Building a search engine for proteins

Roland Faure[1], ... , Rayan Chikhi[1]

[1]Institut Pasteur

November 2026

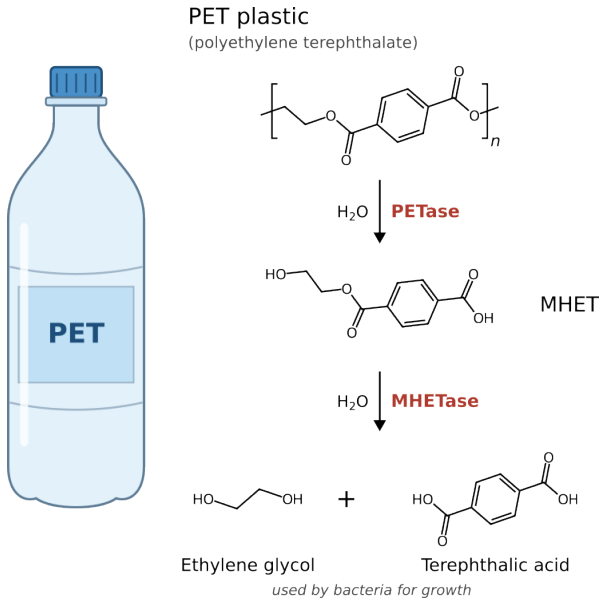
Slides available (CC-BY) at: `rolandfaure.github.io`

You have a favorite protein



You have a favorite protein

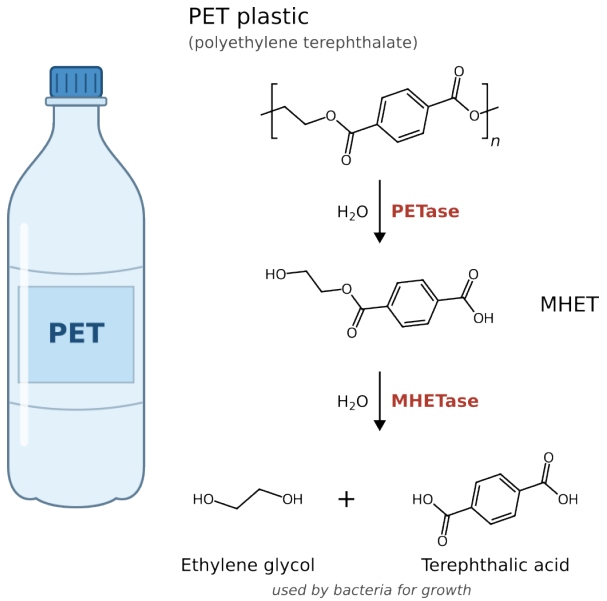




- Wow, a PETase!! I want to find other

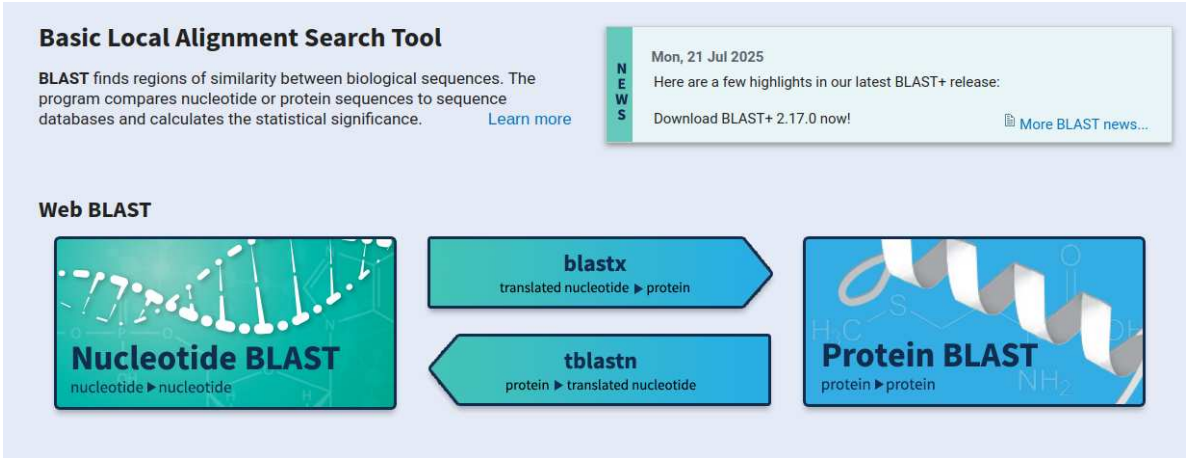
You have a favorite protein





- Wow, a PETase!! I want to find other
- Is it found somewhere else? Where?
- It only works at certain temperature, PH...
- I want to fold it

Let's BLAST the PETase



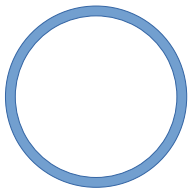
select all *50 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Vibrio gazogenes strain ATCC 43942 chromosome 1, complete sequence	Vibrio gazogenes	1618	1618	100%	0.0	100.00%	3471064	CP018835.1
Vibrio gazogenes strain PB1 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
Vibrio gazogenes strain DSM 21264 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516262	CP151640.1
Vibrio gazogenes ATCC 29988 DNA, chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
Vibrio spartinae CECT 9026 DNA, chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	4003627	AP024907.1
Vibrio spartinae strain 3.6 chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	3817659	CP046268.1

- Found 50 hits from a few different species

What is indexed in BLAST exactly?

In BLAST nr/nt



Well-assembled genomes

Non-redundant

0.1%

NOT in BLAST nr/nt

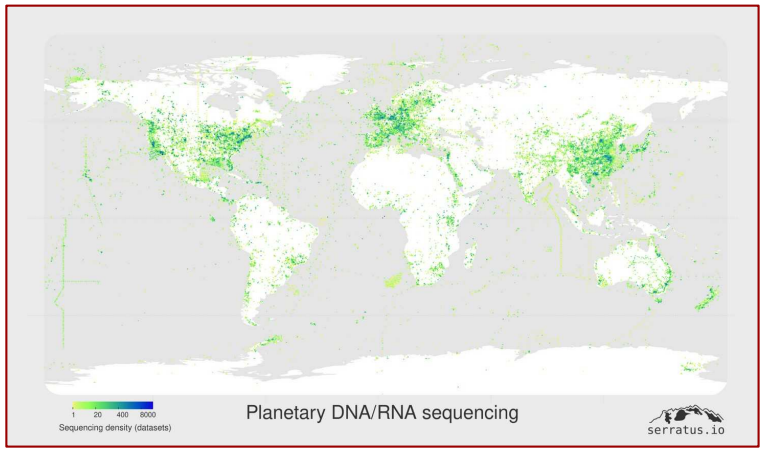


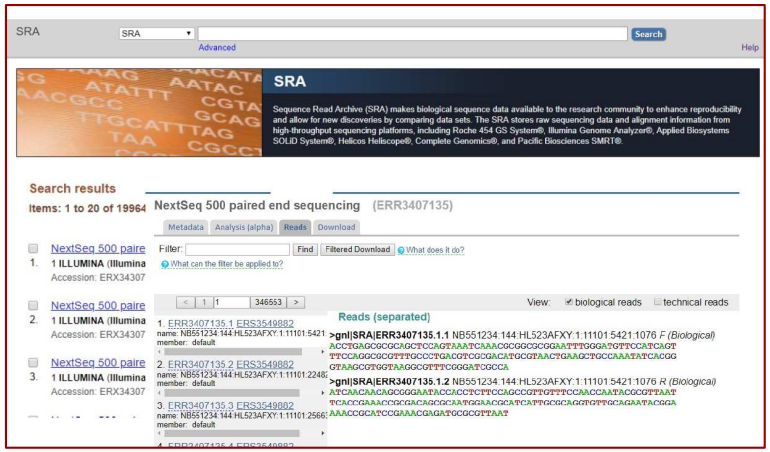
Random pieces of sequenced genomes

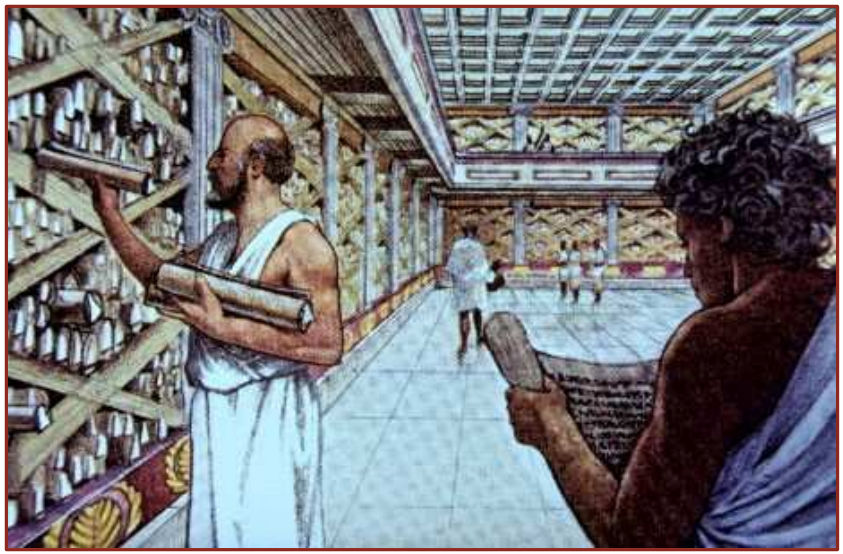
Every version of every sequence

99.9%

The biggest database: **the SRA**







"Library of Alexandria" for genetics

- **All** public sequencing data, 80 PB

At 1 Gbit/s, how much time to download the SRA?

A: 20 days

B: 20 weeks

C: 20 months

D: 20 years

At 1 Gbit/s, how much time to download the SRA?

A: 20 days

B: 20 weeks

C: 20 months

D: 20 years

Let's build a protein catalog of the SRA



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

Let's build a protein catalog of the SRA

SRA



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

Logan-contigs



27 millions accessions, **1 PB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

Let's build a protein catalog of the SRA

SRA



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

Logan-contigs



27 millions accessions, **1 PB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

Logan-proteins



151 billion proteins, **2 TB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

Identified protein-coding gene

Logan-proteins

- All Logan proteins are available online

Dataset	Description	Size
`logan_c1.0_human_proteins.fasta.zst`	All human-associated proteins	283 GiB
`logan_c1.0_nonhuman_proteins.fasta.zst`	All nonhuman-associated proteins	1.8 TiB
`logan50_c1.0_human_complete.fa.zst`	Representative clustered human-associated proteins (50% identity)	5.5 GiB
`logan50_c1.0_nonhuman_complete.fa.zst`	Representative clustered nonhuman-associated proteins (50% identity)	253 GiB

Logan-proteins

logan_proteins.fasta.zst

```
>SRR123456_18294_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR8599441_37164_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR982647_54743_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR10083945_66825_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR364389_308866_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR990230_83216_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
```

Logan-proteins

logan_proteins.fasta.zst

```
>SRR123456_18294_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR8599441_37164_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR982647_54743_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR10083945_66825_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR364389_308866_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR990230_83216_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
```

Proteins are clustered & stored by cluster

2 TB

Logan-proteins

logan_proteins.fasta.zst

```
>SRR123456_18294_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR8599441_37164_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR982647_54743_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR10083945_66825_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR364389_308866_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR990230_83216_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
```

zst frame 1

zst frame 2

2 TB

Logan-proteins

logan_proteins.fasta.zst

```
>SRR123456_18294_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR8599441_37164_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR982647_54743_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR10083945_66825_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR364389_308866_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR990230_83216_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
```

2 TB

cluster_to_frame.tsv

```
DRR000347_123254_1 →0
DRR000347_125518_1 →11456
```

100 GB

Logan-proteins

logan_proteins.fasta.zst

```
>SRR123456_18294_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR8599441_37164_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR982647_54743_3 CL:DRR000347_123254_1
MTHHLGPKCRLDMGYLCPYSFIEQIVF*
>SRR10083945_66825_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR364389_308866_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
>SRR990230_83216_1 CL:DRR000347_125518_1
MIIDTIDTRNVTKKSRGKTGQRFENMEVVITV*
```

2 TB

cluster_to_frame.tsv

```
DRR000347_123254_1 →0
DRR000347_125518_1 →11456
```

100 GB

prot_to_cluster.tsv

SRR10083945_66825_1	DRR000347_125518_1
SRR123456_18294_3	DRR000347_123254_1
SRR364389_308866_1	DRR000347_125518_1
SRR8599441_37164_3	DRR000347_123254_1
SRR982647_54743_3	DRR000347_123254_1
SRR990230_83216_1	DRR000347_125518_1

1.2 TB

In total: 3.3 TB for 151 billion proteins (22 byte/protein)

Let's build a fast protein search engine in Logan-proteins



- Query: one protein
- Result: all homologs in Logan-proteins
- Very fast

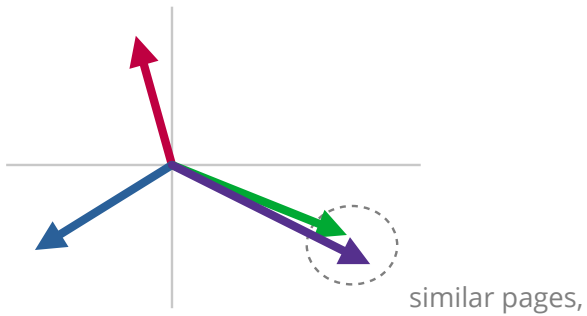
How do web search engine work?

The Web

billions of pages

Bacteria that eat plastic bottles
An enzyme that breaks down PET
The perfect croissant recipe
Champions League: tonight's results
⋮

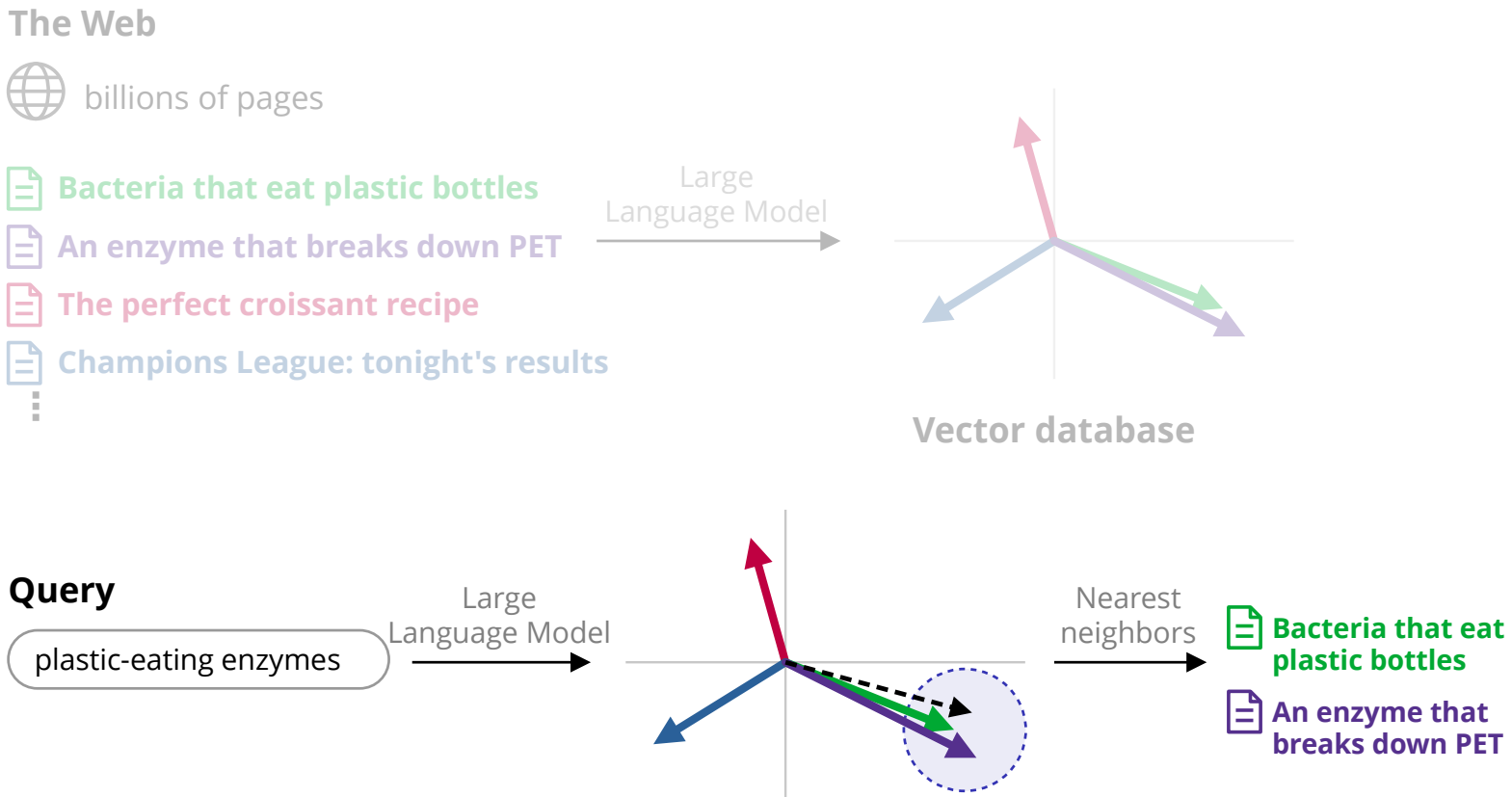
Large
Language Model



similar pages,
close vectors

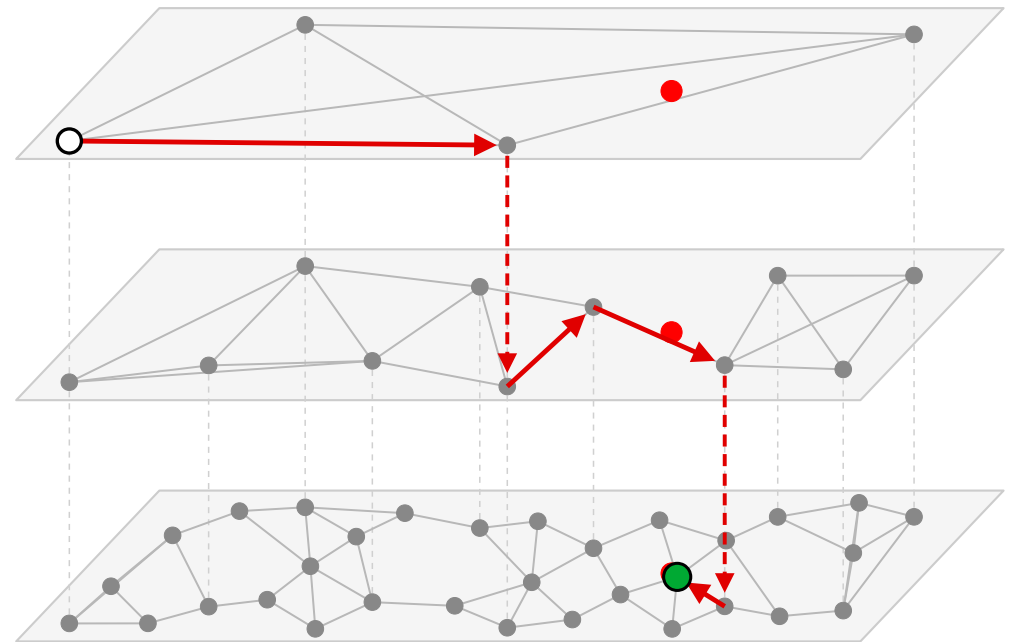
Vector database
(512 dimensions)

How do web search engine work?



How do web search engine work?

Hierarchical Navigable Small World (HNSW) index



index level 2
few vectors

index level 1
more vectors

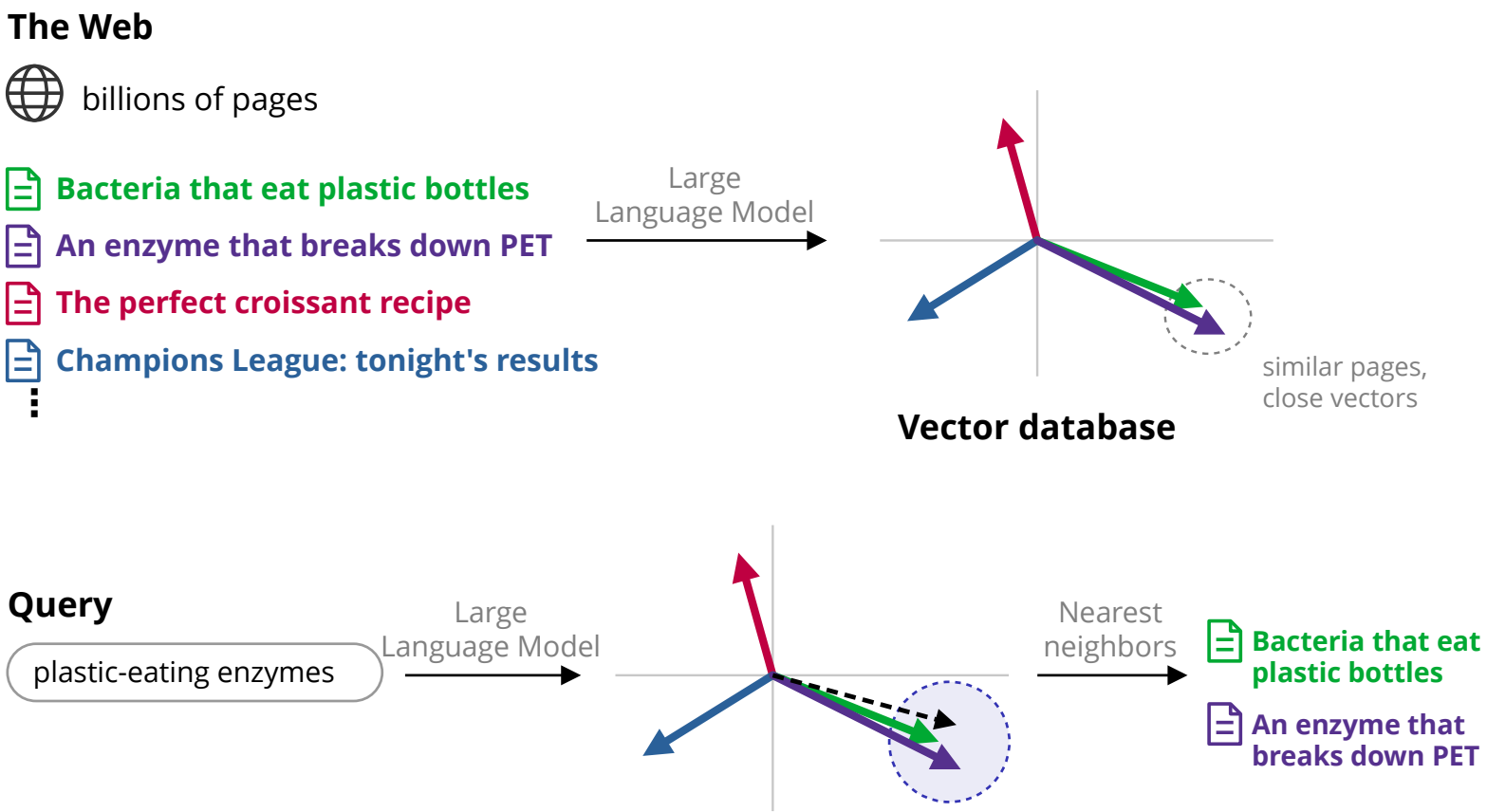
index level 0
all vectors

○ start ● query ● nearest page

every vector comparison: **10^9 distances**

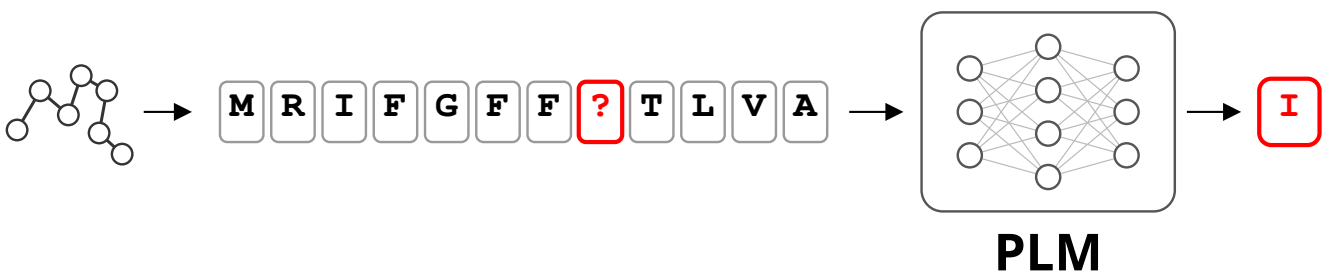
with HNSW: **~10^3 distances**

How to translate this to SRA search?

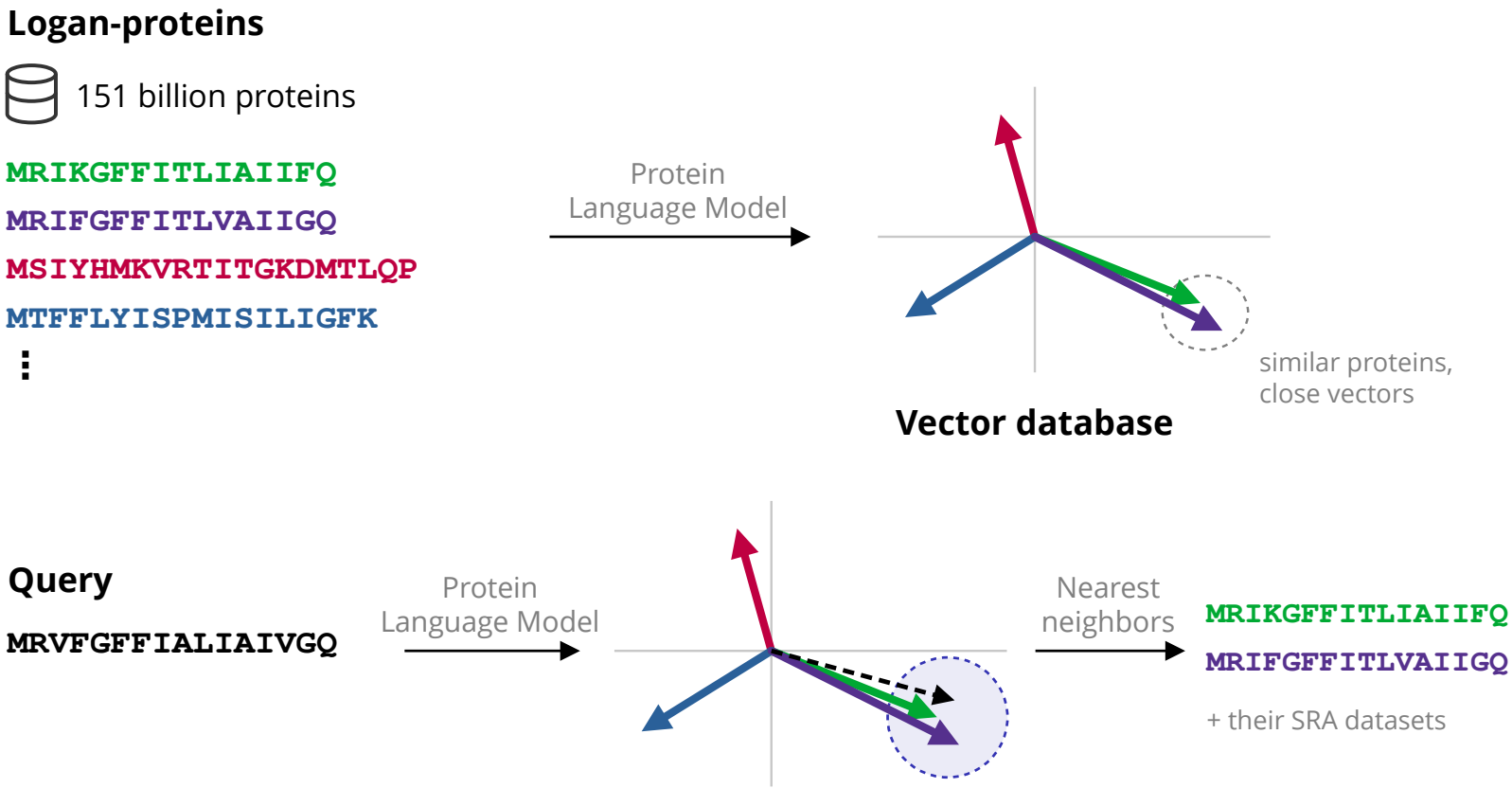


Protein Language Models (PLMs)





Let's build a protein search engine



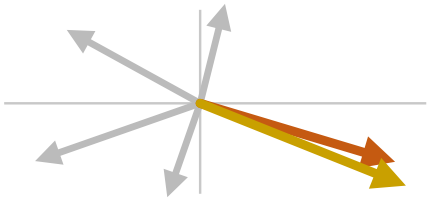
PLMs vs sequence similarity

PETase

SRR25547275_40149_2

```
MMNVLTKCKLALGIIAIFFSLPSFAVPCSDCSNGFERGQVP
|          ||    |
MSRQSRSSVVALACAASIVCVCAVGILSHSPGVARASSV
```

sequence similarity



embedding similarity

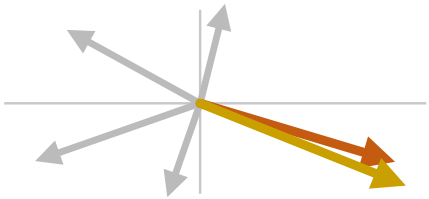
PLMs vs sequence similarity

PETase

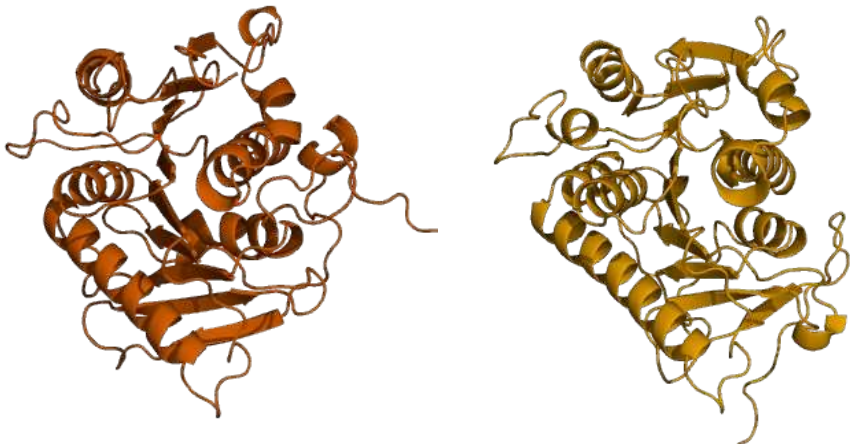
SRR25547275_40149_2

```
MMNVLTKCKLALGIIAIFFSLPSFAVPCSDCSNGFERGQVP
|          ||    |
MSRQSRSSVVALACAASIVCVCAVGILSHSPGVARASSV
```

sequence similarity

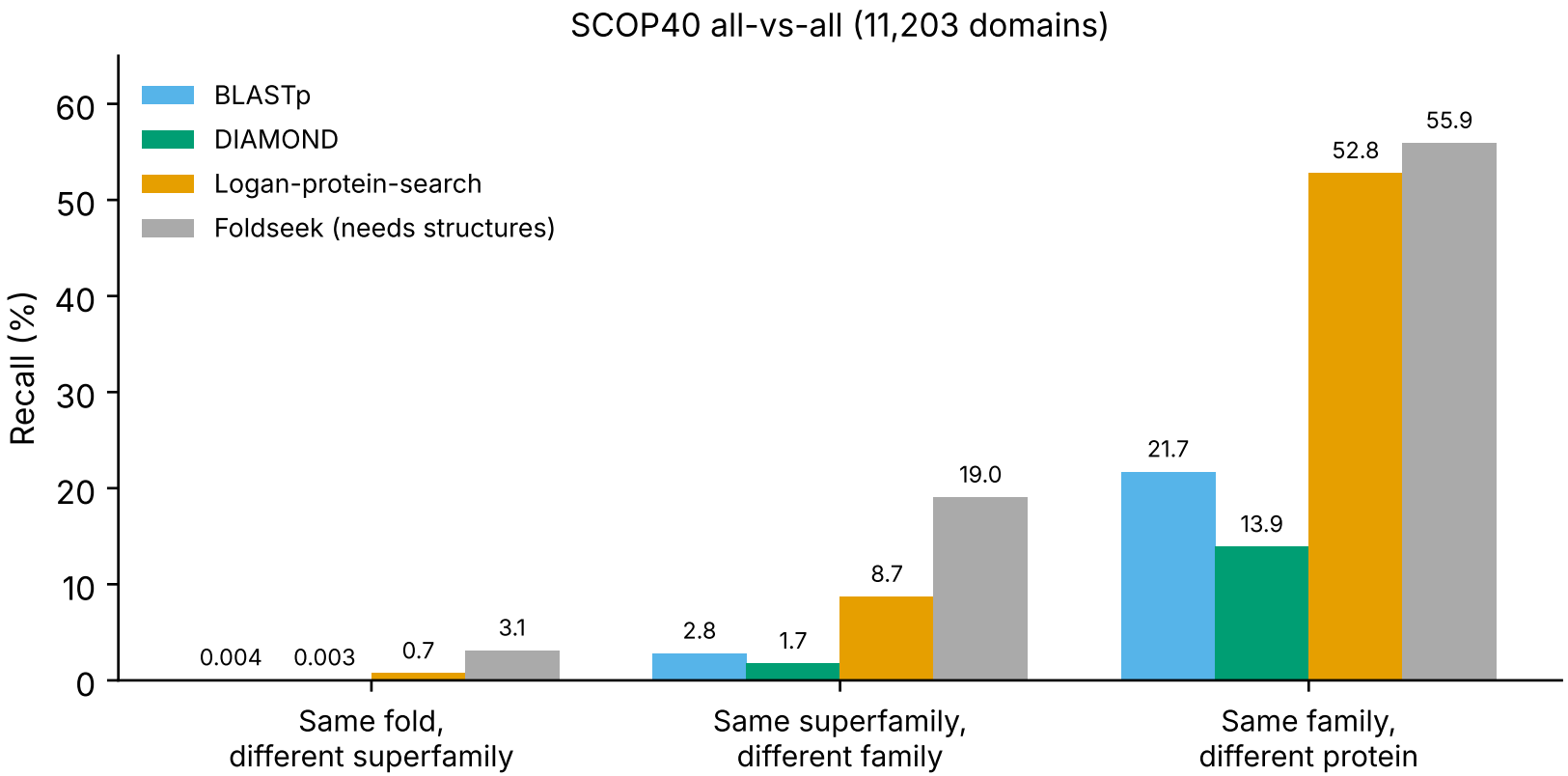


embedding similarity

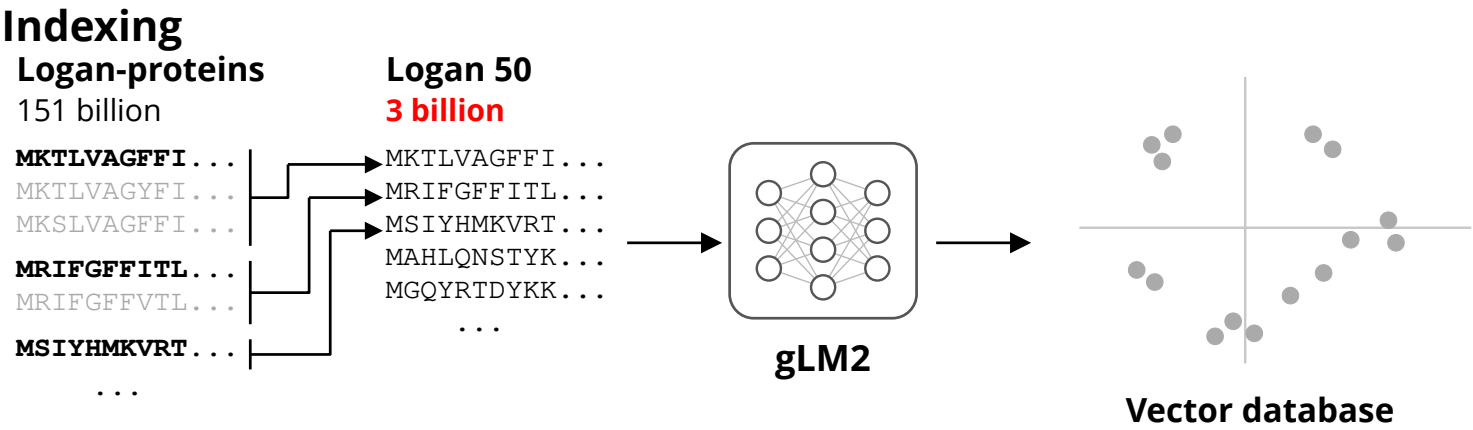


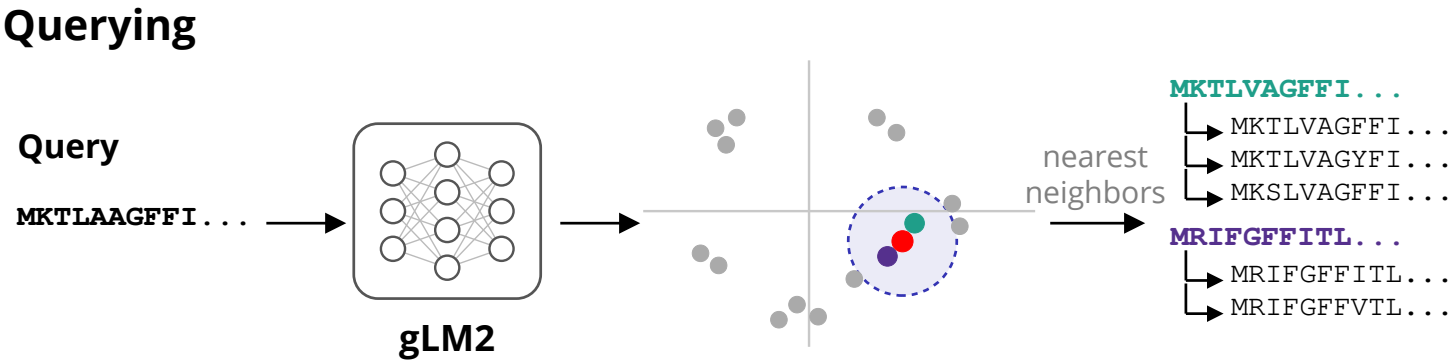
similar fold

PLMs vs sequence similarity



Logan-protein-search: the final search engine





Let's look for the PETase

- One hour computation (not google speed but not too bad)

Let's look for the PETase

- One hour computation (not google speed but not too bad)
- 300k results!

```
Query_10001  150  YHQIDEKEIGIVGYSQGGAGAYNTLEGKDGDK---FKTM-VTVSG---    ------VTESIGEKLHLPVWIYDPSKVTI  213
Query_10002  98   QPEVDANRVALIGWALGGGVVVAEAADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD-DMVRDRPERARSGRSRTV  176
Query_10003  95   EK-TGNPRVGVVGYCAGGGLALMLAAQRP-----------DAVKAVA   PYY---------GLIPWPGAQPDWAAIDA  149
Query_10004  130  SAFVDPERIGVIGFSHGGWTLLDFLGGPPAIHasATATDARDGLRSVV   AVYpycGADVQAGLGKWP--------ADV  198
Query_10005  151  YQKVDTEHIGISGHSQGGVGVFNAISEQPHSN---LYTCAVSLSP---    ------TQQDLAEALKIP---YDPTKTQI  212
Query_10006  99   DP-RCTGKVGIVGFCMGGGFTLLLAPRG-------------IFDAAA   PNY---------GVLP---RDLSALSSSC  148
Query_10007  99   LEFVDPDRIGVLGVCAGGGYSVNAAMTEHRIKavGTVVGANIG-----[4]ENNpiqTLEAIGKQRTAEANGAEPMIINW  174
Query_10008  136  GKFILSNKIAVIGHSMGGYTALALAGGIPWTQeaERVETSSDARVKAI    ----------VLMAPGAGWFMNSLSKVTI  202
```

Let's look for the PETase

- One hour computation (not google speed but not too bad)
- 300k results!

```
Query_10001  150  YHQIDEKEIGIVGYSQGGAGAYNTLEGKDGDK---FKTM-VTVSG---    ------VTESIGEKLHLPVWIYDPSKVTI  213
Query_10002  98   QPEVDANRVALIGWALGGGVVVAEAADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD-DMVRDRPERARSGRSRTV  176
Query_10003  95   EK-TGNPRVGVVGYCAGGGLALMLAAQRP-----------DAVKAVA   PYY---------GLIPWPGAQPDWAAIDA  149
Query_10004  130  SAFVDPERIGVIGFSHGGWTLLDFLGGPPAIHasATATDARDGLRSVV   AVYpycGADVQAGLGKWP--------ADV  198
Query_10005  151  YQKVDTEHIGISGHSQGGVGVFNAISEQPHSN---LYTCAVSLSP---    ------TQQDLAEALKIP---YDPTKTQI  212
Query_10006  99   DP-RCTGKVGIVGFCMGGGFTLLLAPRG-------------IFDAAA   PNY---------GVLP---RDLSALSSSC  148
Query_10007  99   LEFVDPDRIGVLGVCAGGGYSVNAAMTEHRIKavGTVVGANIG-----[4]ENNpiqTLEAIGKQRTAEANGAEPMIINW  174
Query_10008  136  GKFILSNKIAVIGHSMGGYTALALAGGIPWTQeaERVETSSDARVKAI    ----------VLMAPGAGWFMNSLSKVTI  202
```

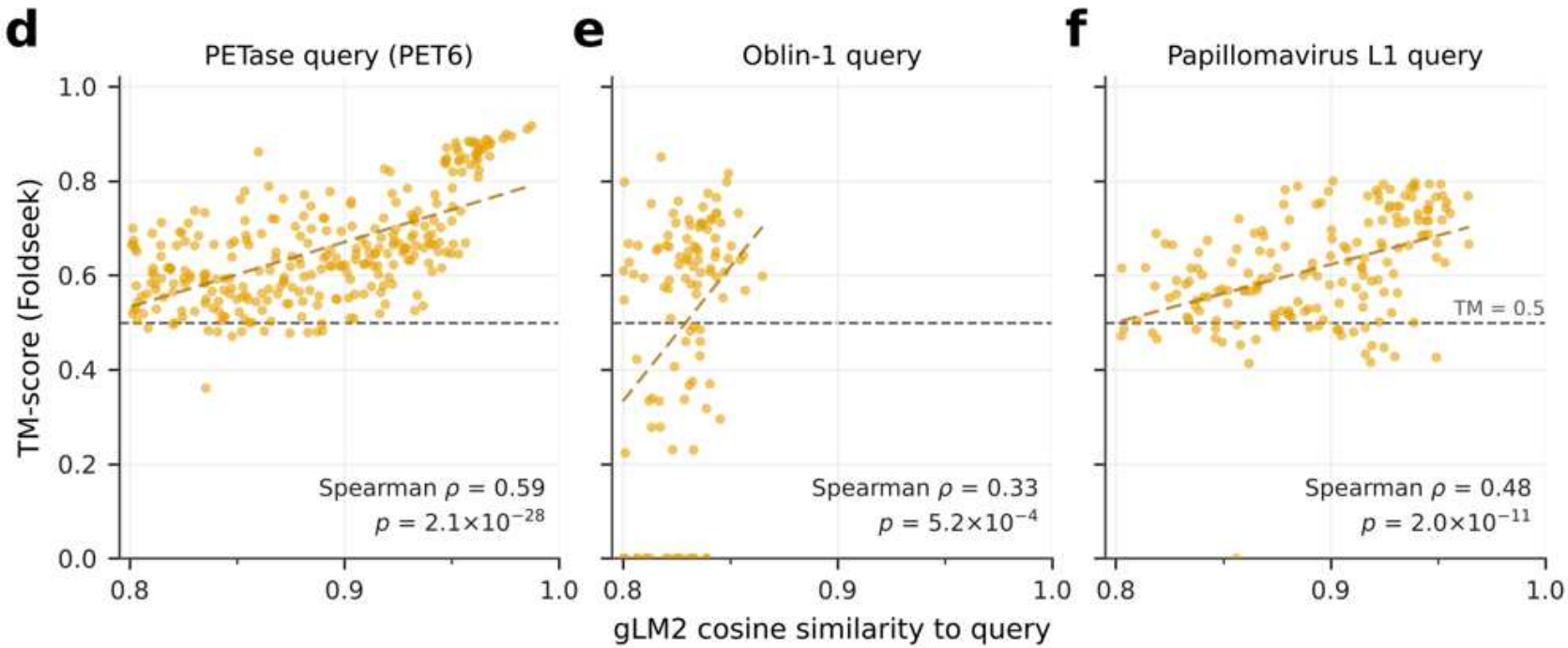
PETase found e.g. in "Fermented Xuecai"



Source: chillcrispbyxueci.substack.com

Limits of this index

- Limited by the Protein Language Model



Take-home messages

- Logan-proteins is the biggest protein database ever, and it is available

Take-home messages

- Logan-proteins is the biggest protein database ever, and it is available
- Searchable with Logan-protein-search, available on Galaxy

Take-home messages

- Logan-proteins is the biggest protein database ever, and it is available
- Searchable with Logan-protein-search, available on Galaxy
- PLM + vector search is fast and sensitive

Acknowledgments

- Rayan Chikhi
- Galaxy team: Anton Nekrutenko, Björn Grüning, Danielle Callan, Saim Momin, Patrik Smeds, Nate Coraor
- TACC team: Kelsey Beavers



Rayan Chikhi

Galaxy PROJECT

TACC
TEXAS ADVANCED COMPUTING CENTER

aws