bioinforbricklayer
Papers of the week

nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01618-2

Large language models generate functional protein sequences across diverse families

Received: 12 July 2022

Accepted: 17 November 2022

Published online: 26 January 2023

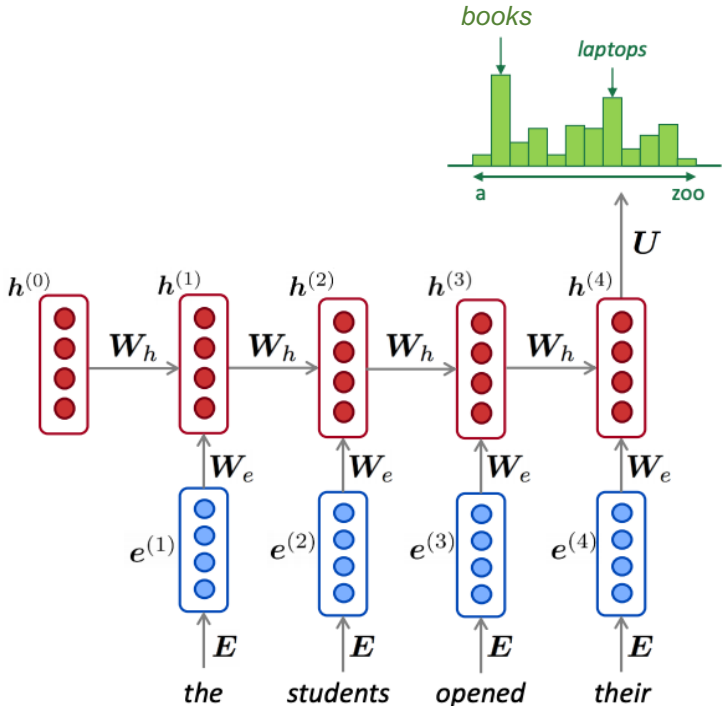
Check for updates

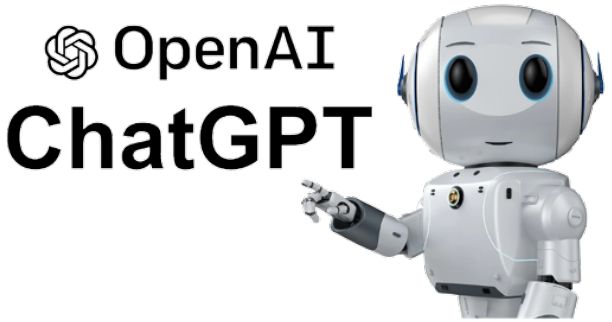
Ali Madani[1,2], Ben Krause[1,10], Eric R. Greene[3,10], Subu Subramanian[4,5], Benjamin P. Mohr[6], James M. Holton[7,8,9], Jose Luis Olmos Jr.[3], Caiming Xiong[1], Zachary Z. Sun[6], Richard Socher[1], James S. Fraser[3] & Nikhil Naik[1]

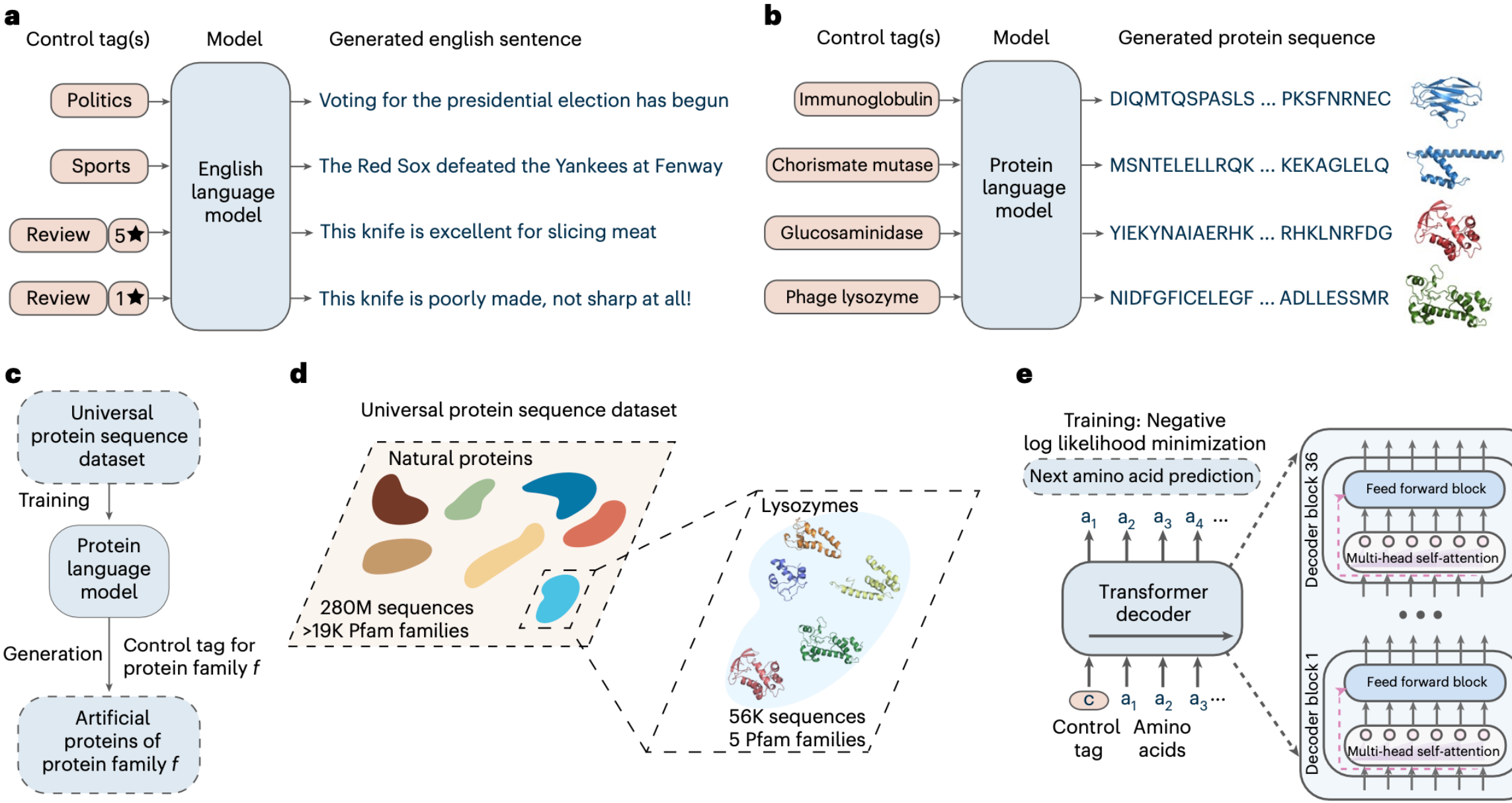
Deep-learning language models have shown promise in various biotechnological applications, including protein design and engineering. Here we describe ProGen, a language model that can generate protein sequences with a predictable function across large protein families, akin to generating grammatically and semantically correct natural language sentences on diverse topics. The model was trained on 280 million protein sequences from >19,000 families and is augmented with control tags specifying protein properties. ProGen can be further fine-tuned to curated sequences and tags to improve controllable generation performance of proteins from families with sufficient homologous samples. Artificial proteins fine-tuned to five distinct lysozyme families showed similar catalytic efficiencies as natural lysozymes, with sequence identity to natural proteins as low as 31.4%. ProGen is readily adapted to diverse protein families, as we demonstrate with chorismate mutase and malate dehydrogenase.

What's large language models?

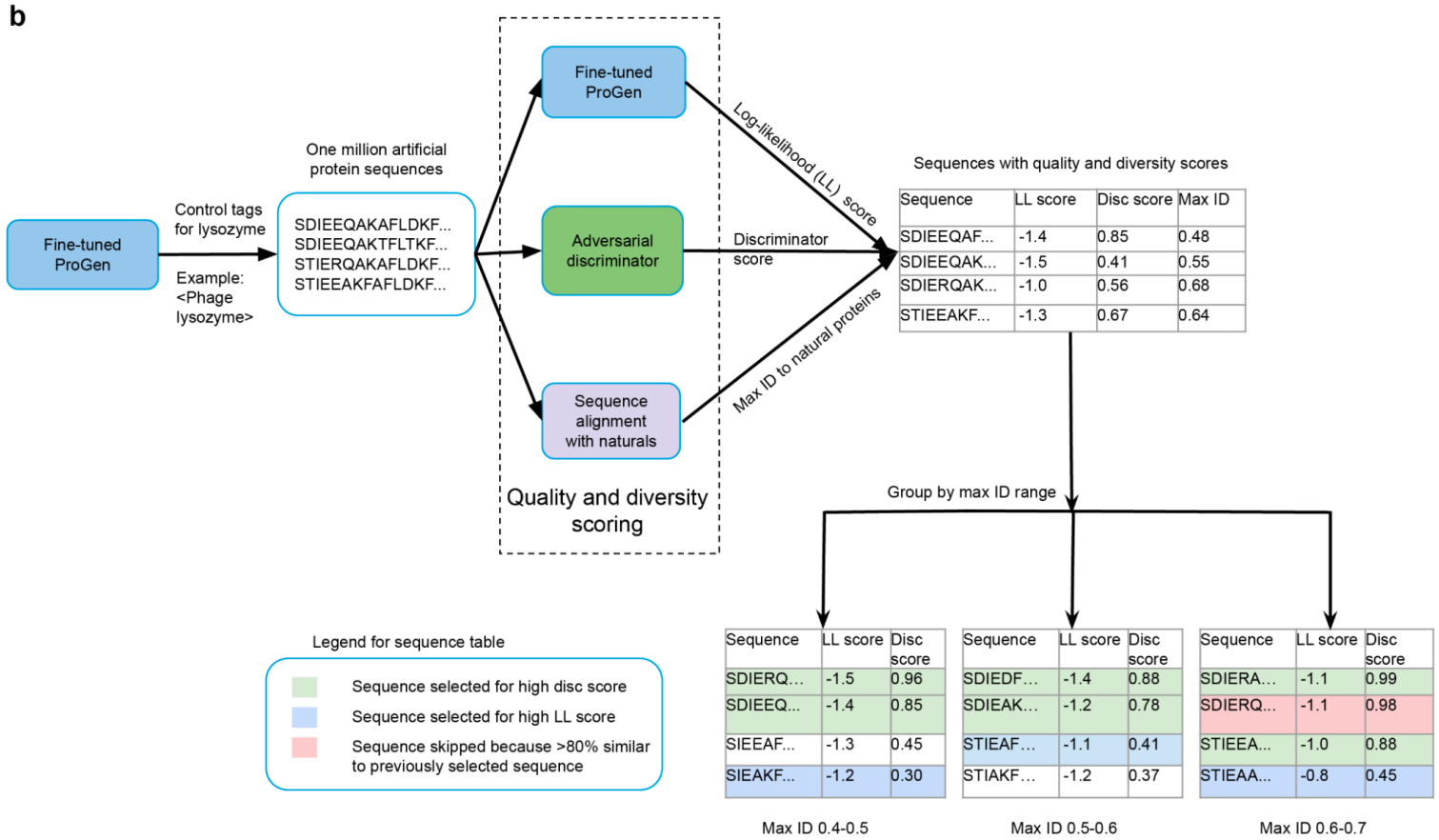
- Large language models are a type of artificial intelligence (AI) system that are trained on vast amounts of text data and can generate natural language responses to a wide range of inputs.
- A deep learning model trained on a large corpus of text data, which enables the model to learn the patterns and structures of language at a very high level of abstraction, allowing it to generate new text that is grammatically correct and semantically meaningful.
- One of the most well-known large language models is GPT-3, which was developed by OpenAI and contains over 175 billion parameters. GPT-3 has been trained on a vast amount of text data, including books, articles, and websites, and can generate responses to a wide range of prompts, such as questions, writing prompts, or even computer code.

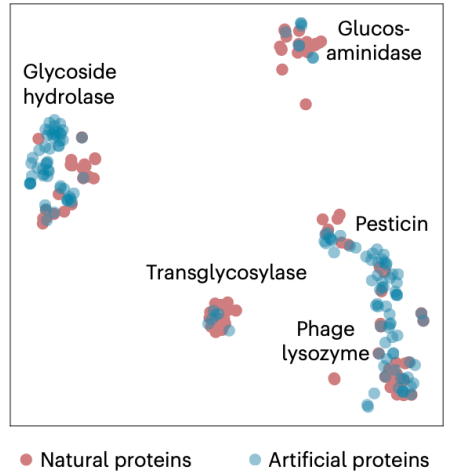


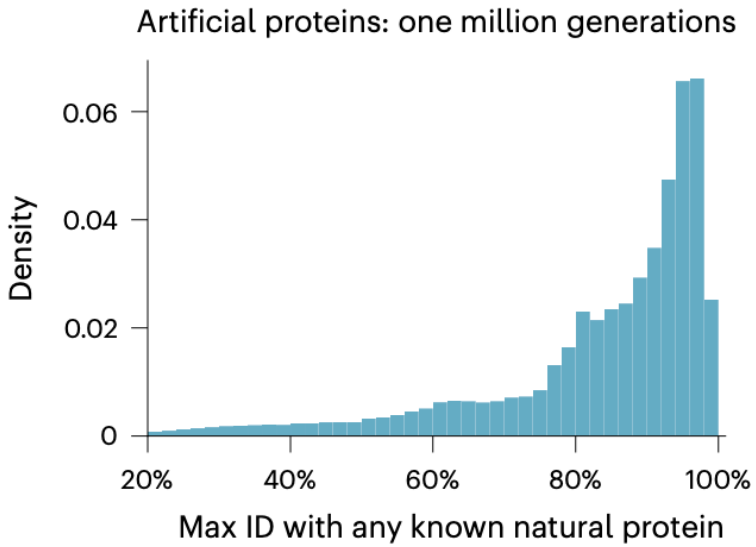


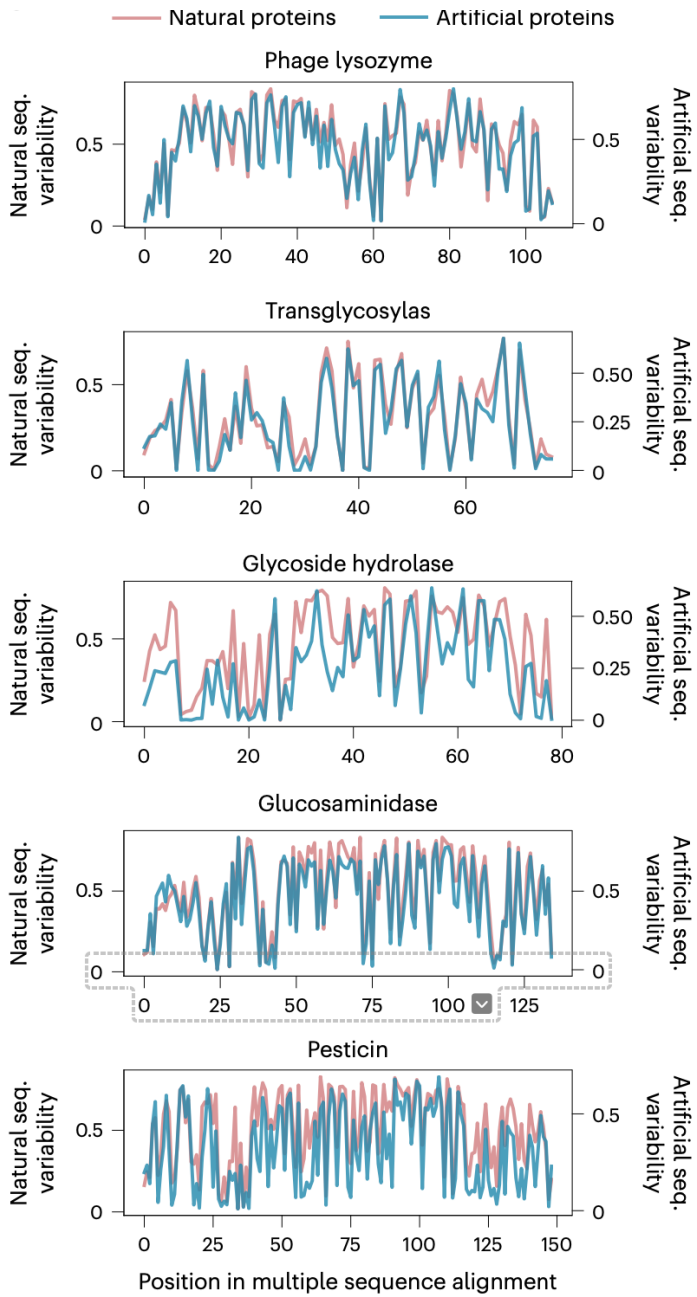

a
Control tag(s)
Model
Generated english sentence
Politics
Sports
Review 5★
Review 1★
English language model
Voting for the presidential election has begun
The Red Sox defeated the Yankees at Fenway
This knife is excellent for slicing meat
This knife is poorly made, not sharp at all!
b
Control tag(s)
Model
Generated protein sequence
Immunoglobulin
Chorismate mutase
Glucosaminidase
Phage lysozyme
Protein language model
DIQMTQSPASLS ... PKSFNRNEC
MSNTELELLRQK ... KEKAGLELQ
YIEKYNAIAERHK ... RHKLNRFDG
NIDFGFICELEGF ... ADLLESSMR
c
Universal protein sequence dataset
Training
Protein language model
Generation
Control tag for protein family f
Artificial proteins of protein family f
d
Universal protein sequence dataset
Natural proteins
280M sequences
>19K Pfam families
Lysozymes
56K sequences
5 Pfam families
e
Training: Negative log likelihood minimization
Next amino acid prediction
a1 a2 a3 a4 ...
Transformer decoder
c a1 a2 a3 ...
Control tag
Amino acids
Decoder block 36
Decoder block 1
Feed forward block
Multi-head self-attention
Feed forward block
Multi-head self-attention

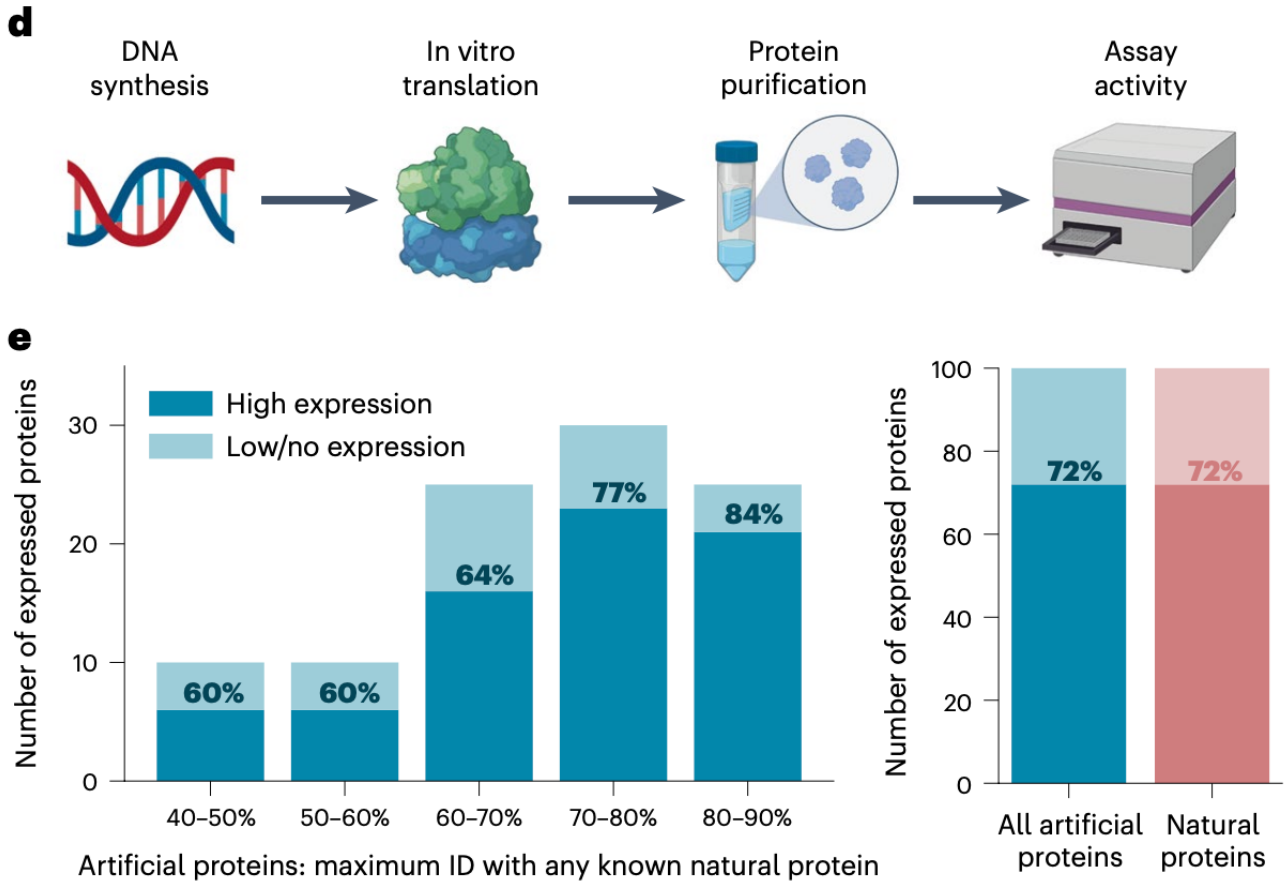
Generate artificial antibacterial proteins using ProGen

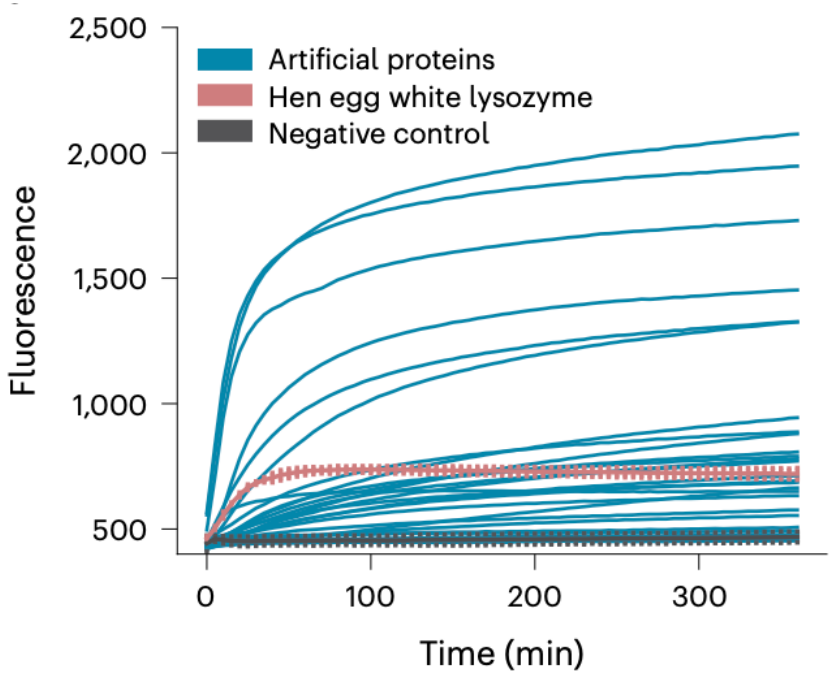


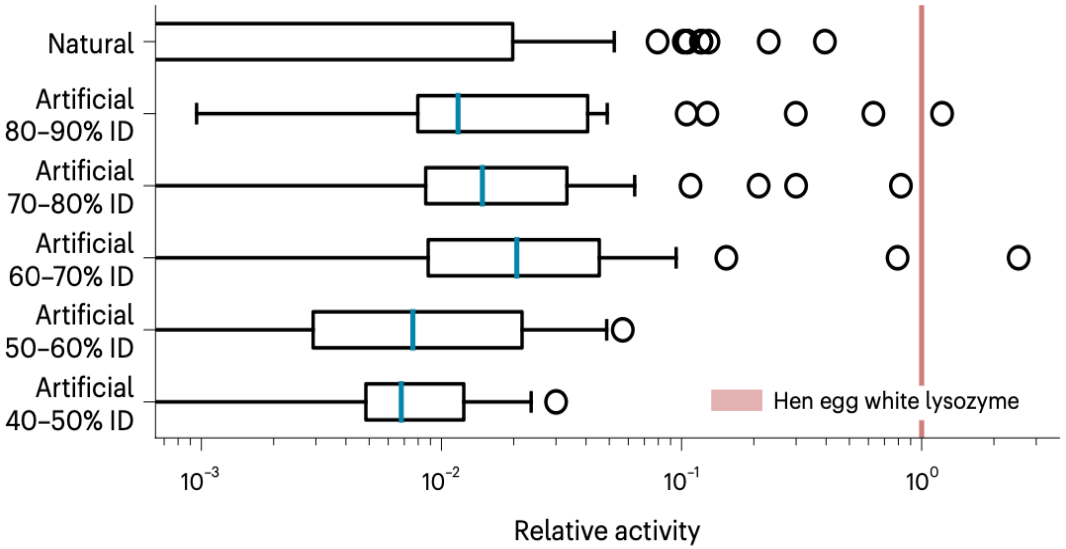


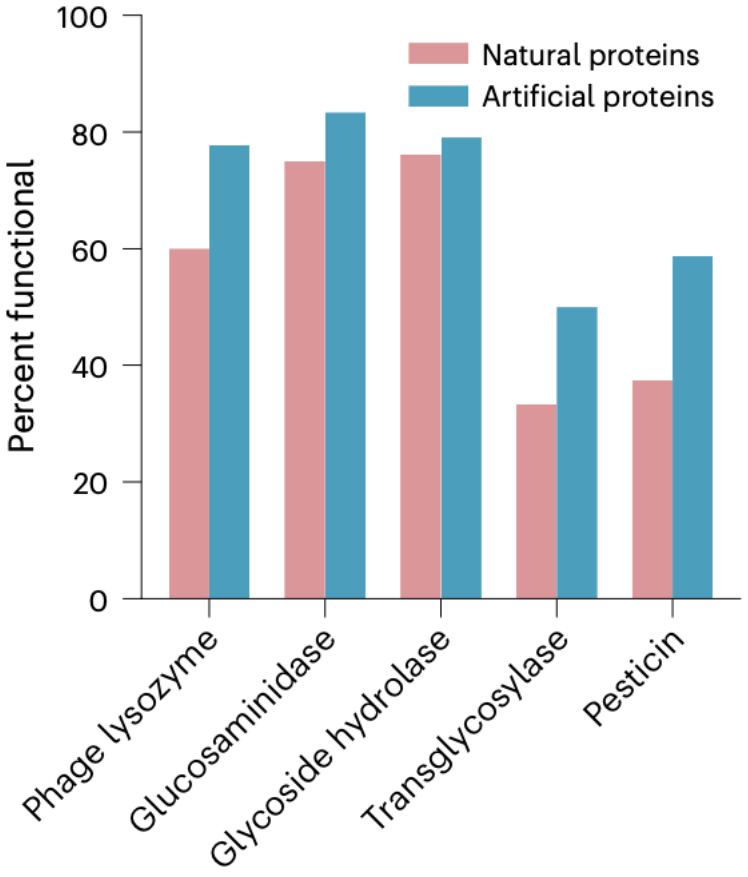



Natural proteins
Artificial proteins
Phage lysozyme
Transglycosylas
Glycoside hydrolase
Glucosaminidase
Pesticin
Natural seq. variability
Artificial seq. variability
Position in multiple sequence alignment

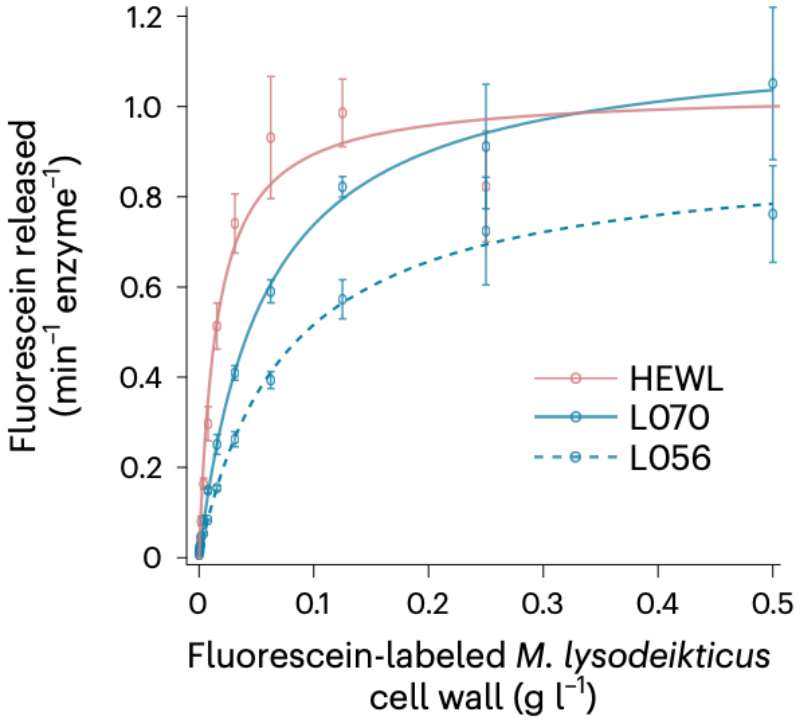

d
DNA synthesis
In vitro translation
Protein purification
Assay activity
e
High expression
Low/no expression
Number of expressed proteins
60%
60%
64%
77%
84%
40–50%
50–60%
60–70%
70–80%
80–90%
Artificial proteins: maximum ID with any known natural protein
72%
72%
All artificial proteins
Natural proteins

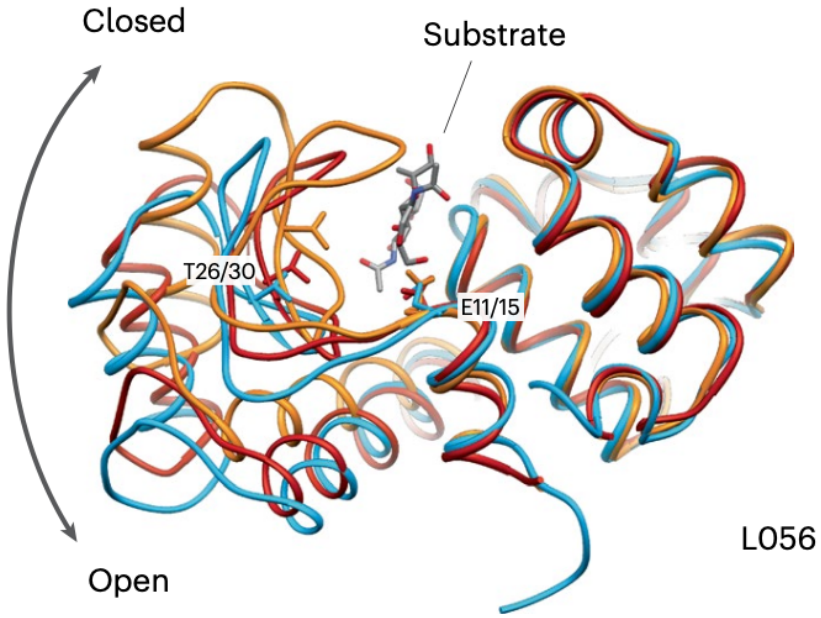

2,500
2,000
1,500
1,000
500
Fluorescence
Artificial proteins
Hen egg white lysozyme
Negative control
0
100
200
300
Time (min)


Natural
Artificial 80–90% ID
Artificial 70–80% ID
Artificial 60–70% ID
Artificial 50–60% ID
Artificial 40–50% ID
Hen egg white lysozyme
10^{-3}
10^{-2}
10^{-1}
10^{0}
Relative activity


100
80
60
40
20
0
Percent functional
Natural proteins
Artificial proteins
Phage lysozyme
Glucosaminidase
Glycoside hydrolase
Transglycosylase
Pesticin

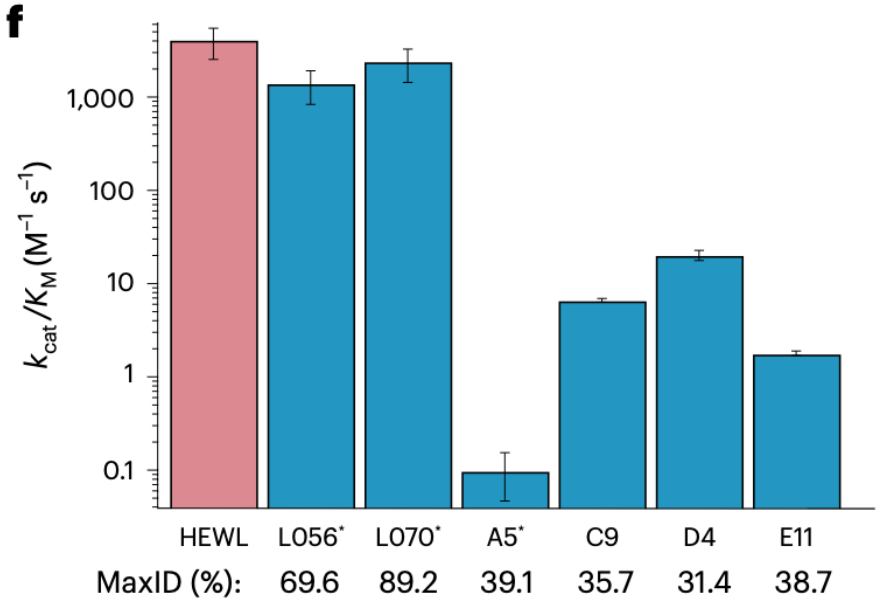
Two synthesized protein L070 and L056 (<40% max ID) gain competent enzyme activity to HEWL with 53 and 18 amino acid different from the nearest natural protein. L070 and L056 only share 17.9% sequence identity.



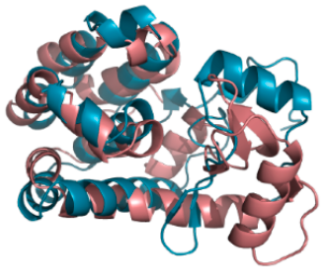


A global overlay of L056 crystal structure with two representative T4 lysozyme conformations is shown with L056 presented in sky blue, 'open' conformation of M6I T4 lysozyme (PDB accession 150L) in dark red, 'closed' conformation of wild-type T4 lysozyme (PDB accession 3FA0) in orange,

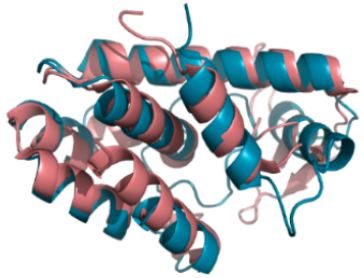
6 artificial sequences (<40% max ID) with gained lysosome activity



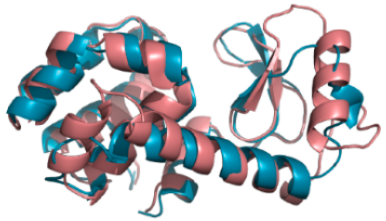
The 3D structure roughly math with nearest natural protein despite low 2D similarity



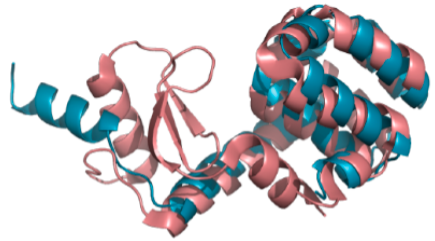
A5
maxID: 39.1%
ID-to-PDB_1L36: 25.9%



C9
maxID: 35.7%
ID-to-PDB_6ET6: 30.7%

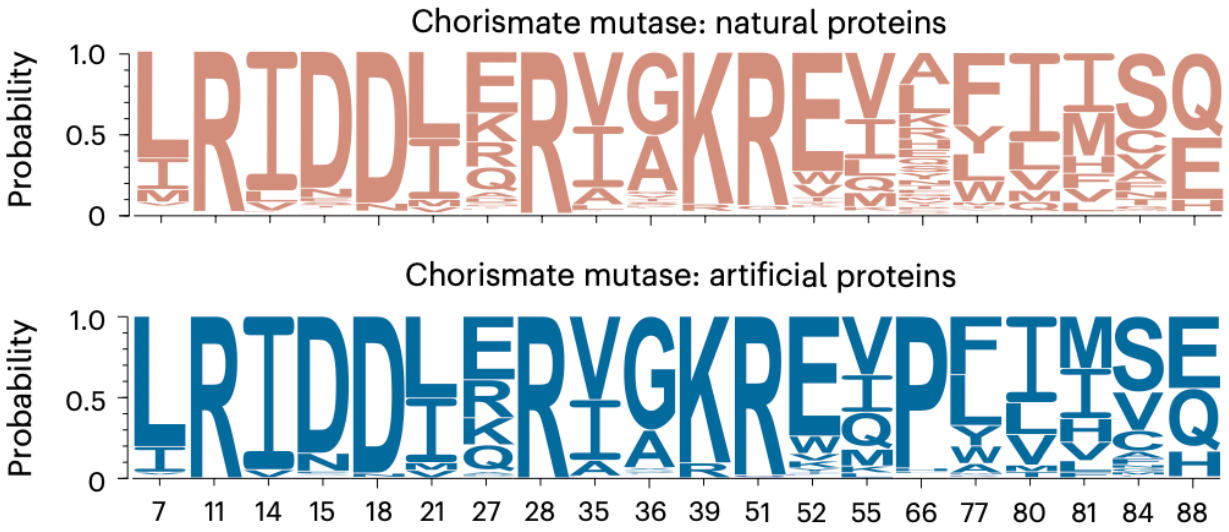


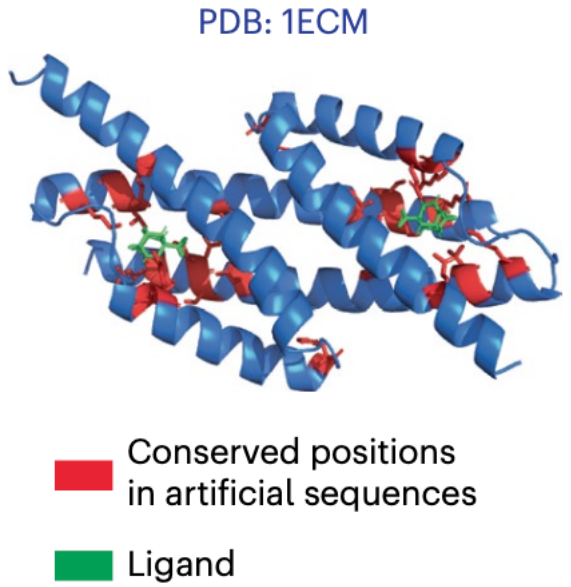
D4
maxID: 31.4%
ID-to-PDB_241L: 29.6%

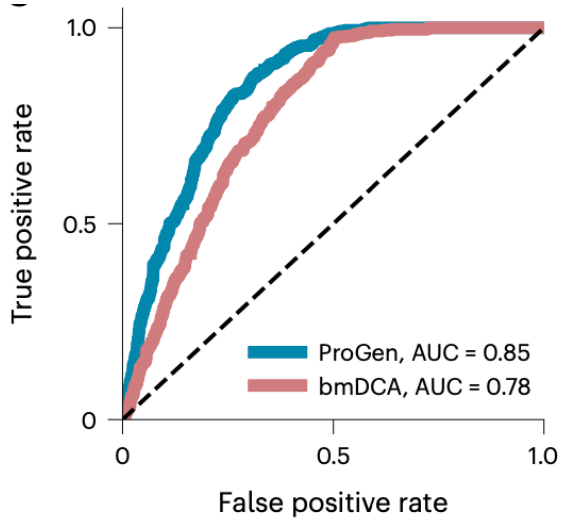


E11
maxID: 38.7%
ID-to-PDB_1QT5: 30.3%

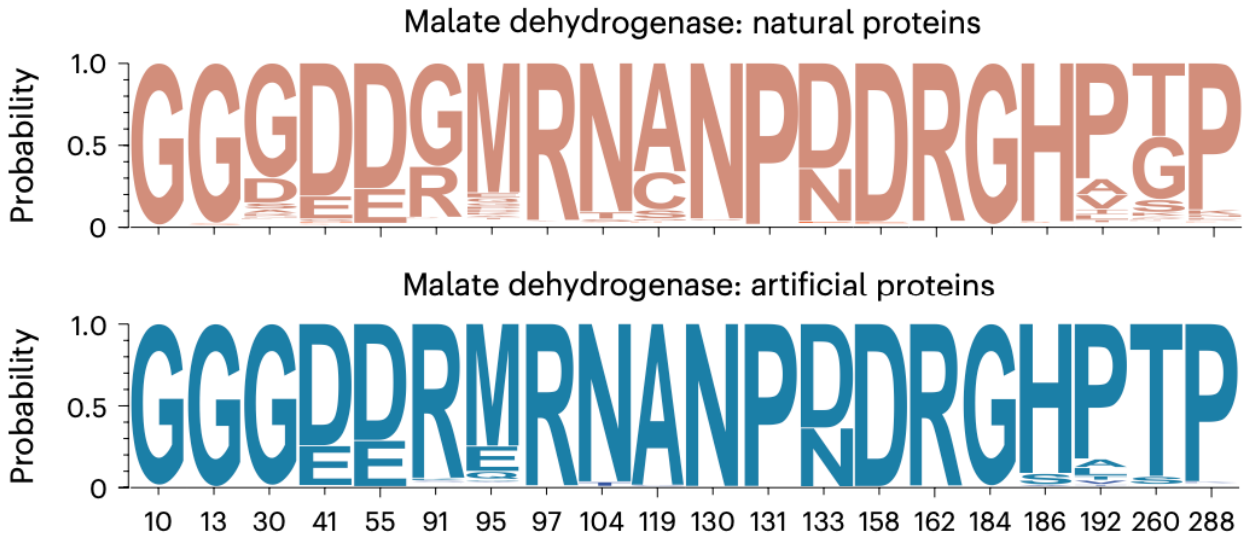
Generate protein sequences for chorismate mutase protein family using ProGen

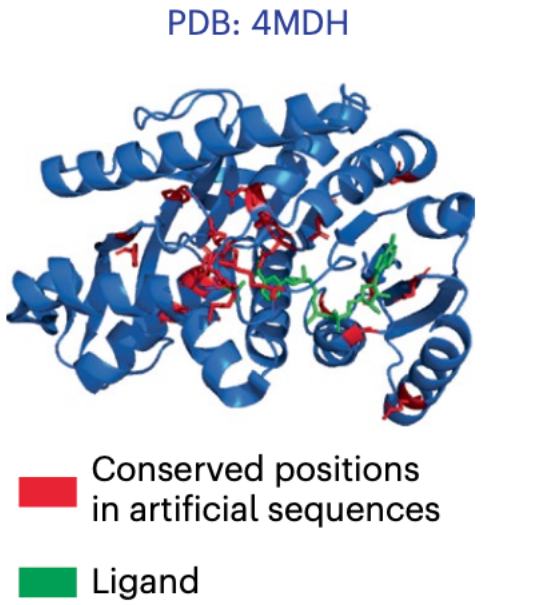


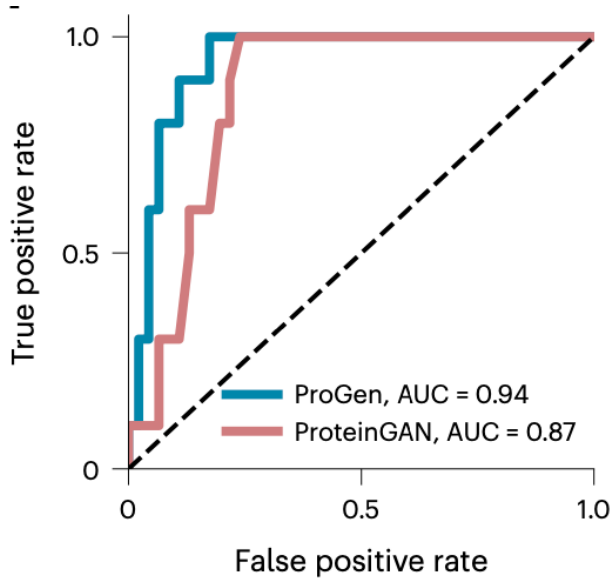




Generate protein sequences for malate dehydrogenase protein family using ProGen







nature methods

Resource

https://doi.org/10.1038/s41592-022-01685-y

AlphaFill: enriching AlphaFold models with ligands and cofactors

Received: 10 December 2021

Accepted: 18 October 2022

Published online: 24 November 2022

Check for updates

Maarten L. Hekkelman [1,2], Ida de Vries [1,2], Robbie P. Joosten [1,3] ✉ & Anastassis Perrakis [1,3] ✉

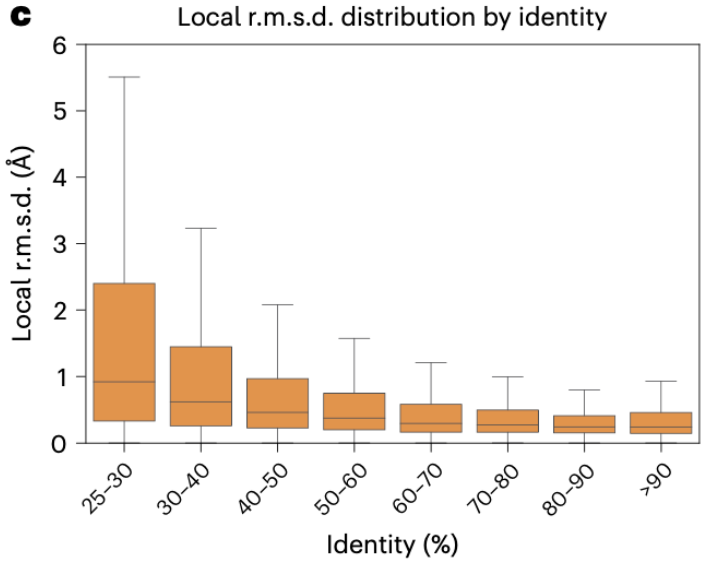
Artificial intelligence-based protein structure prediction approaches have had a transformative effect on biomolecular sciences. The predicted protein models in the AlphaFold protein structure database, however, all lack coordinates for small molecules, essential for molecular structure or function: hemoglobin lacks bound heme; zinc-finger motifs lack zinc ions essential for structural integrity and metalloproteases lack metal ions needed for catalysis. Ligands important for biological function are absent too; no ADP or ATP is bound to any of the ATPases or kinases. Here we present AlphaFill, an algorithm that uses sequence and structure similarity to 'transplant' such 'missing' small molecules and ions from experimentally determined structures to predicted protein models. The algorithm was successfully validated against experimental structures. A total of 12,029,789 transplants were performed on 995,411 AlphaFold models and are available together with associated validation metrics in the alphafill.eu databank, a resource to help scientists make new hypotheses and design targeted experiments.

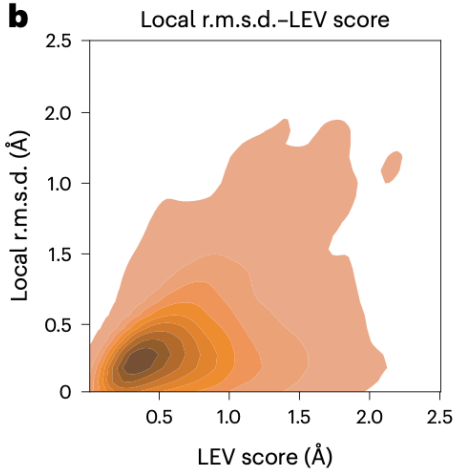
Process for ligand co-factor transplant

First, we search for sequence homologs for each structure in the Alpha- Fold database in the PDB-REDO databank. We consider structures with identity higher than 25% over an aligned sequence of at least 85 residues as hits.

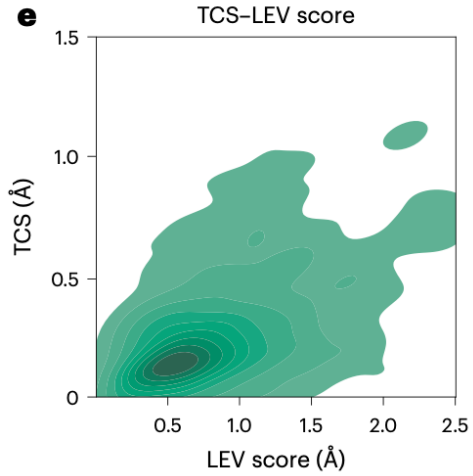
Next, the selection of structures with compounds of interest are structurally aligned on the Cα-atoms of the AlphaFold model, and the root-mean-square deviation (r.m.s.d.) is calculated (global r.m.s.d.).

Starting from the closest homolog, all backbone atoms within 6 Å from the atoms of each compound that will be considered for ‘transplantation’ are selected and used for a local structural alignment to the AlphaFold model; the r.m.s.d. from this alignment is also calculated (local r.m.s.d.). Compounds are then transplanted into the AlphaFold model to make the AlphaFill model.





Local environment validation (LEV) score as the all-atom r.m.s.d. of any ligand atom and all proteins’ atoms within 6.0 Å from the ligand, between the AlphaFill and experimental complexes.

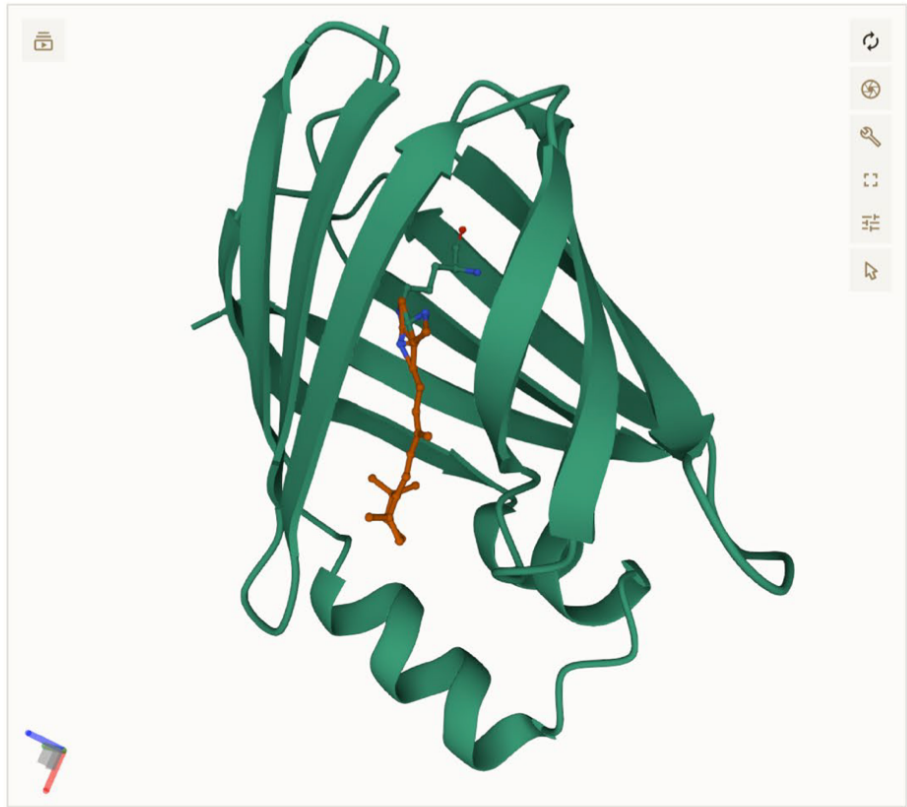


Transplant clash score (TCS) as a function Of the van der Waals overlaps between a transplanted ligand and its binding site

P29373

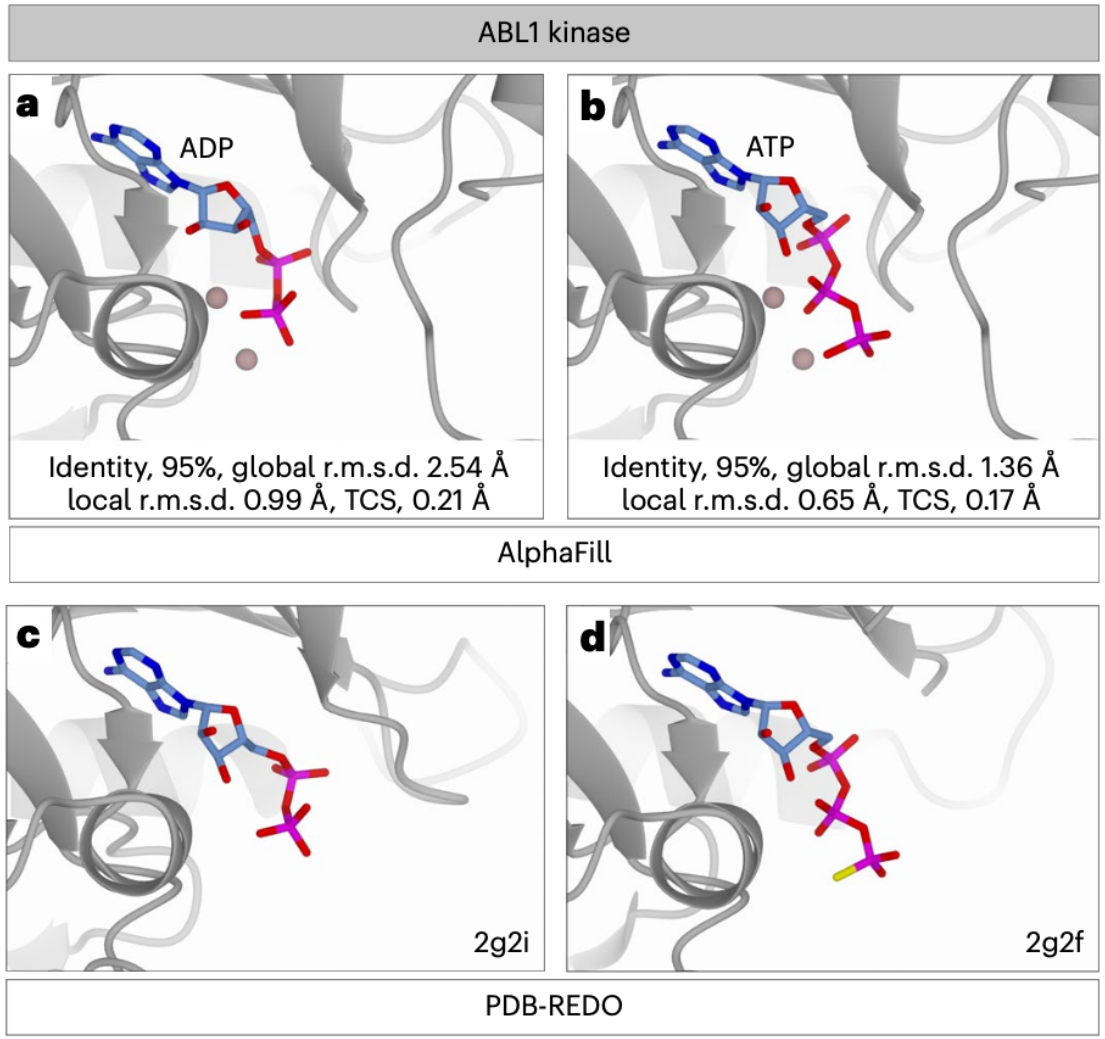
Cellular retinoic acid-binding protein 2

Structure file	https://alphafill.eu/v1/aff/P29373-F1
Metadata	https://alphafill.eu/v1/aff/P29373-F1/json
Original AlphaFold model	https://alphafold.ebi.ac.uk/entry/P29373



25% identity | 30% identity | 40% identity | 50% identity | 60% identity | **70% identity**

Compound	PDBID	Global RMSd	Asym	Local RMSd	TCS	Show
B3P	3fel.A	0.56	M	0.20	0.09	☐
			L	0.21	0.30	☐
NA	2g78.A	0.36	G	0.19	0.00	☐
	2fs6.A	0.37	C	0.17	0.32	☐
	2g79.A	0.46	I	0.04	0.00	☐
			J	0.08	0.00	☐
			H	0.09	0.00	☐
	2g7b.A	0.52	O	0.13	0.00	☐
			N	0.23	0.34	☐
	2frs.A	0.69	E	0.22	0.14	☐
			D	0.25	0.09	☐
			F	0.96 ?	0.32	☐
	3d97.B	0.94	P	0.04	0.00	☐
			Q	0.05	0.48	☐
REA	1cbs.A	0.36	B	0.52	0.21	☐
	1cbr.A	0.62	T	0.04	0.00	☐
RET	2g79.A	0.46	K	0.59	0.14	☐
	4i9s.A	0.76	R	0.78	0.99 ?	☑ optimise



For the human tyrosine-protein kinase ABL1 (AF-P00519) the AlphaFill model shows an ADP molecule and an ATP molecule (Fig. 5a,b) allowing different hypotheses for the functional state of this model. The global r.m.s.d. for the ADP source is 2.54 and for ATP 1.36 Å, while the local r.m.s.d. for ADP is 0.99 Å and for ATP 0.65 Å. This suggests that the structure is more representative of the ATP-bound state.

How do you think the roles of AI for scientific research in the future and what we can do to take advantage of it?

We need more creative thinking, which is beyond the existing knowledge and raise questions. Ask a question is more important than solving it.