq alone without control tags
- max_seq_len=512, batch_size=2048, 1 million iterations
- Adagrad, linear warmup from 0 to 1×10^{-2} over the initial 40,000 steps, linear decay for the remainder of training
- The model was initialized with pretrained weights of CTRL trained on English corpus

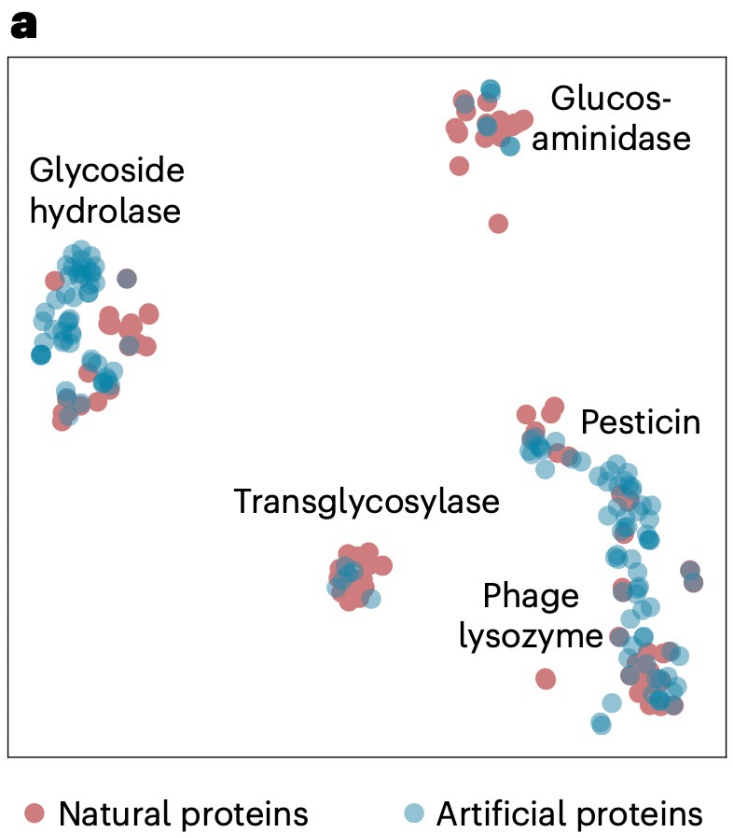
Generated artificial antibacterial proteins are diverse …

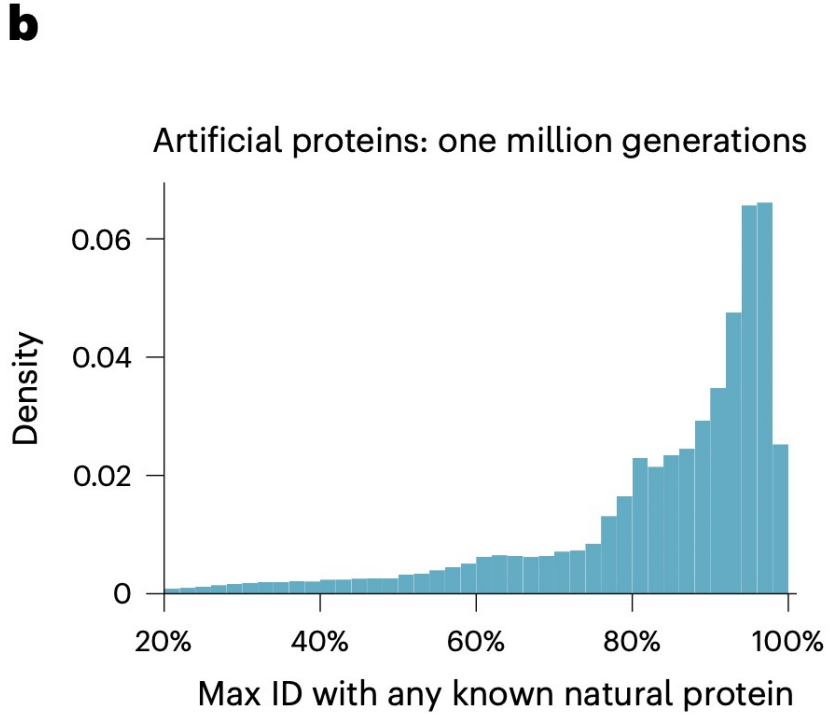
Fine-tune ProGen on 55,948 seqs from 5 lysozyme families

a. Artificial lysozymes span the seq landscape of natural lysozymes across 5 families

b. Generate 1 million artificial seqs using each of the 5 lysozyme families as a control tag, and applying top-*p* sampling → The generated seqs diverge from natural proteins

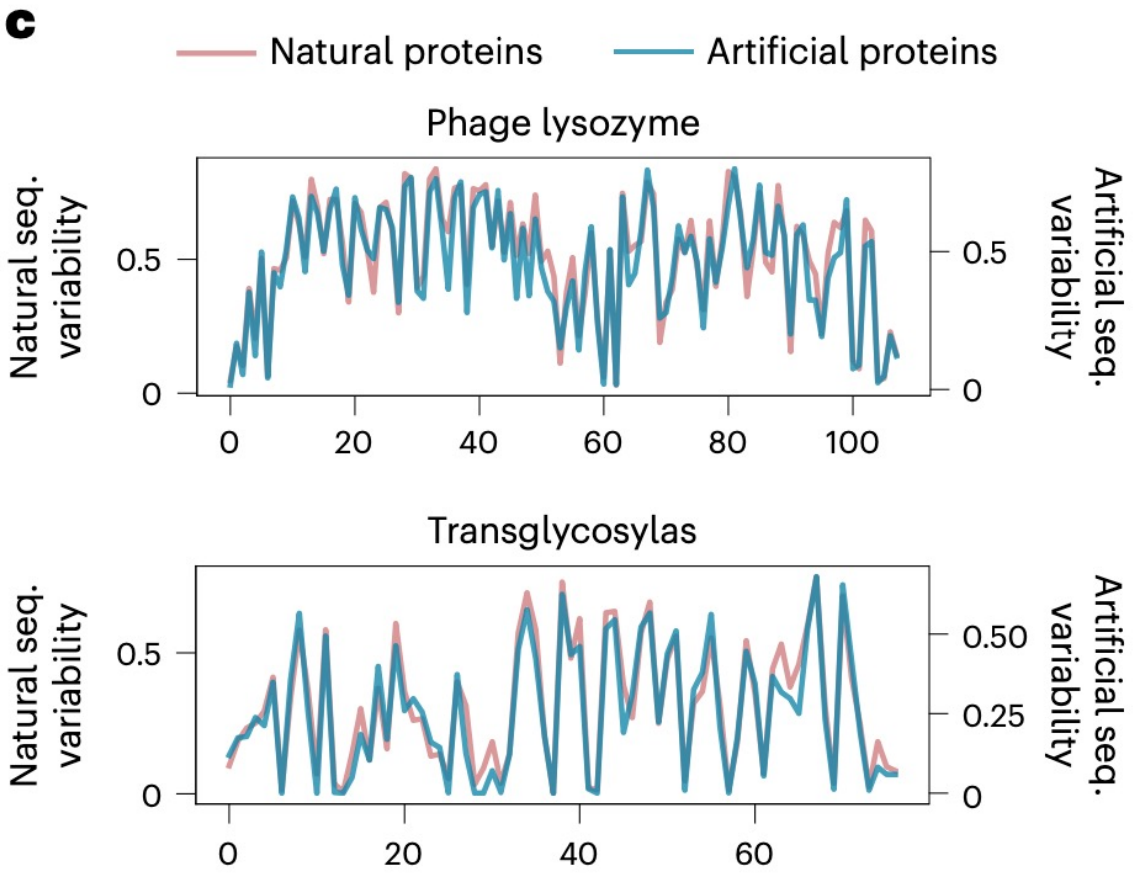
Pfam	#train	#test
Phage	16,488	20
Glyco	5,857	45
Gluco	23,238	8
Trans	9,824	8
Pesti	541	19

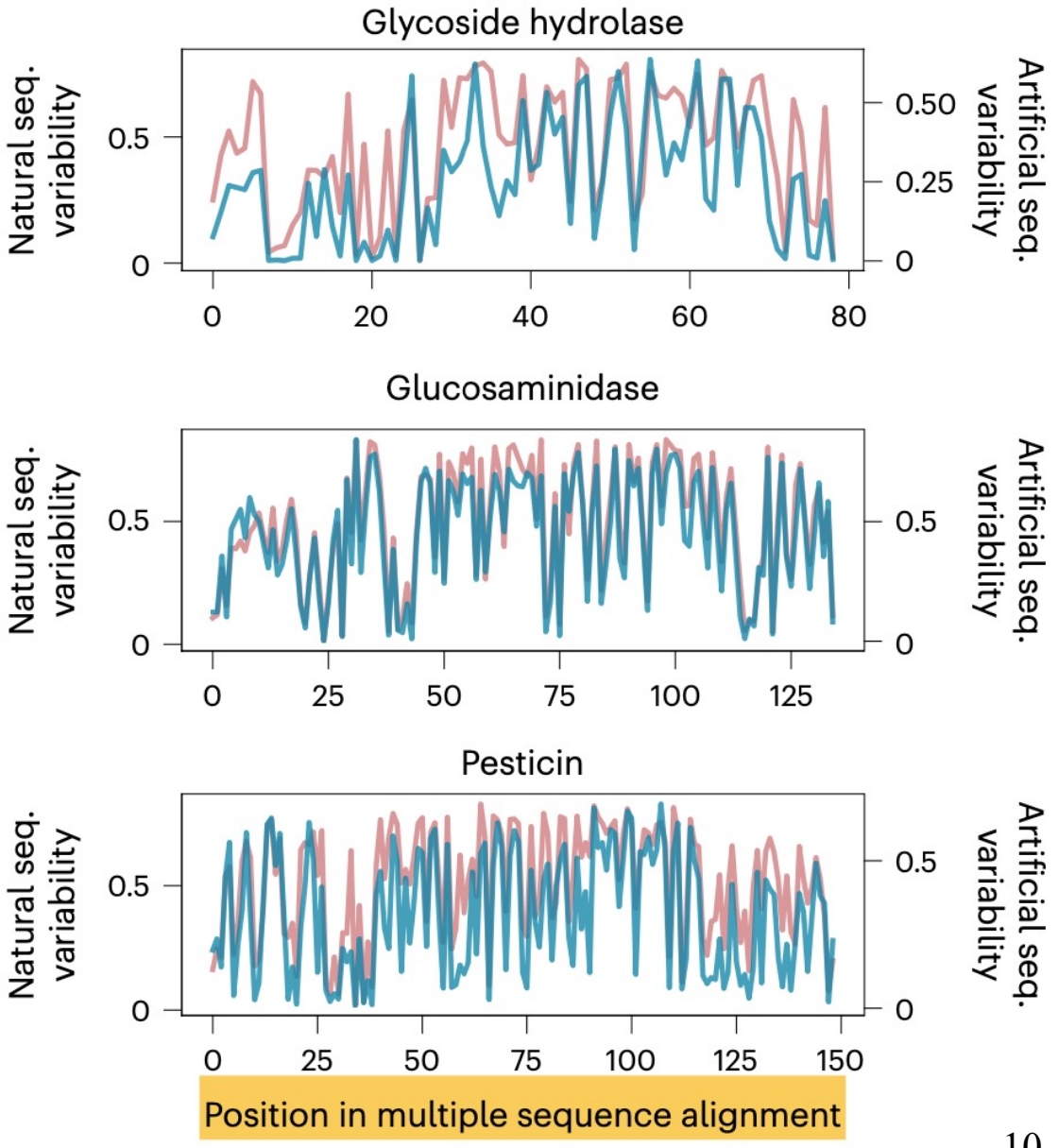




yet maintain similar evolutionary conservation patterns

They demonstrate similar residue position entropies when forming separate MSAs of natural and artificial proteins within each family.

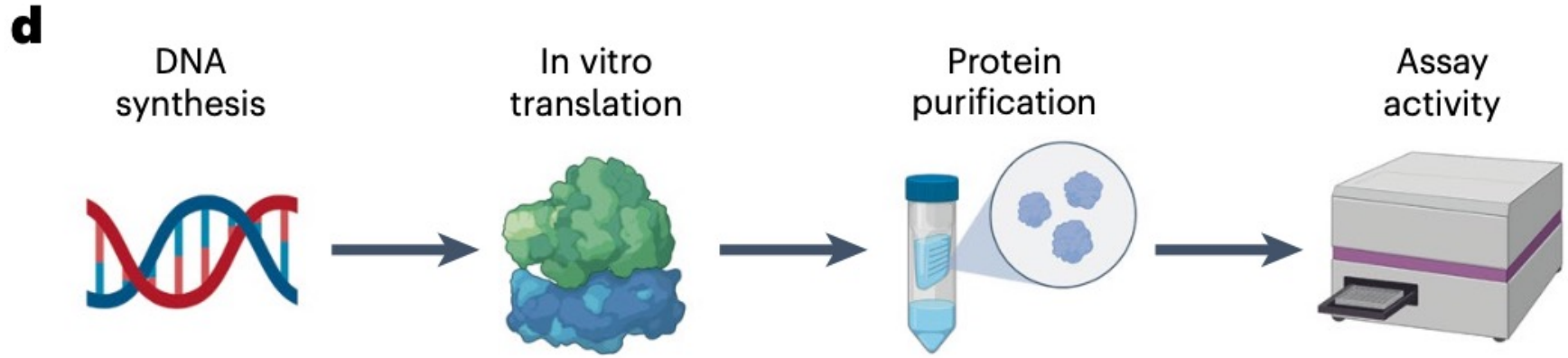




Select 100 artificial lysozyme proteins for wet experiments

Ranking the sequence using the combination of an adversarial discriminator and generative model log-likelihood scoring.

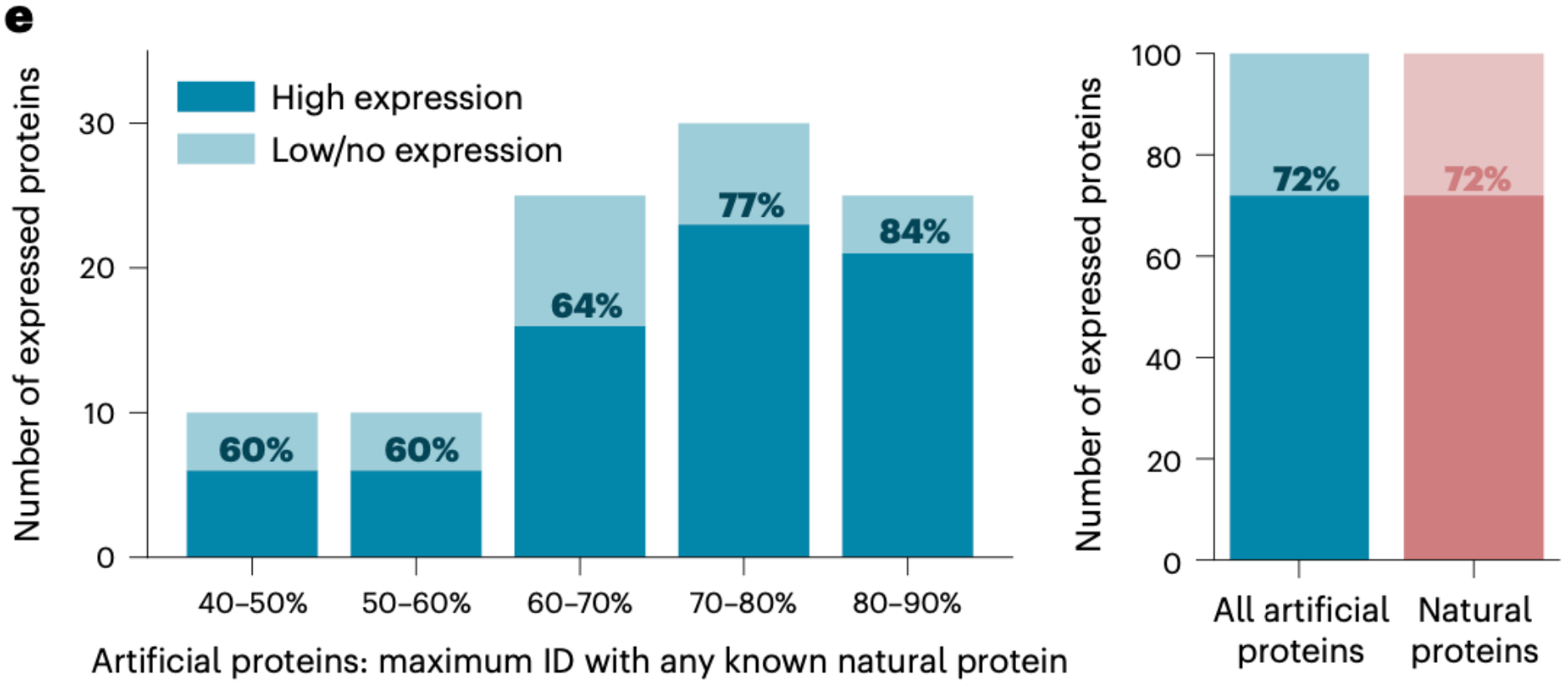
- We trained three adversarial discriminator (TAPE-BERT[1]) to distinguish between natural lysozymes and ProGen-generated lysozymes.
- To train the discriminator, we generated a batch of samples from fine-tuned ProGen that was the same size and distribution of families as our dataset of natural lysozymes.
- We assigned each sequence a discriminator score as the geometric mean of the probability of the sample being a natural sequence as predicted by the three discriminators.



[1] Evaluating protein transfer learning with TAPE. NeurIPS 2019.

Wet experiment results -- expression

- Artificial proteins express well even with increasing dissimilarity from nature (40–50% max ID) and yield comparable expression quality to one hundred representative natural proteins.

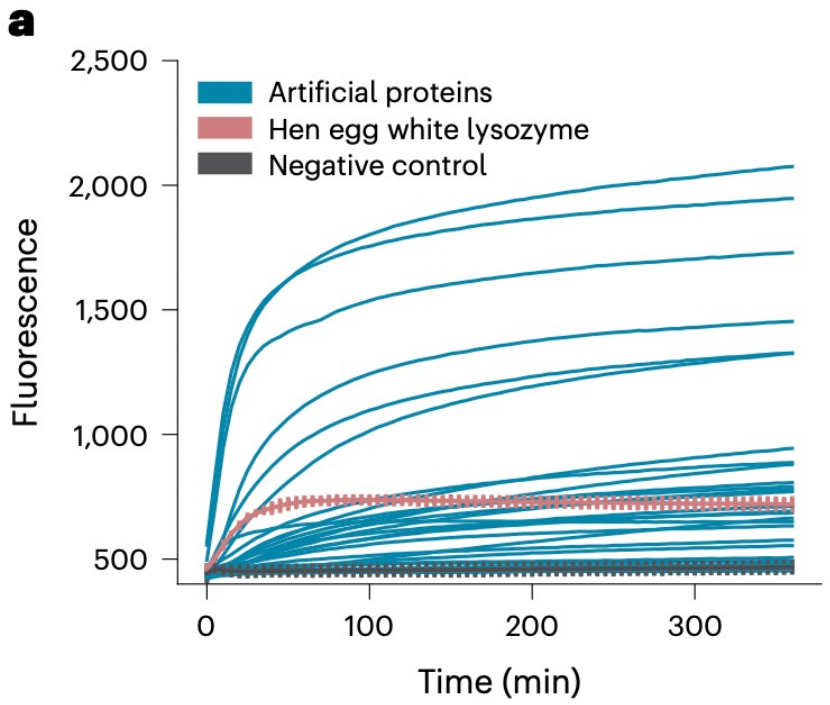


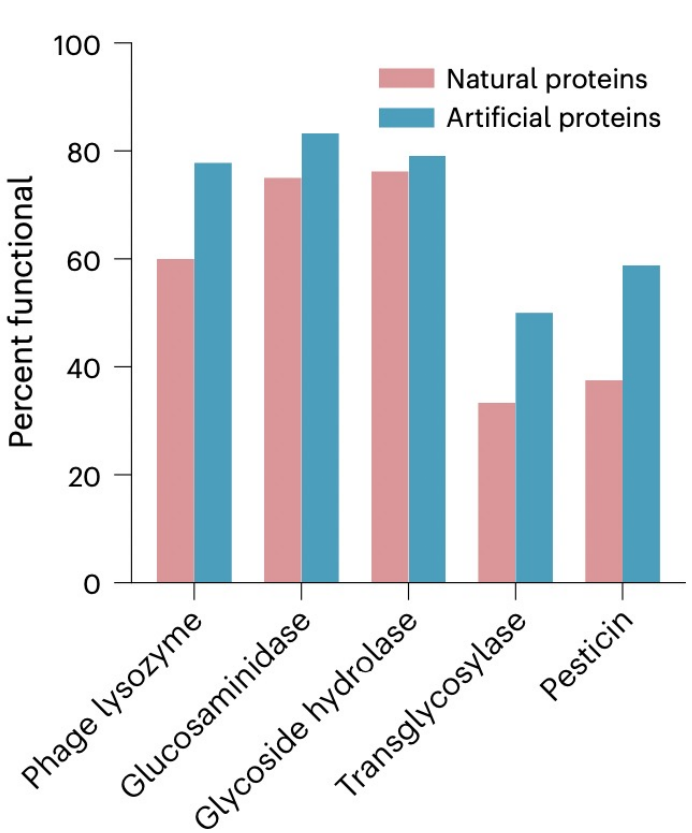
Wet experiment results -- function

a. Artificial proteins exhibit high fluorescence responses over time

c. Artificial proteins are functional across protein families

- Functional is defined as a fluorescence one standard deviation above the maximum value of all negative controls.





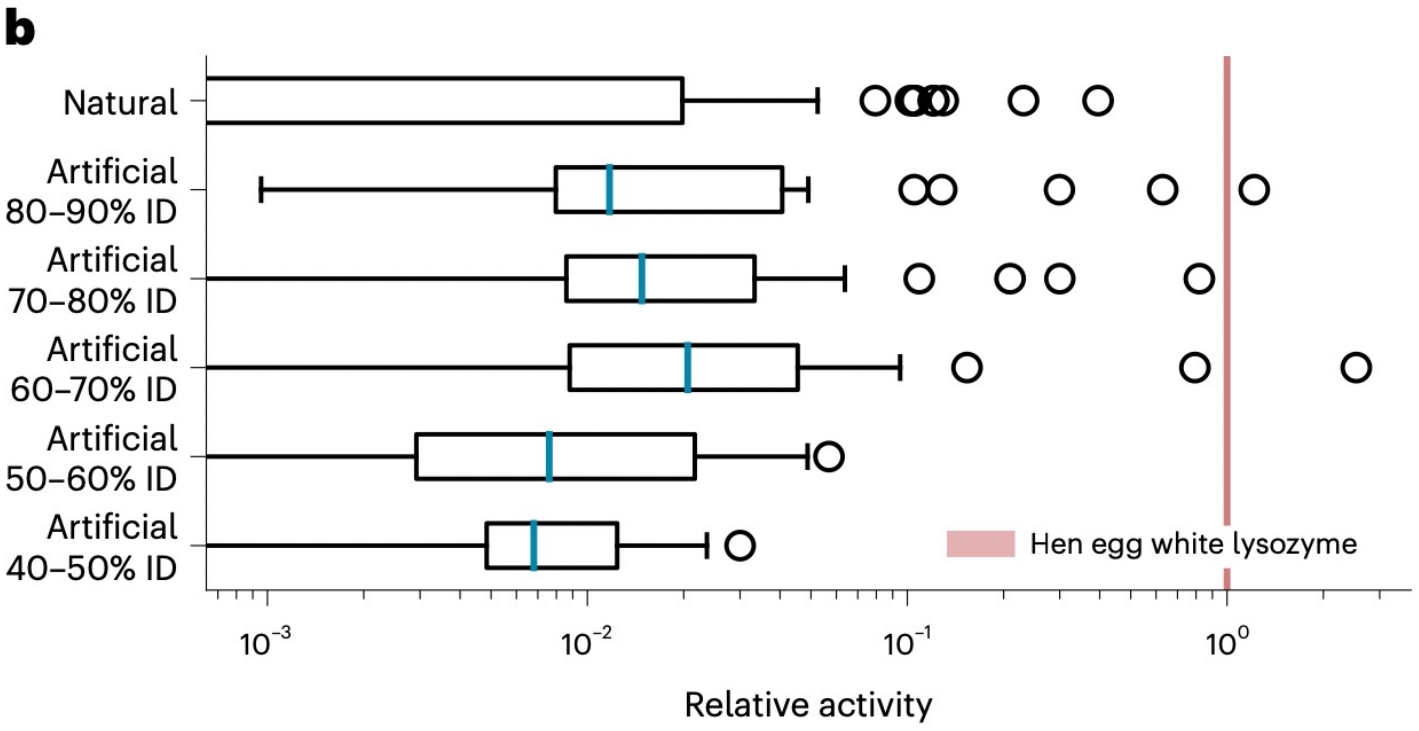
Natural:
59%
53/90

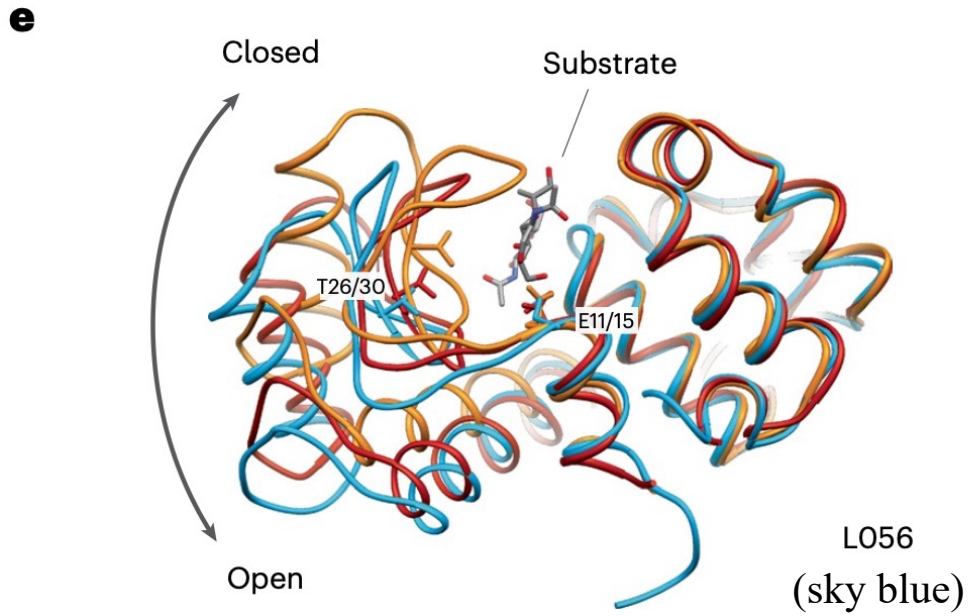
Artificial:
73%
66/90

Wet experiment results -- activity

b. Artificial proteins match activity levels of natural proteins even at lower levels of sequence identity to any known natural protein

e. L056 (max ID 69.6%) express well and incurred bactericidal activities towards the E. coli BL21(DE3) strain





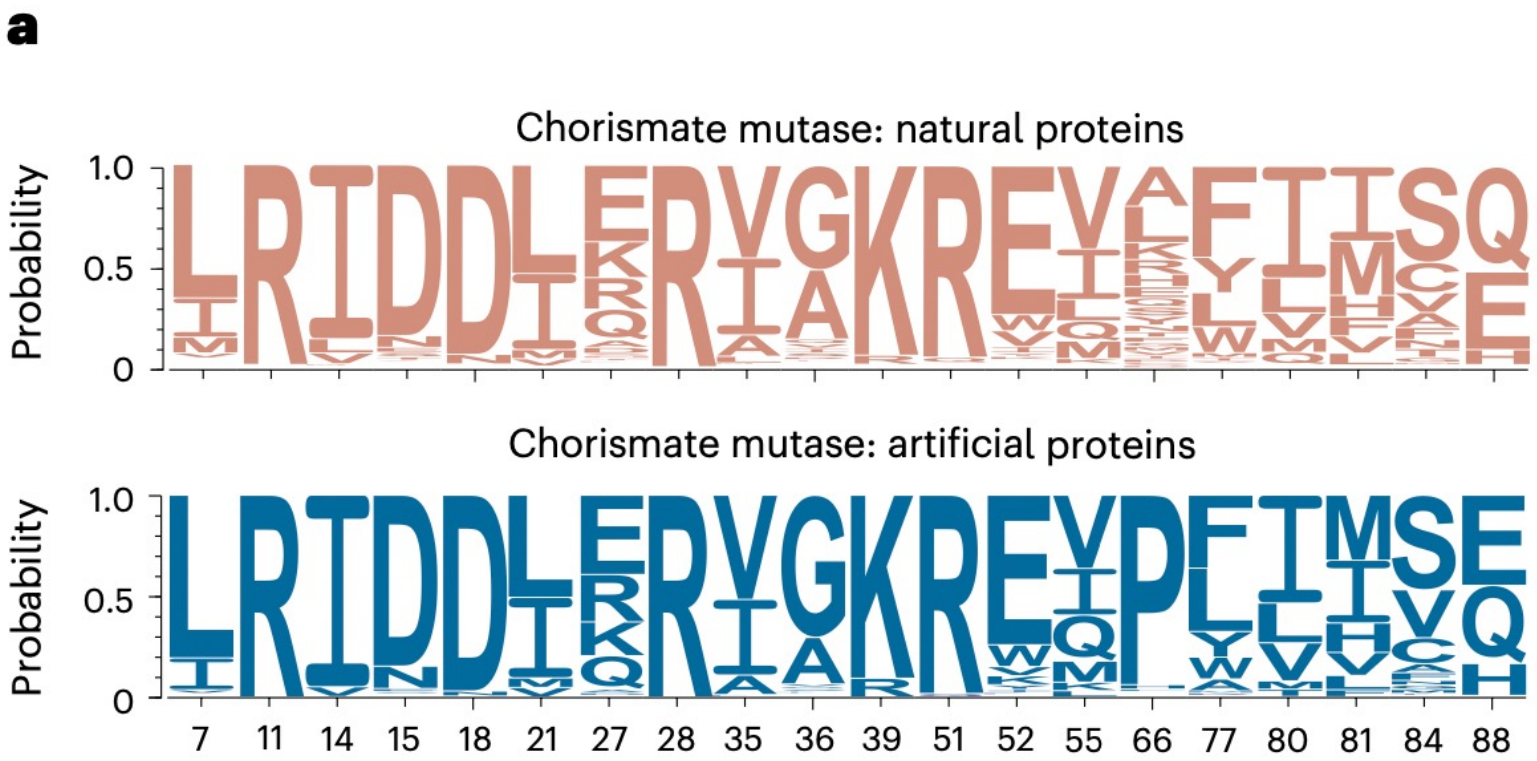
The highly active outliers demonstrate the potential for ProGen to generate sequences that may rival natural proteins.

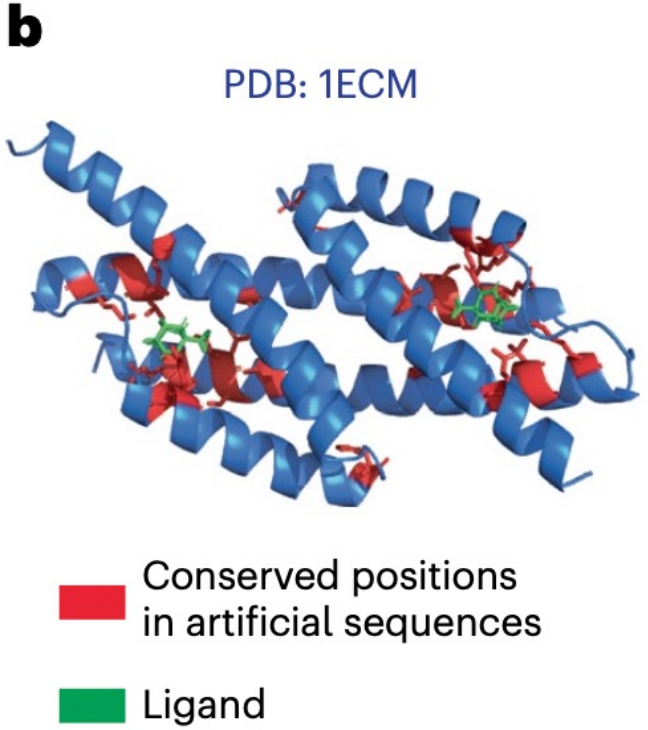
2.5Å resolution crystal of L056 artificial lysozyme, the active site cleft is well formed

Applicability of ProGen to other protein systems

a. ProGen can generate CM enzymes that exhibit a similar residue distribution to nature

b. The conserved residues among generated sequences correlate to ligand-binding sites





Ablation study of pre-training and fine-tuning

- We measured per-token log-likelihoods for artificial sequences using ProGen and used them to predict if artificial sequences should function
- Dataset: previous CM and MDH experimentally measured array data

(AUC)	**Training data**	**CM**	**MDH**
Fine-tuned ProGen	**Universal sequences & CM/MDH**	**0.85**	**0.94**
Pretrained ProGen	Universal sequences	0.18↓	0.08↓
Direct-trained ProGen	CM/MDH	0.11↓	0.04↓

- Both pretraining in the universal sequence dataset and fine-tuning on the protein family of interest contribute significantly to the final model performance.

Conclusion

- ProGen: 1.2B transformer-based conditional protein language model
- Trained only with evolutionary sequence data and can generate functional artificial proteins across protein families
- Use control tags to steer the generation of proteins with desired properties
- Applications of ProGen could include generating synthetic libraries of highly likely functional proteins for discovery or iterative optimization

- Code: https://zenodo.org/record/7296780
- Pretrained checkpoint (4.6GB): https://zenodo.org/record/7309036