# Building a search engine for proteins

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

[1]Institut Pasteur

November 2026

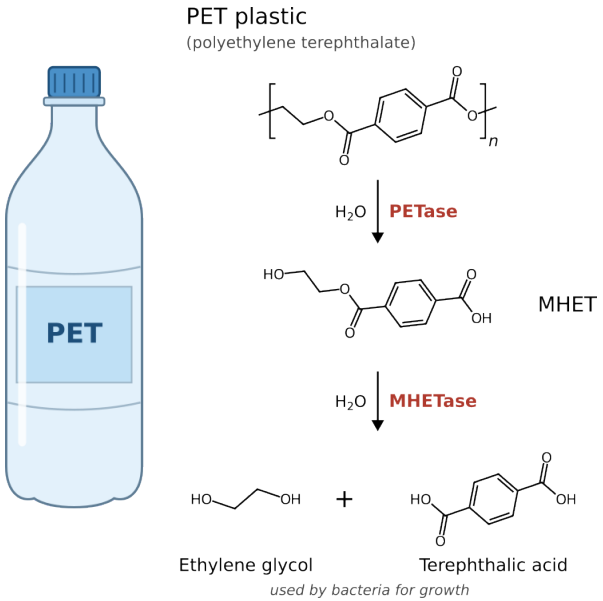
Slides available (CC-BY) at: `rolandfaure.github.io`

# You have a favorite protein



# You have a favorite protein

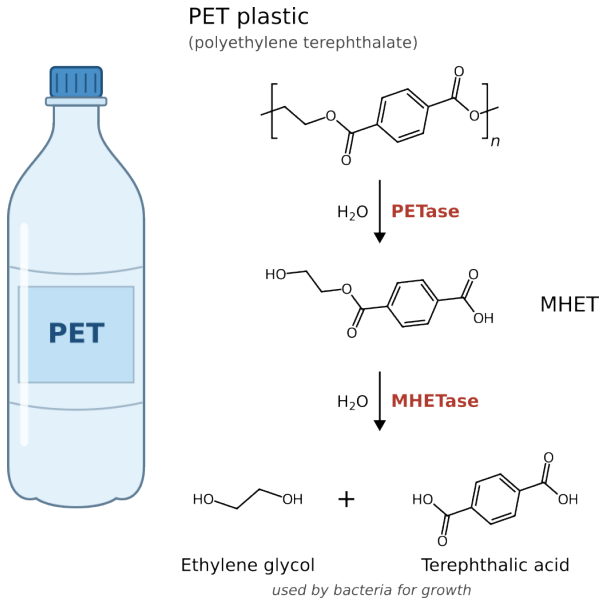



- Wow, a PETase!!

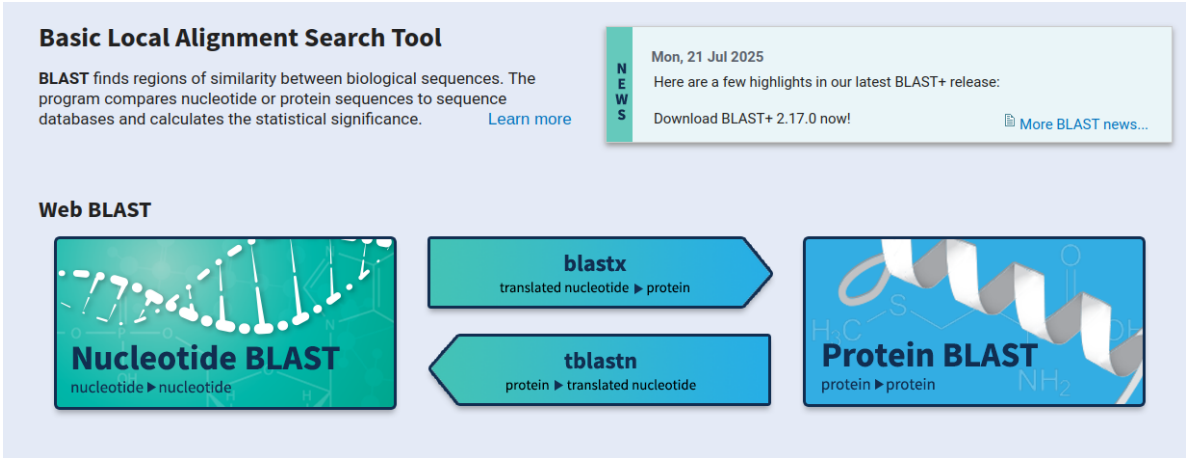
# You have a favorite protein




- Wow, a PETase!!
- Has this already been described?
- Is it found somewhere else? Where?
- I want to find other PETases!

# Let's BLAST the PETase

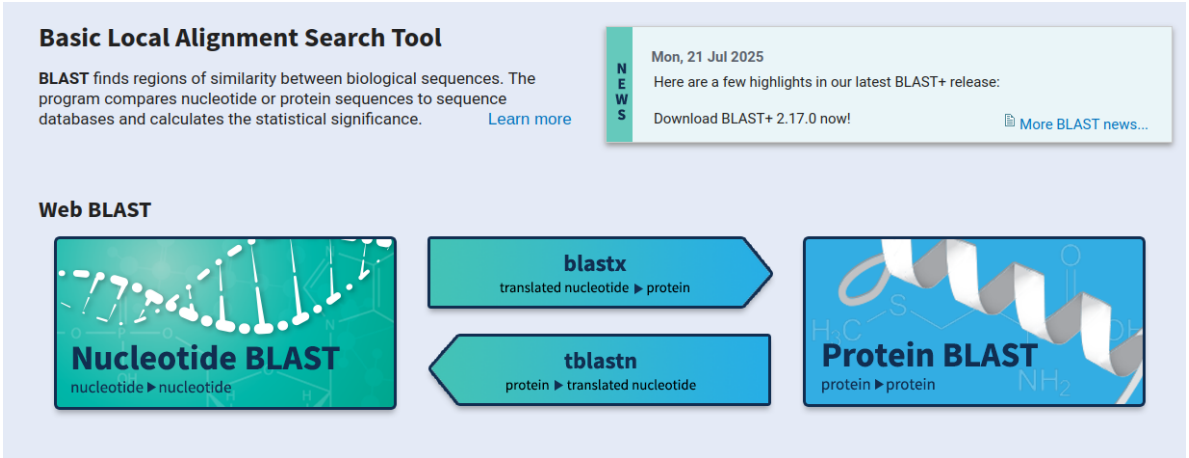





select all 50 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

| 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 |

# Let's BLAST the PETase





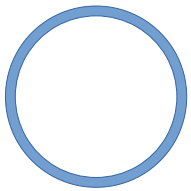

select all 50 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

| 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

**NOT in BLAST nr/nt**

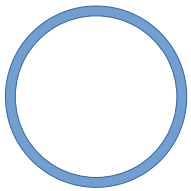


Random pieces of sequenced genomes

Every version of every sequence

# What is indexed in BLAST exactly?

**In BLAST nr/nt**



Well-assembled genomes

Non-redundant

**0.1%**

**NOT in BLAST nr/nt**

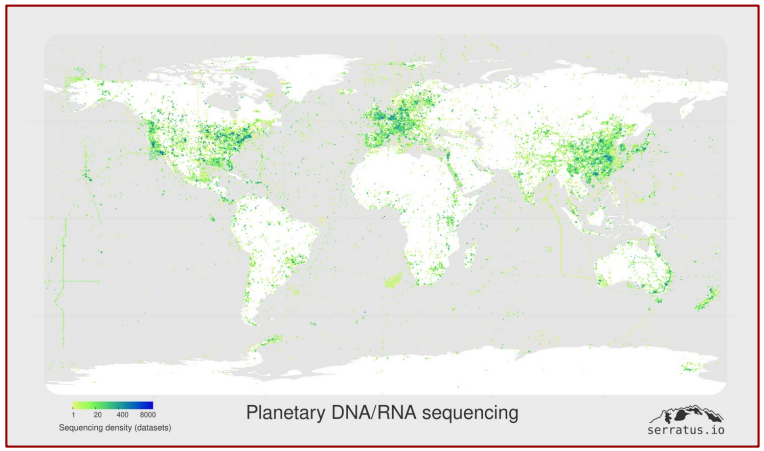


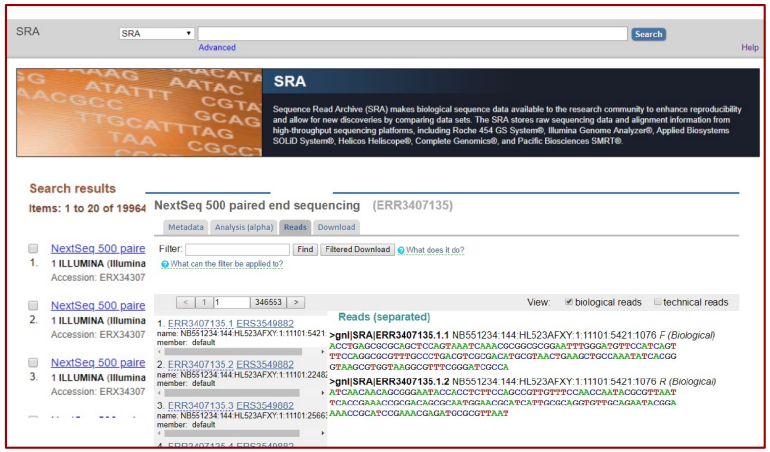
Random pieces of sequenced genomes

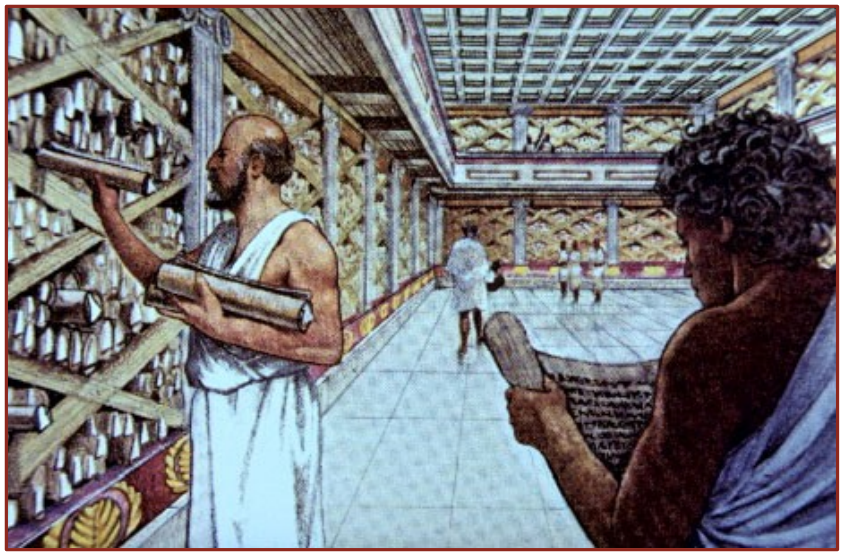
Every version of every sequence

**99.9%**

# The biggest database: **the SRA**



"Library of Alexandria" for genetics

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

# Let's align the PETase with the data of the SRA

is MMNVLTKCKLA in  ?

Sequencing dataset



Repeat for every dataset

# Let's align the PETase with the data of the SRA

is MMNVLTKCKLA in  ? 

Sequencing dataset

Repeat for every dataset



# 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?

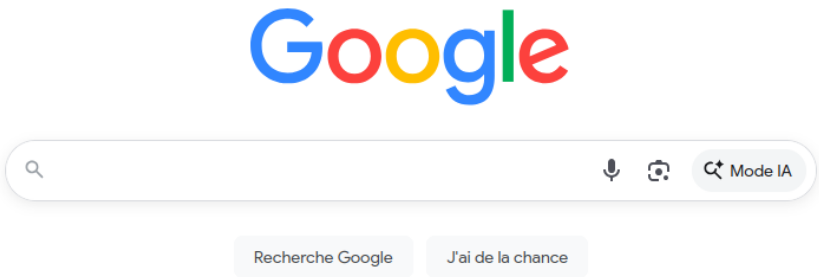
A: 20 days

B: 20 weeks

C: 20 months

D: 20 years

# Goal: let's build a fast protein search engine




- Return every protein in the SRA homolog to the query
- Very fast

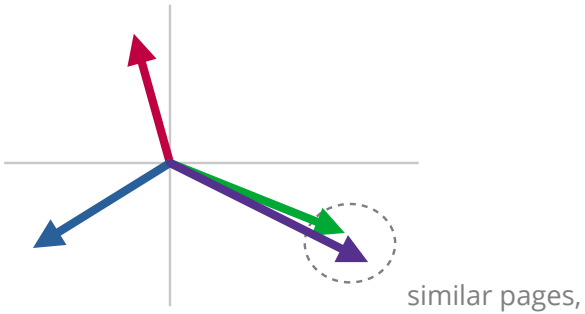
# How do web search engine work?

**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**

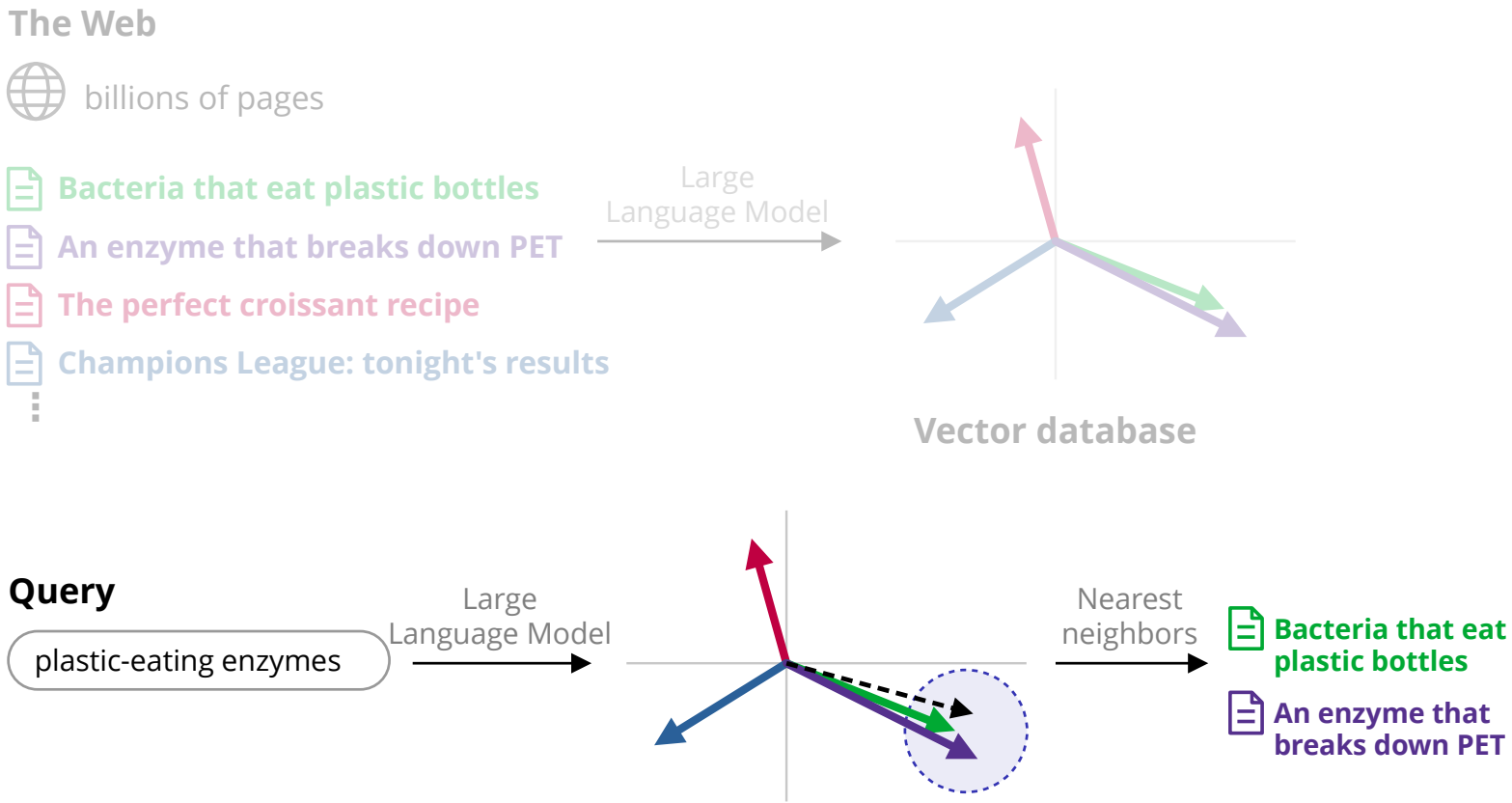
Large Language Model



similar pages, close vectors

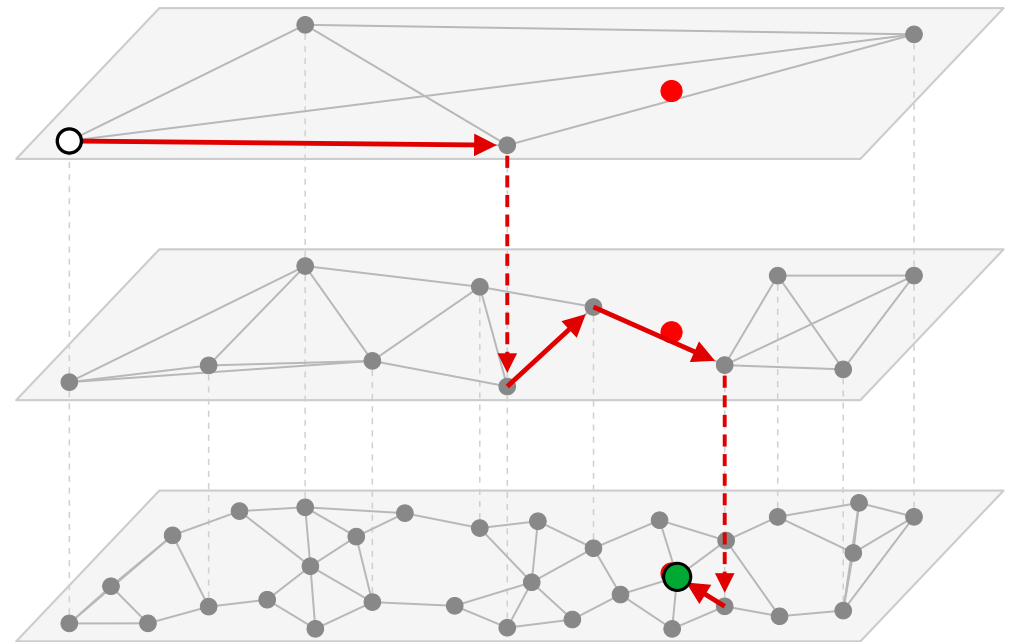
**Vector database**
(512 dimensions)

# How do web search engine work?



# How do web search engine work? Fast vector search

**Hierarchical Navigable Small World (HNSW) index**



**index level 2**
few vectors

**index level 1**
more vectors

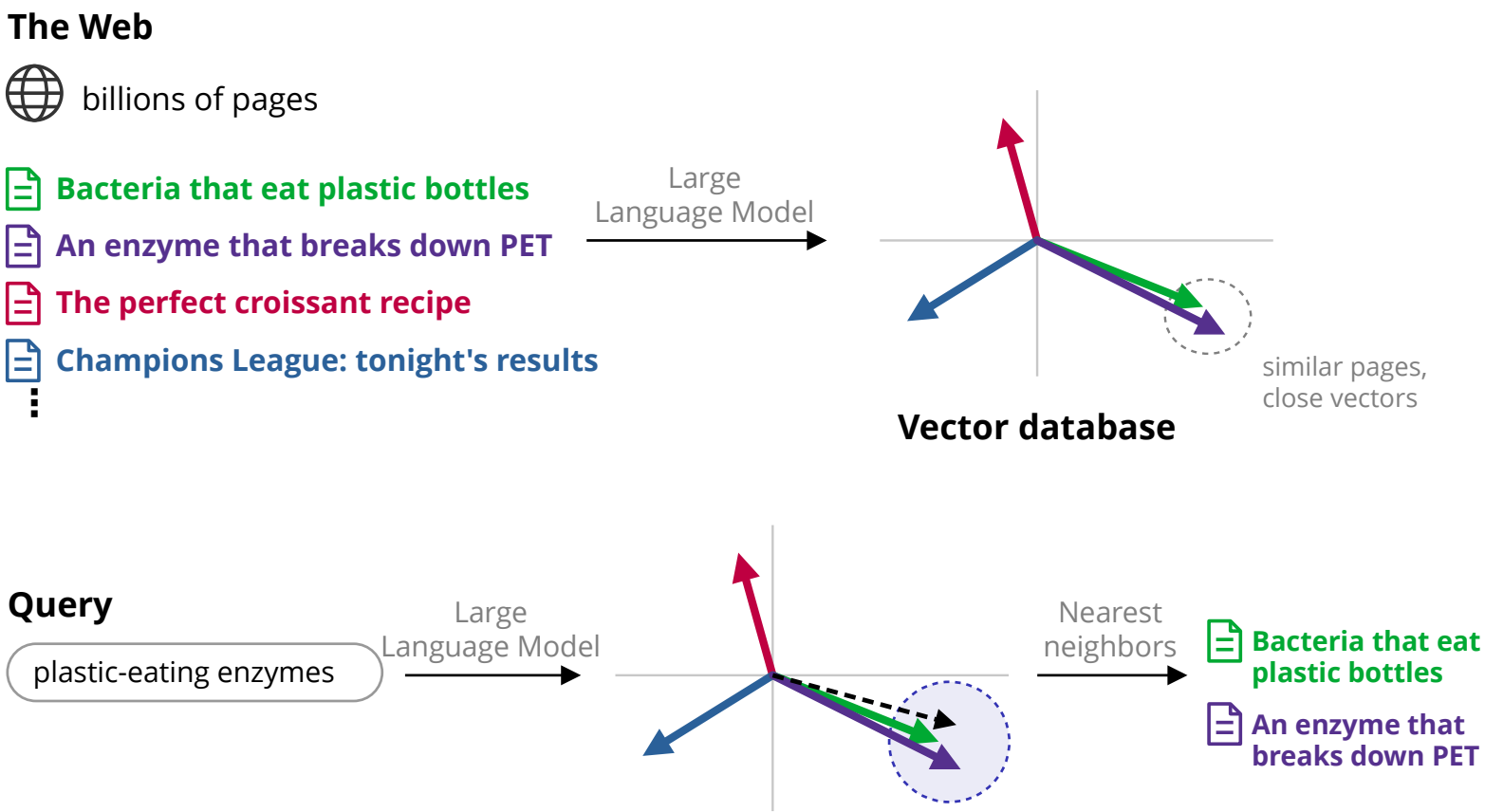
**index level 0**
all vectors

○ start ● query ● nearest page

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

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

# How to translate this to SRA search?



# First step: reducing the size of the data




27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

# First step: reducing the size of the data

**SRA**



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

**Logan-contigs**



27 millions accessions, **1 PB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

# First step: reducing the size of the data

**SRA**



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

**Logan-contigs**



27 millions accessions, **1 PB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

**Logan-proteins**



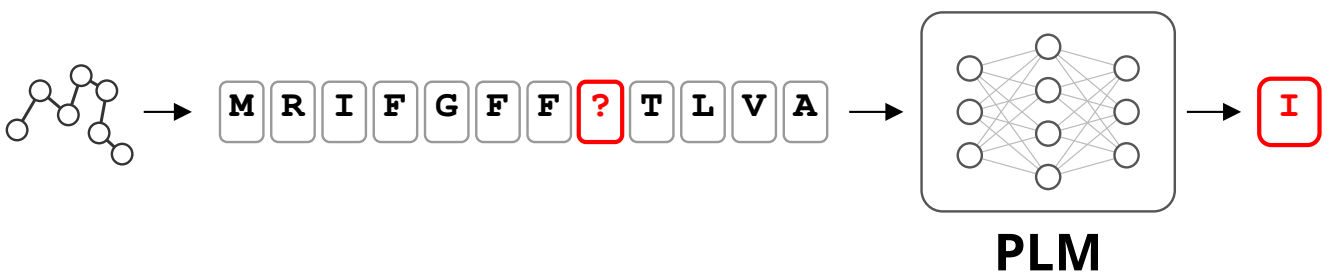
151 billion proteins, **25 TB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

Identified protein-coding gene

# Protein Language Models (PLMs)



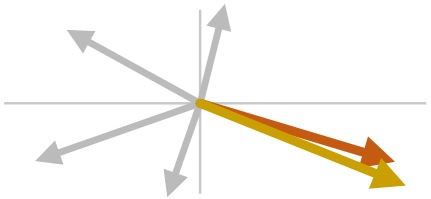
# PLMs vs sequence similarity

**PETase**

**SRR25547275_40149_2**

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

**sequence similarity**



**embedding similarity**

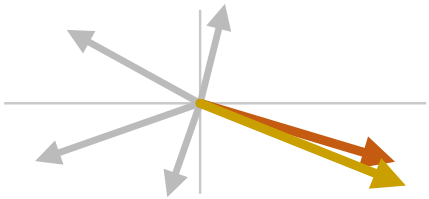
# PLMs vs sequence similarity

**PETase**

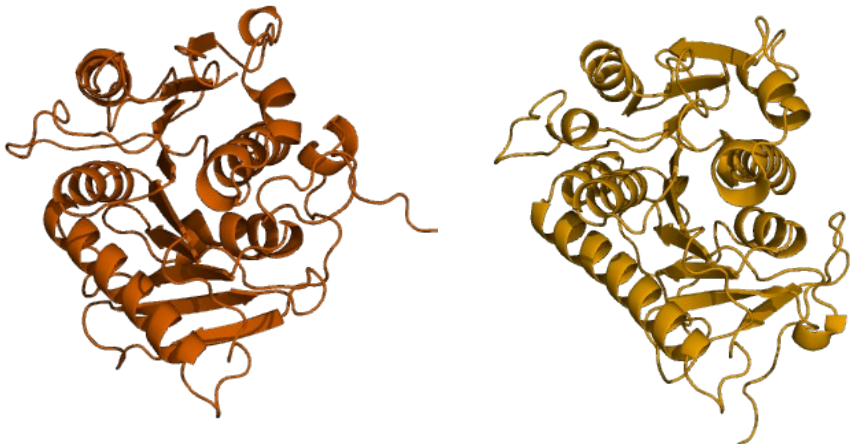
**SRR25547275_40149_2**

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

**sequence similarity**



**embedding similarity**



**similar fold**

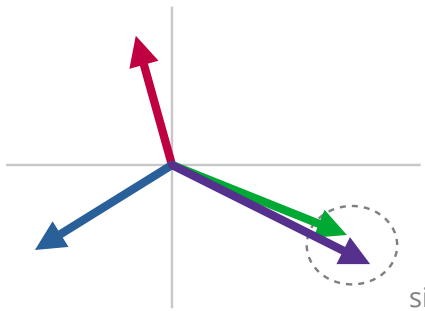
# Let's build a protein search engine

**Logan-proteins**

151 billion proteins, **25 TB**

MRIKGFFITLIAIIFQ
MRIFGFFITLVAIIGQ
MSIYHMKVRTITGKDMTLQP
MTFFLYISPMISILIGFK
⋮

Protein Language Model



similar proteins, close vectors

**Vector database**

# Let's build a protein search engine

**Logan-proteins**

151 billion proteins, **25 TB**

MRIKGFFITLIAIIFQ
MRIFGFFITLVAIIGQ
MSIYHMKVRTITGKDMTLQP
MTFFLYISPMISILIGFK
⋮

Protein Language Model

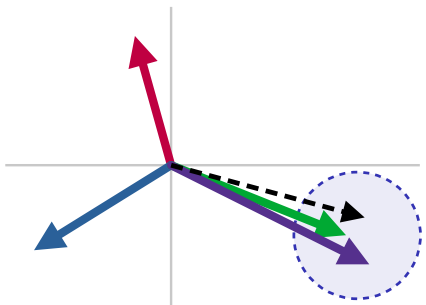


**Vector database**

**Query**

**MRVFGFFIALIAIVGQ**

Protein Language Model



Nearest neighbors

**MRIKGFFITLIAIIFQ**
**MRIFGFFITLVAIIGQ**
+ their SRA datasets

# At 100 embeddings/s, how much time to embed the 151 billion proteins?

A: 5 days

B: 5 months

C: 5 years

D: 50 years

# At 100 embeddings/s, how much time to embed the 151 billion proteins?

A: 5 days

B: 5 months

C: 5 years

D: 50 years

# Embedding 150 billion proteins is too much

- 100 embedding/s on a GPU

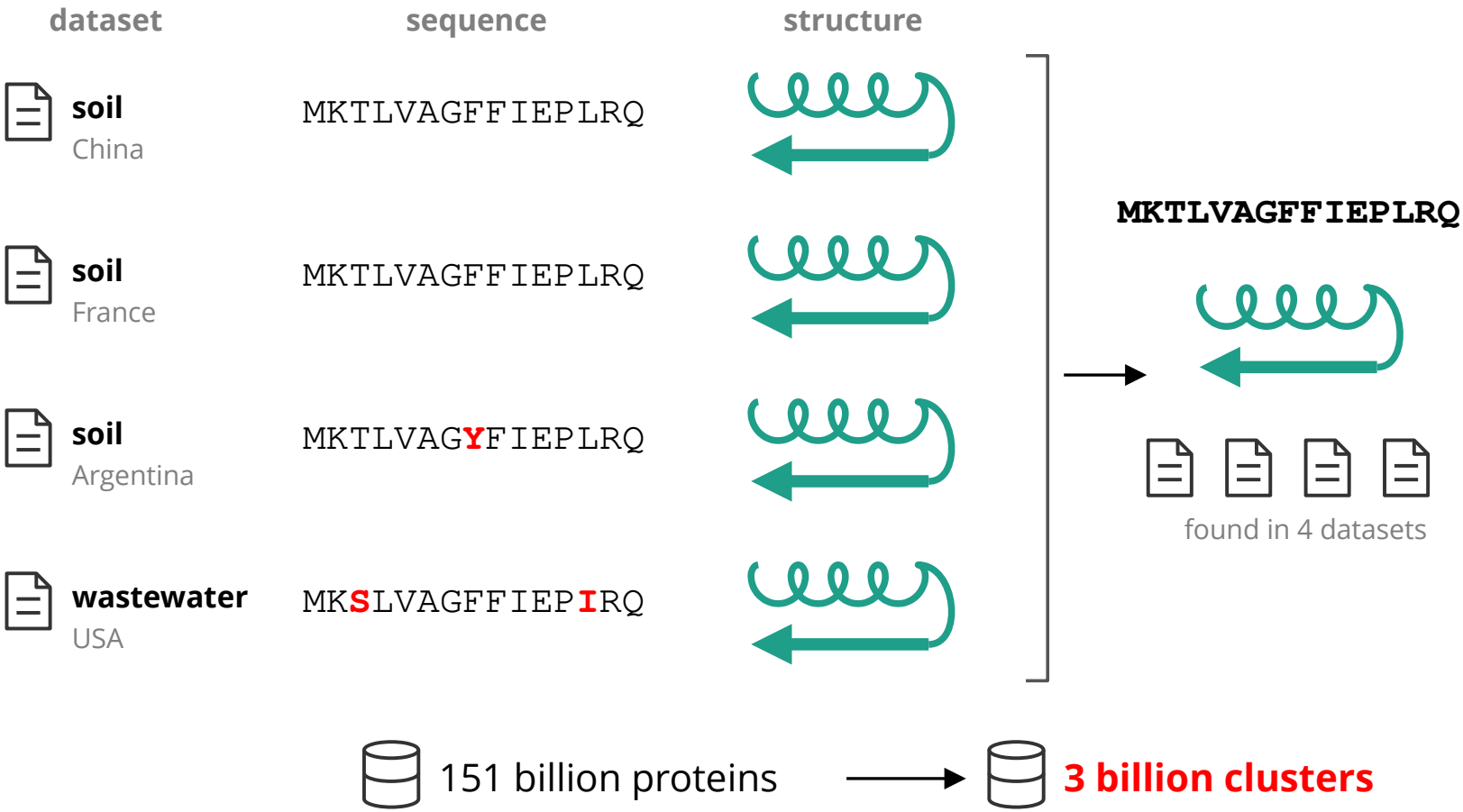
# Embedding 150 billion proteins is too much

- 100 embedding/s on a GPU
- 151 billion embeddings $\sim$ 50 GPU.years

# Embedding 150 billion proteins is too much

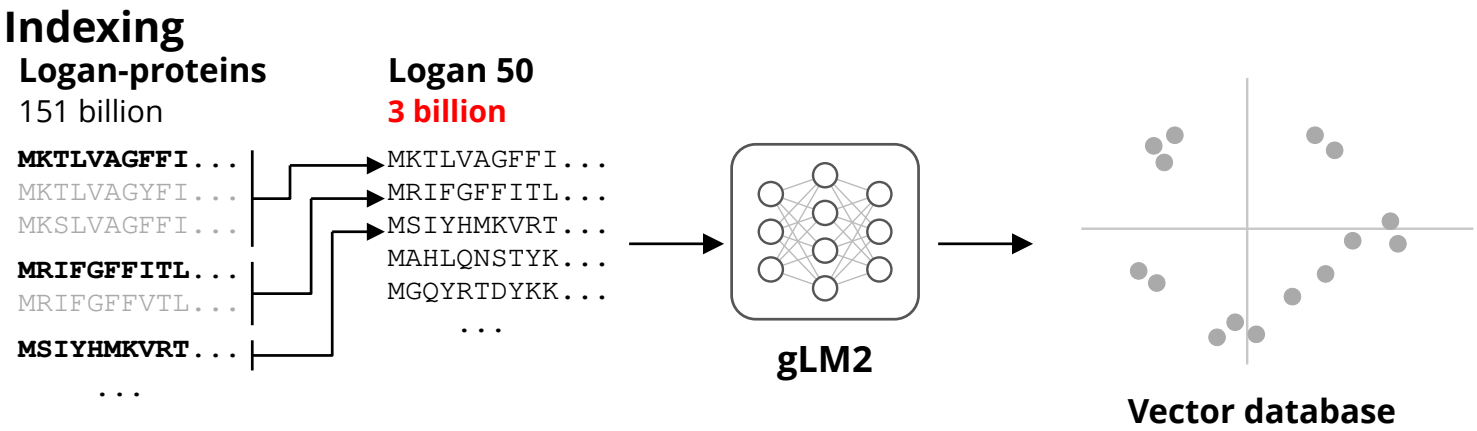
- 100 embedding/s on a GPU
- 151 billion embeddings $\sim$ 50 GPU.years
- $\sim$ \$500,000

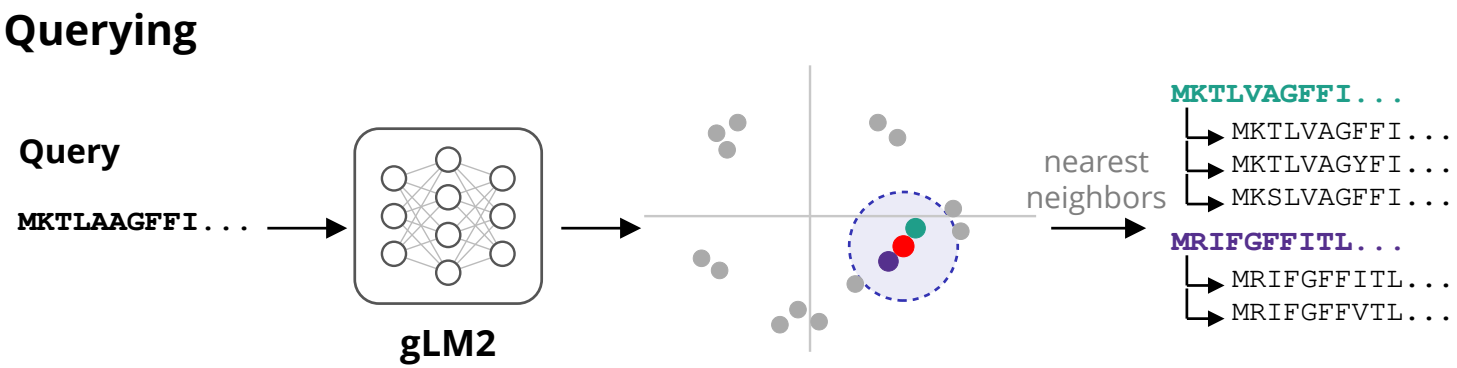
# Cluster the proteins




- Index only the clusters in the vector database

# 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
```

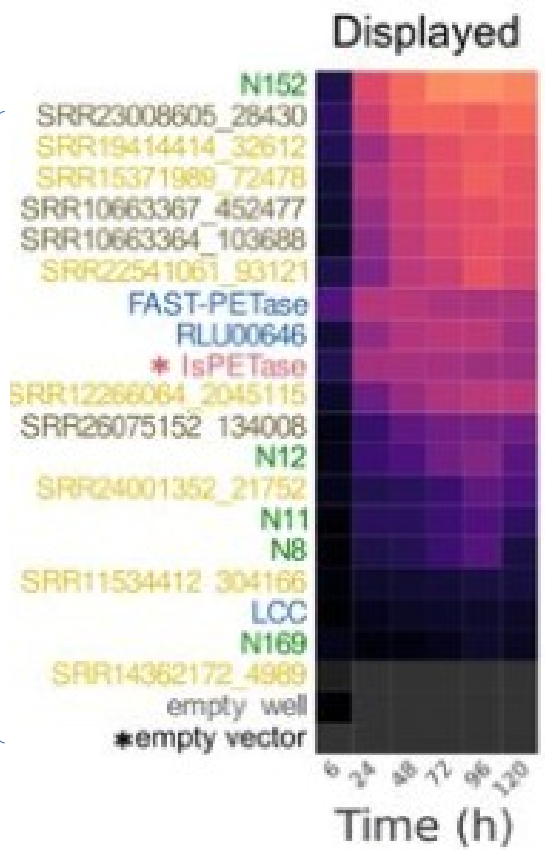
PETase found e.g. in "Fermented Xuecai"




Source: chillcrispbyxueci.substack.com

# Let's test the plastic-degrading activity of the PETase

PETases homologs

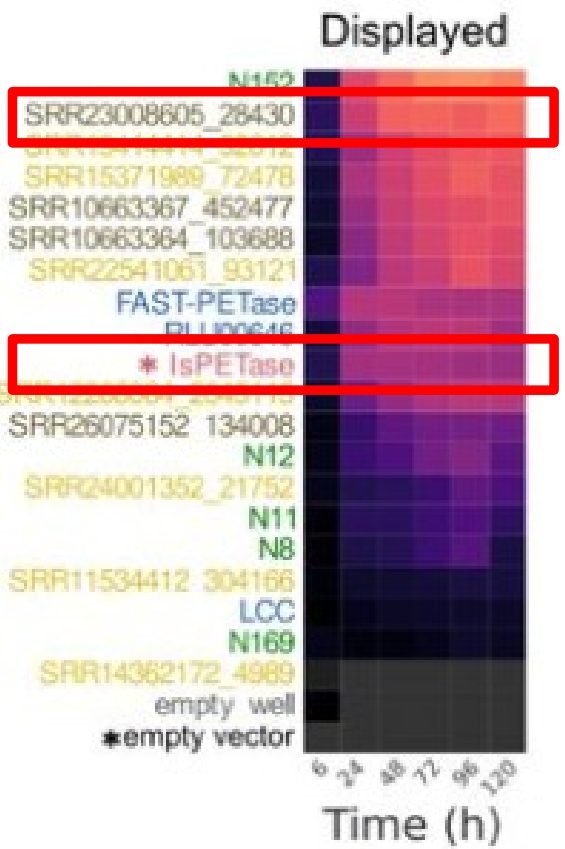



Expressed on cell surface

Plastic-degrading activity

*Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity*, Chikhi et al., 2025

# Let's test the plastic-degrading activity of the PETase




**PETase from SRR23008605 degrades plastic faster than previously known PETase!!**

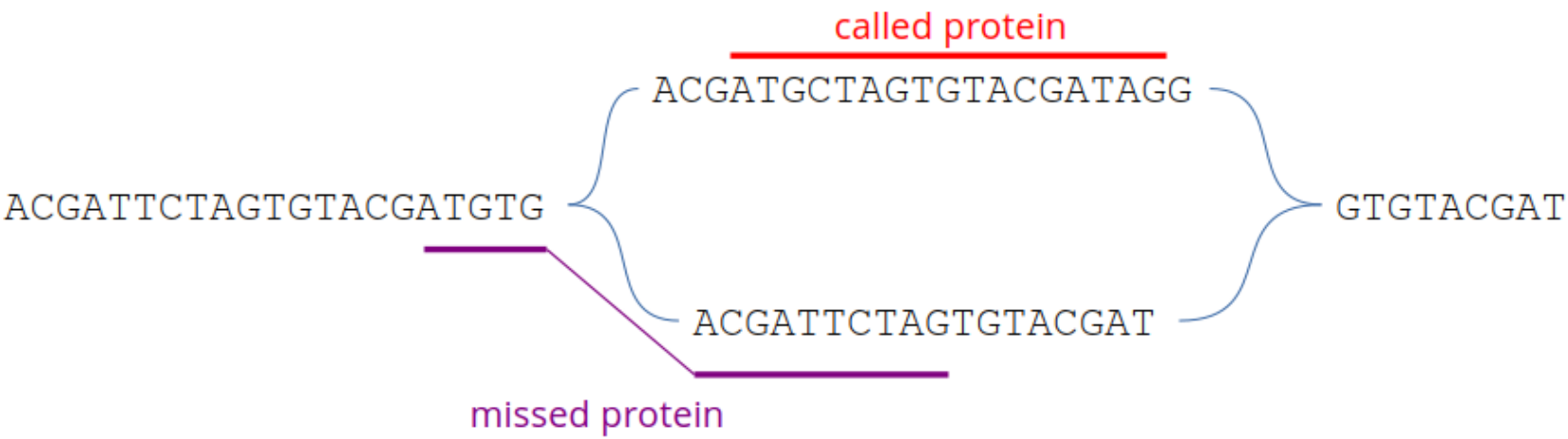
*Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity*, Chikhi et al., 2025

# Future: industrial applications



# Limits of this index

- Only full proteins
- Could be faster (?)
- 90% proteins missing in the database because of the protein calling



# Take-home messages

- Most sequences are not in reference databases, they are in random samples

# Take-home messages

- Most sequences are not in reference databases, they are in random samples
- The sequenced diversity of proteins is huge and difficult to explore

# Take-home messages

- Most sequences are not in reference databases, they are in random samples
- The sequenced diversity of proteins is huge and difficult to explore
- Protein Language Models can be used to represent proteins as vectors

# Take-home messages

- Most sequences are not in reference databases, they are in random samples
- The sequenced diversity of proteins is huge and difficult to explore
- Protein Language Models can be used to represent proteins as vectors
- Web-inspired search engines are a solution, but still not perfect

# 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