SRA-scale search engines: k-mers versus vectors

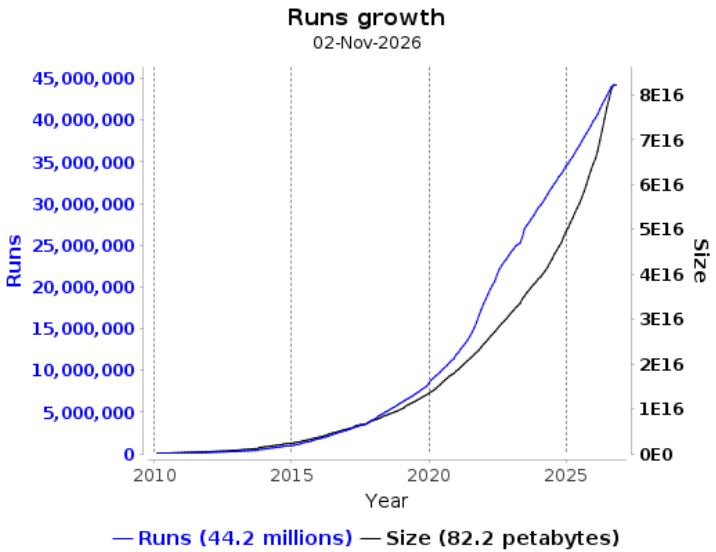
Roland Faure[1]

[1]Institut Pasteur

November 2026

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

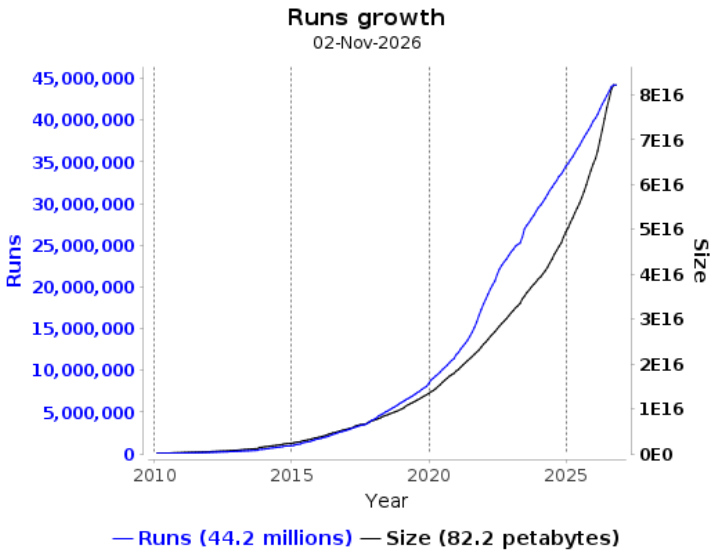
DNA sequencing over the years





- Fast growth of sequencing data

DNA sequencing over the years

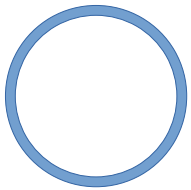




- Fast growth of sequencing data
- More and more **metagenomic** data: random pieces of genomes, under-analyzed
- Reference databases (BLAST nt/nr) contain only a tiny fraction of it

What is indexed in BLAST exactly?

In BLAST nr/nt



Well-assembled genomes

Non-redundant

NOT in BLAST nr/nt

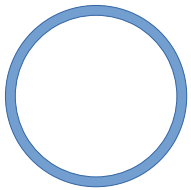


Random pieces of sequenced genomes

Every version of every sequence

What is indexed in BLAST exactly?

In BLAST nr/nt

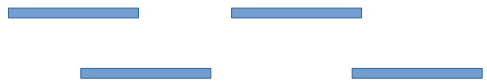


Well-assembled genomes

Non-redundant

0.1%

NOT in BLAST nr/nt

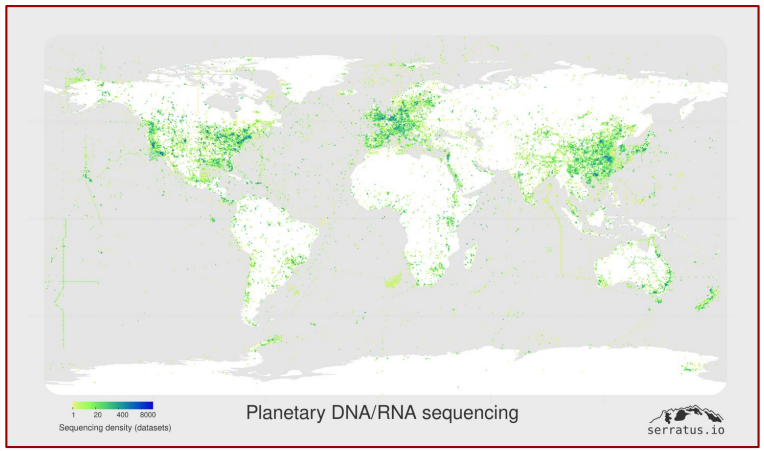


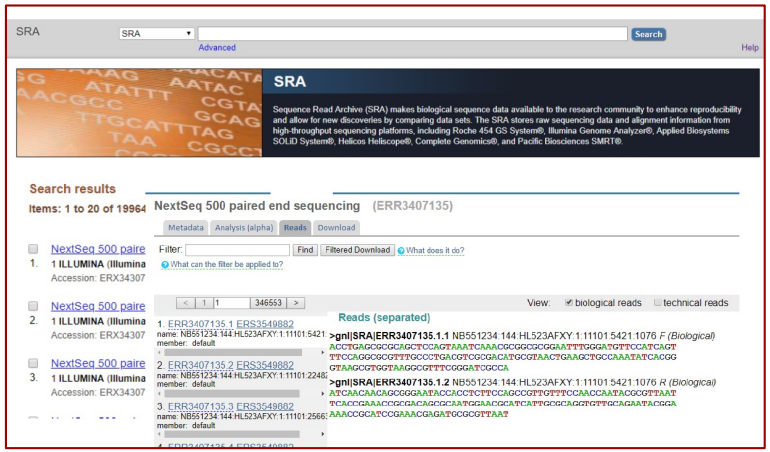
Random pieces of sequenced genomes

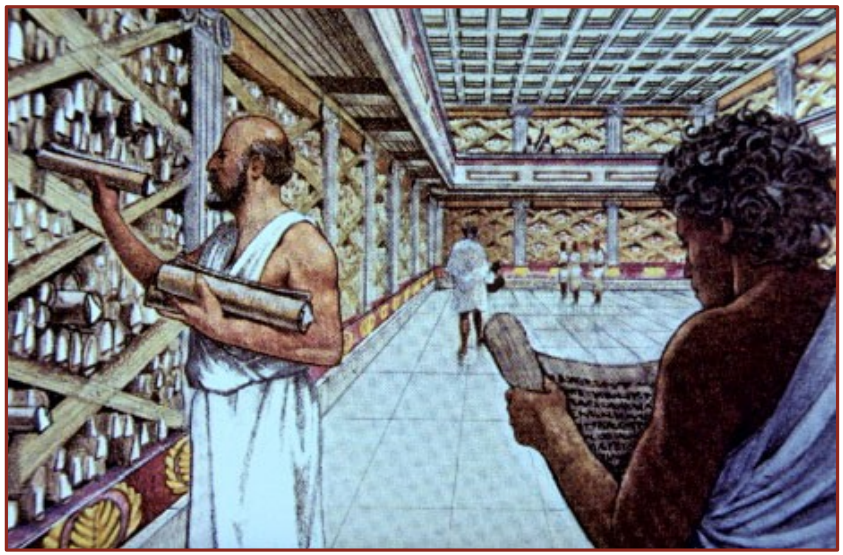
Every version of every sequence

99.9%

The Sequence Read Archive (SRA)



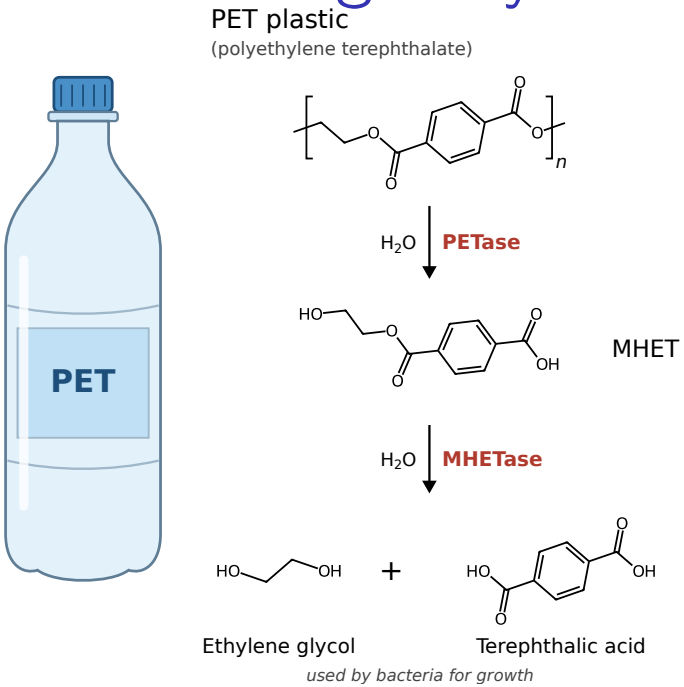




"Library of Alexandria" for genetics

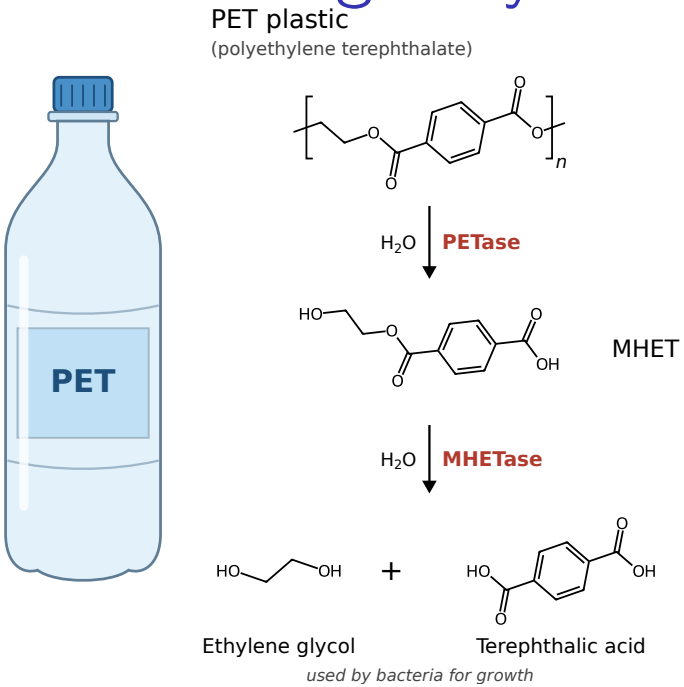
- **All** public sequencing data, 80Pbp

Plastic-eating enzymes: PETases



- PET6: a PET hydrolase from *Vibrio gazogenes*, a halophilic marine bacterium

Plastic-eating enzymes: PETases

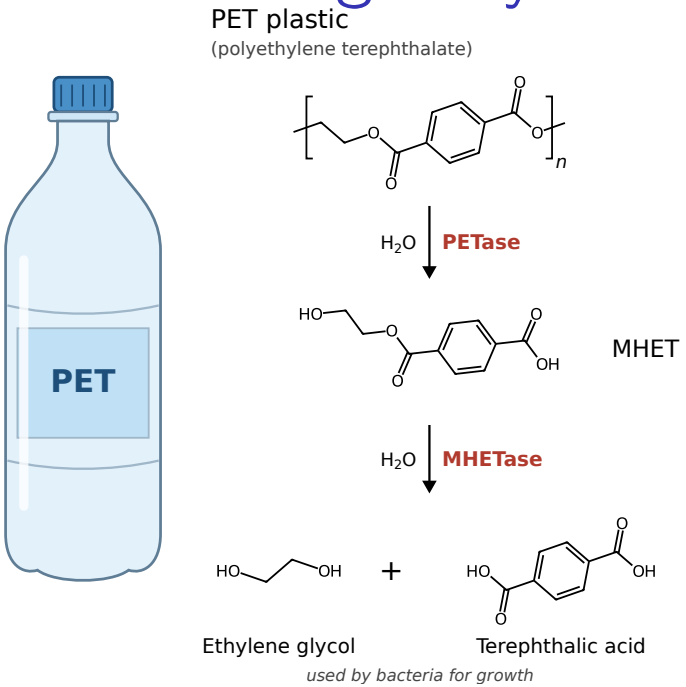


- PET6: a PET hydrolase from *Vibrio gazogenes*, a halophilic marine bacterium
- Found by **homology search** (HMM) in UniProtKB

Danso et al., *Appl. Environ. Microbiol.* 84:e02773-17 (2018)

Weigert et al., *Protein Science* 31:e4500 (2022)

Plastic-eating enzymes: PETases



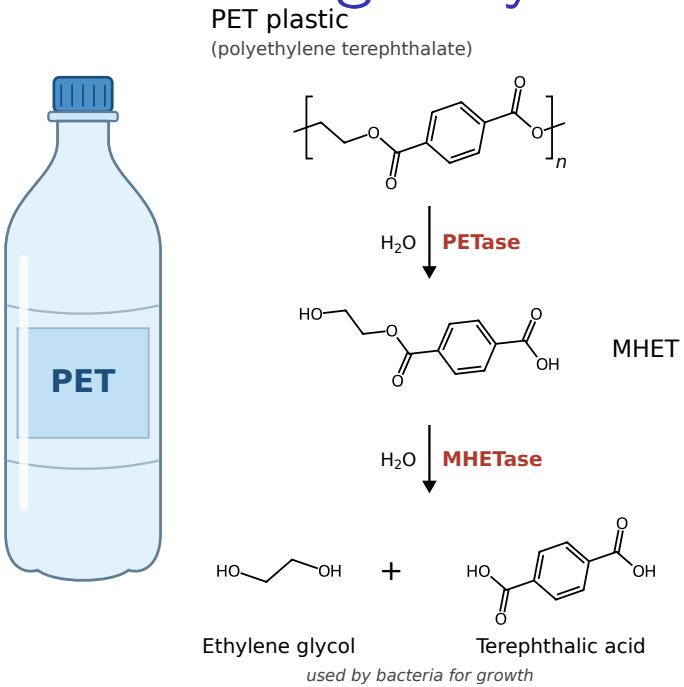
- PET6: a PET hydrolase from *Vibrio gazogenes*, a halophilic marine bacterium
- Found by **homology search** (HMM) in UniProtKB

Danso et al., *Appl. Environ. Microbiol.* 84:e02773-17 (2018)

Weigert et al., *Protein Science* 31:e4500 (2022)

- Found by searching **reference databases**, a tiny fraction of what is sequenced...
- ...so there must be many more PETases in the millions of metagenomic samples!

Plastic-eating enzymes: PETases



- PET6: a PET hydrolase from *Vibrio gazogenes*, a halophilic marine bacterium
- Found by **homology search** (HMM) in UniProtKB

Danso et al., *Appl. Environ. Microbiol.* 84:e02773-17 (2018)

Weigert et al., *Protein Science* 31:e4500 (2022)

- Found by searching **reference databases**, a tiny fraction of what is sequenced...
- ...so there must be many more PETases in the millions of metagenomic samples!
- Let's search over all the SRA/Logan → focus of this talk

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

The Logan project: reducing the size of the data



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

The Logan project: reducing the size of the data

SRA



27 millions accessions, **50 PB**

```
CAGCACGTAGCTAGCTCACGATC
   CACGTAGCTAGCTCACGATCTTT
      GTAGCTAGCTCACGATCTTTAC
```

Logan-contigs



27 millions accessions, **1 PB**

```
CAGCACGTAGCTAGCTCACGATCTTTAC
```

The Logan project: 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

How much did it cost to assemble Logan?

A: $5,000

B: $50,000

C: $500,000

D: $5,000,000

How much did it cost to assemble Logan?

A: $5,000

B: $50,000

C: $500,000

D: $5,000,000

The Logan project

- Goal: make the SRA **accessible**
- All the assemblies and proteins are freely available online

Downloading

To download one accession, type:

```
wget https://s3.amazonaws.com/logan-pub/c/[accession]/[accession].contigs.fa.zst
```

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

The Logan project

- Goal: make the SRA **accessible**
- All the assemblies and proteins are freely available online

Downloading

To download one accession, type:

```
wget https://s3.amazonaws.com/logan-pub/c/[accession]/[accession].contigs.fa.zst
```

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

- Let's search through these databases

Seed-based strategies

- Almost all aligners are **seed-and-extend**: BLAST, DIAMOND, MMseqs2, minimap2, LexicMap...

Seed-based strategies

- Almost all aligners are **seed-and-extend**: BLAST, DIAMOND, MMseqs2, minimap2, LexicMap...
- Index short exact matches (**seeds**, i.e. k-mers) of the database
- Only extend/align candidates sharing seeds with the query



Indexing Logan contigs: logan-search.org

SRR00001	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
SRR00002	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1	0	1	0	0	1	1
SRR00003	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
SRR00004	0	1	1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	0	1	0	0	0	0
SRR00005	1	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	1	0	0
SRR00006	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
SRR00007	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1

- One Bloom filter per dataset (31-mers), 27 million datasets, 1PB index

Indexing Logan contigs: logan-search.org

SRR00001	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
SRR00002	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1	0	1	0	0	1	1
SRR00003	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
SRR00004	0	1	1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	0	1	0	0	0	0
SRR00005	1	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	1	0	0
SRR00006	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
SRR00007	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1

- One Bloom filter per dataset (31-mers), 27 million datasets, 1PB index (kmindex)

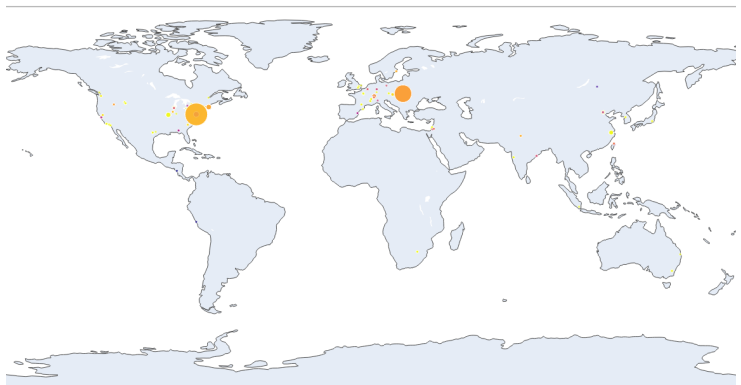
Indexing Logan contigs: logan-search.org

- "Find all datasets that share x% of 31-mers with my query"

Table | Map | Plot | BLAST-like alignment | Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample
SRR16173100 (SRA|OV)	1	PRJNA768258 (SRA|OV)	SAMN22020249 (SRA|OV)
SRR14416387 (SRA|OV)	1	PRJNA727098 (SRA|OV)	SAMN19006852 (SRA|OV)
SRR15166888 (SRA|OV)	1	PRJNA746342 (SRA|OV)	SAMN20203734 (SRA|OV)
ERR3505101 (SRA|OV)	1	PRJEB27179 (SRA|OV)	SAMEA5930711 (SRA|OV)
SRR3740047 (SRA|OV)	1	PRJNA327431 (SRA|OV)	SAMN05335350 (SRA|OV)
ERR2398673 (SRA|OV)	1	PRJEB22684 (SRA|OV)	SAMEA104683220 (SRA|OV)
SRR14655031 (SRA|OV)	1	PRJNA732827 (SRA|OV)	SAMN13221434 (SRA|OV)
SRR13306852 (SRA|OV)	1	PRJNA687219 (SRA|OV)	SAMN17141292 (SRA|OV)
SRR14416320 (SRA|OV)	1	PRJNA727098 (SRA|OV)	SAMN19006784 (SRA|OV)
ERR2399820 (SRA|OV)	1	PRJEB22684 (SRA|OV)	SAMEA104685216 (SRA|OV)
SRR13011679 (SRA|OV)	1	PRJNA668406 (SRA|OV)	SAMN16710006 (SRA|OV)
ERR2395390 (SRA|OV)	1	PRJEB22684 (SRA|OV)	SAMEA104681707 (SRA|OV)
ERR2394036 (SRA|OV)	1	PRJEB22684 (SRA|OV)	SAMEA104680353 (SRA|OV)
SRR25617432 (SRA|OV)	1	PRJNA1004049 (SRA|OV)	SAMN36920974 (SRA|OV)
SRR25659779 (SRA|OV)	1	PRJNA1006165 (SRA|OV)	SAMN37013649 (SRA|OV)
SRR8369137 (SRA|OV)	1	PRJNA511728 (SRA|OV)	SAMN03178087 (SRA|OV)
SRR11318504 (SRA|OV)	1	PRJNA612868 (SRA|OV)	SAMN14389256 (SRA|OV)
SRR25507401 (SRA|OV)	1	PRJNA1001958 (SRA|OV)	SAMN36828017 (SRA|OV)



How much does it cost to query logan-search?

A: $1

B: $10

C: $100

D: $1,000

How much does it cost to query logan-search?

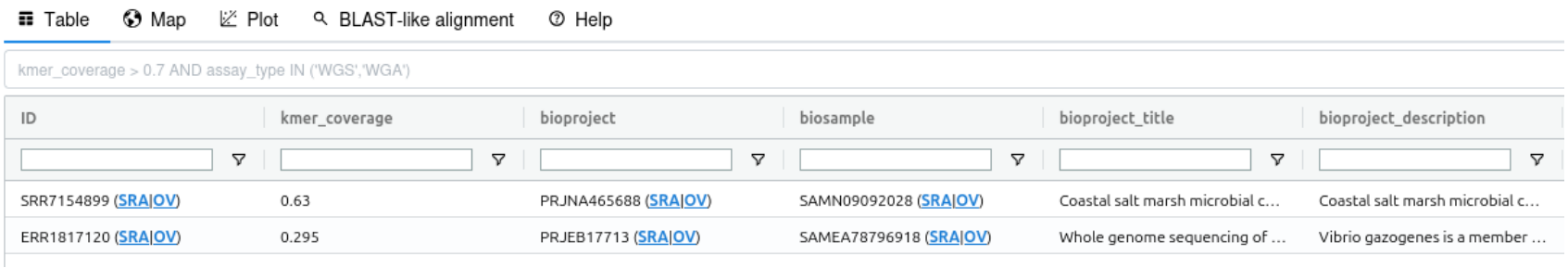
A: $1

B: $10

C: $100

D: $1,000

Searching for PET6 with logan-search.org

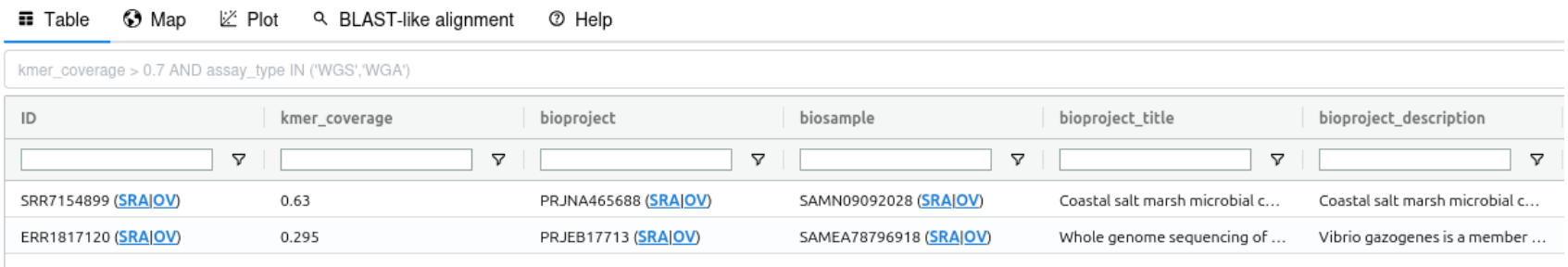

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 (SRA|OV)	0.63	PRJNA465688 (SRA|OV)	SAMN09092028 (SRA|OV)	Coastal salt marsh microbial c…	Coastal salt marsh microbial c…
ERR1817120 (SRA|OV)	0.295	PRJEB17713 (SRA|OV)	SAMEA78796918 (SRA|OV)	Whole genome sequencing of …	Vibrio gazogenes is a member …

- Found a hit in a coastal salt marsh microbial community!

Searching for PET6 with logan-search.org

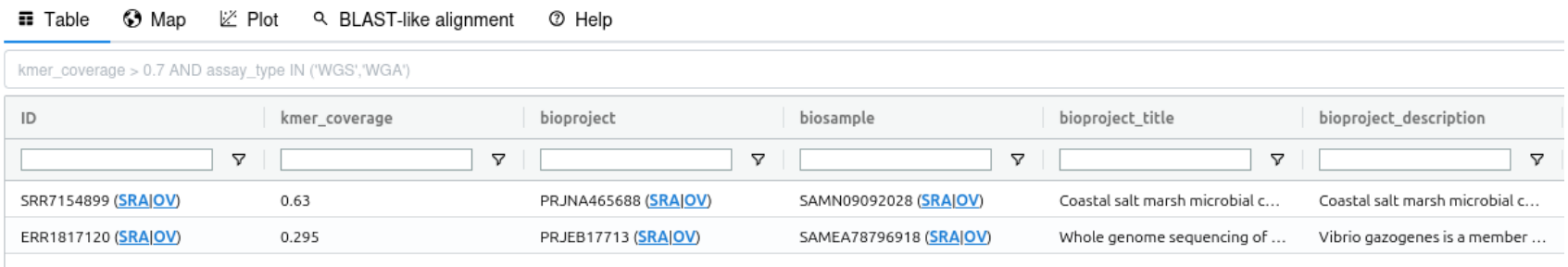

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 (SRA|OV)	0.63	PRJNA465688 (SRA|OV)	SAMN09092028 (SRA|OV)	Coastal salt marsh microbial c…	Coastal salt marsh microbial c…
ERR1817120 (SRA|OV)	0.295	PRJEB17713 (SRA|OV)	SAMEA78796918 (SRA|OV)	Whole genome sequencing of …	Vibrio gazogenes is a member …

- Found a hit in a coastal salt marsh microbial community!
- Wait, only two hits !??

Searching for PET6 with logan-search.org


Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 (SRA|OV)	0.63	PRJNA465688 (SRA|OV)	SAMN09092028 (SRA|OV)	Coastal salt marsh microbial c…	Coastal salt marsh microbial c…
ERR1817120 (SRA|OV)	0.295	PRJEB17713 (SRA|OV)	SAMEA78796918 (SRA|OV)	Whole genome sequencing of …	Vibrio gazogenes is a member …

- Found a hit in a coastal salt marsh microbial community!
- Wait, only two hits !??
- Only finds very high identity sequences (shared 31-mers)

More sensitive strategy: LexicMap

Article | Open access | Published: 10 September 2025

Efficient sequence alignment against millions of prokaryotic genomes with LexicMap

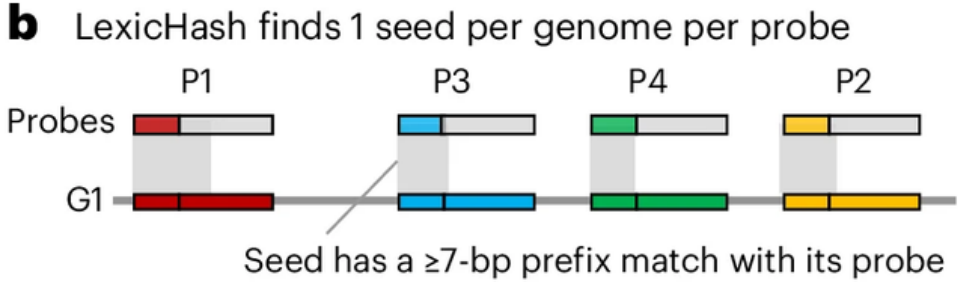
Wei Shen ✉, John A. Lees & Zamin Iqbal ✉

More sensitive strategy: LexicMap

Article | Open access | Published: 10 September 2025

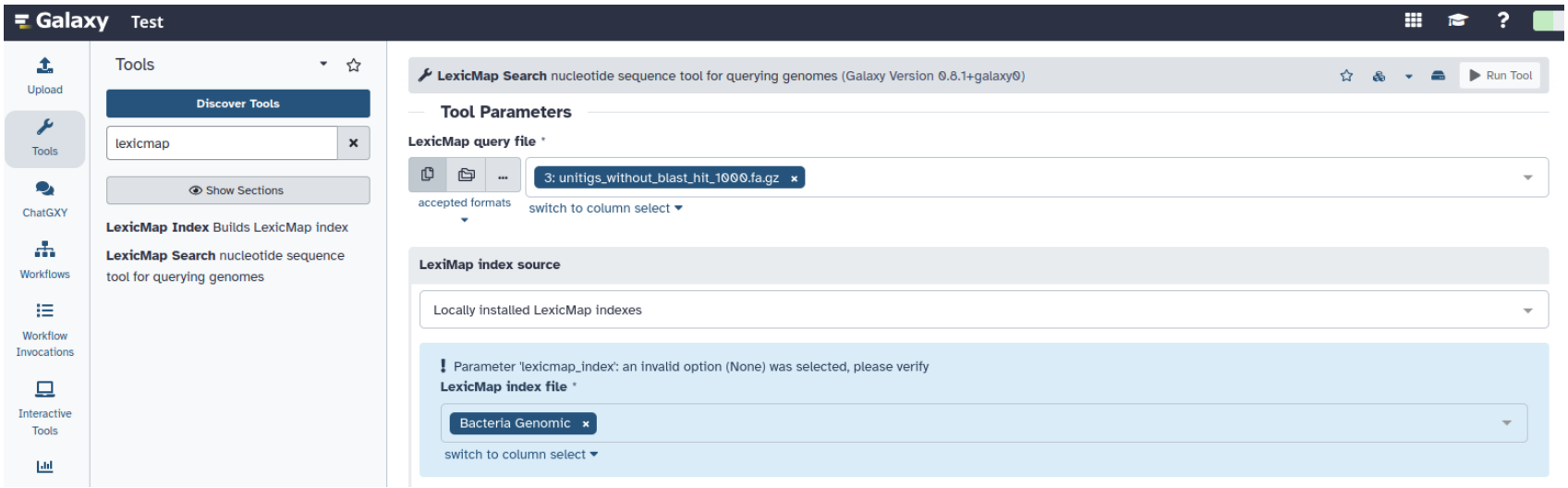
Efficient sequence alignment against millions of prokaryotic genomes with LexicMap

Wei Shen ✉, John A. Lees & Zamin Iqbal ✉



- Detects sequences with $\geq 70\%$ similarity

LexicMap-Logan



- Available on Galaxy
- Full alignment against all Logan metagenomes

Searching for PET6 with LexicMap-Logan

- One hour computation to return results
- 200 results

Searching for PET6 with LexicMap-Logan

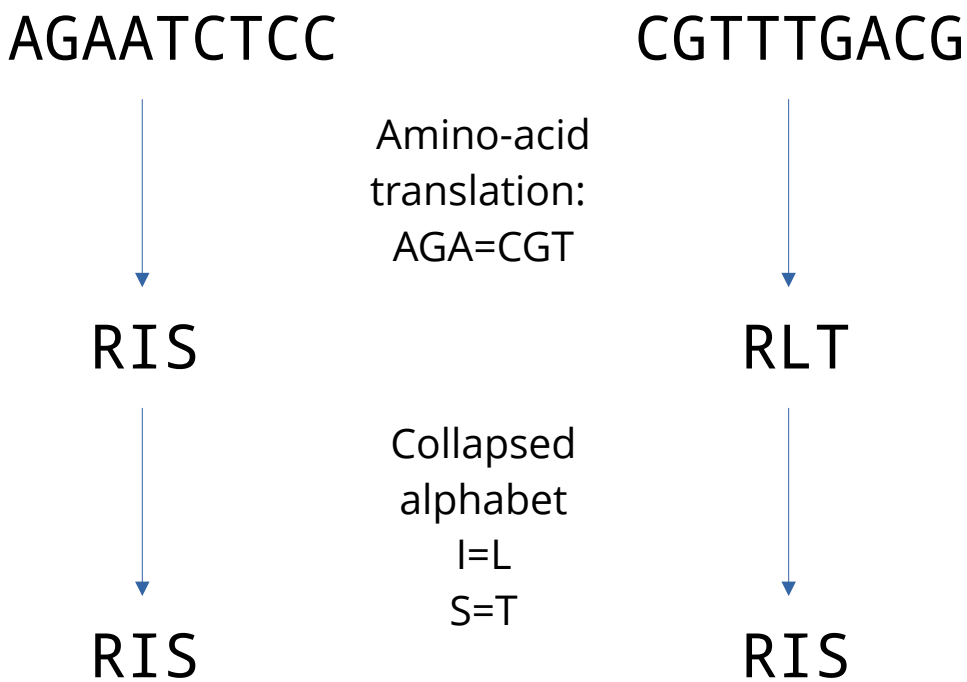
- One hour computation to return results
- 200 results
- PETase found e.g. in “Fermented Xuecai”



Source: chillcrispbyxueci.substack.com

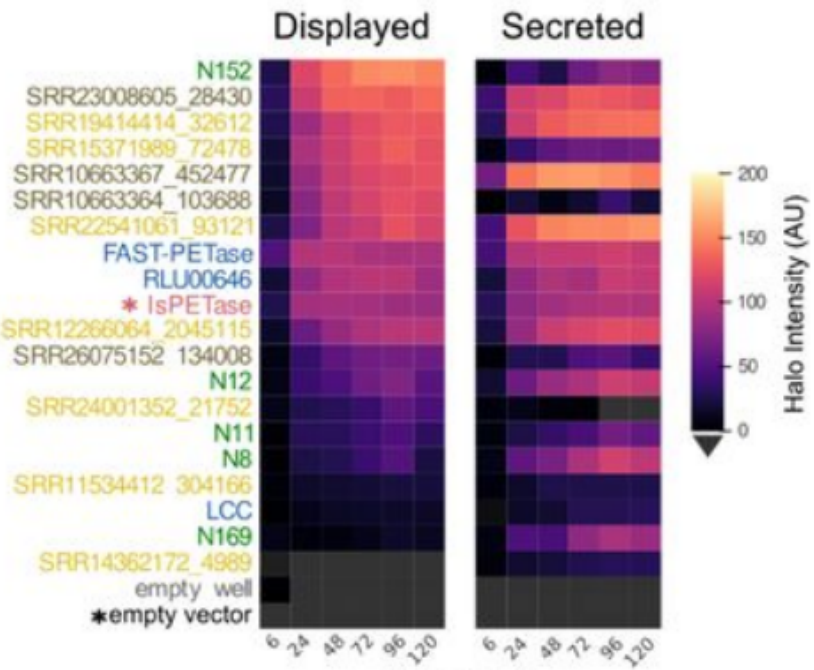
Even more sensitive strategy: DIAMOND

- amino-acid space
- reduced alphabet



Searching for PET6 with DIAMOND

- Logan paper: aligned against all 0.9 Pbp of Logan contigs
- 358,694 aligned proteins
- Few hundred candidates selected for testing
- Some discovered enzymes better than engineered ones



How much did it cost to DIAMOND on Logan?

A: $10

B: $100

C: $1,000

D: $10,000

How much did it cost to DIAMOND on Logan?

A: $10

B: $100

C: $1,000

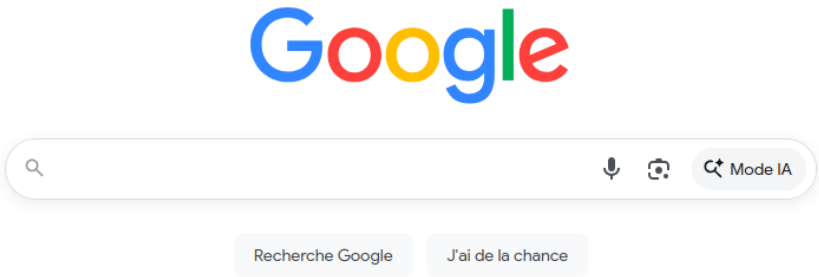
D: $10,000

Seed-based strategies: strength and limits

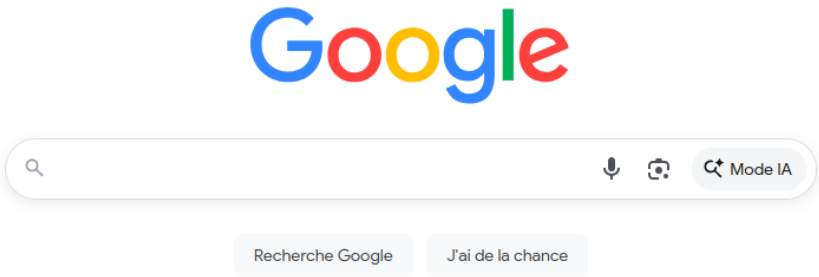
Tool	Seeds	Cost/query	Sensitivity
logan-search	31-mers (nt)	minutes, ∼$1	near-identical
LexicMap-Logan	k-mers (nt)	hours, ∼$10	$\geq$ 70% nucl id.
DIAMOND	aa seeds	weeks, ∼$10,000	$\geq$ 50% aa id.

- Very fast when low sensitivity, much slower when higher sensitivity
- Based on local matches, much stricter than e.g. structural similarity.

Web searches



Web searches



- Historically: inverted index of **words** ($\approx$ k-mers!) + ranking
- Now (since $\sim$2019): **vector search**

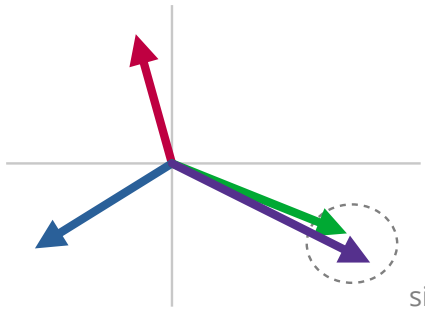
How do web search engine work?

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

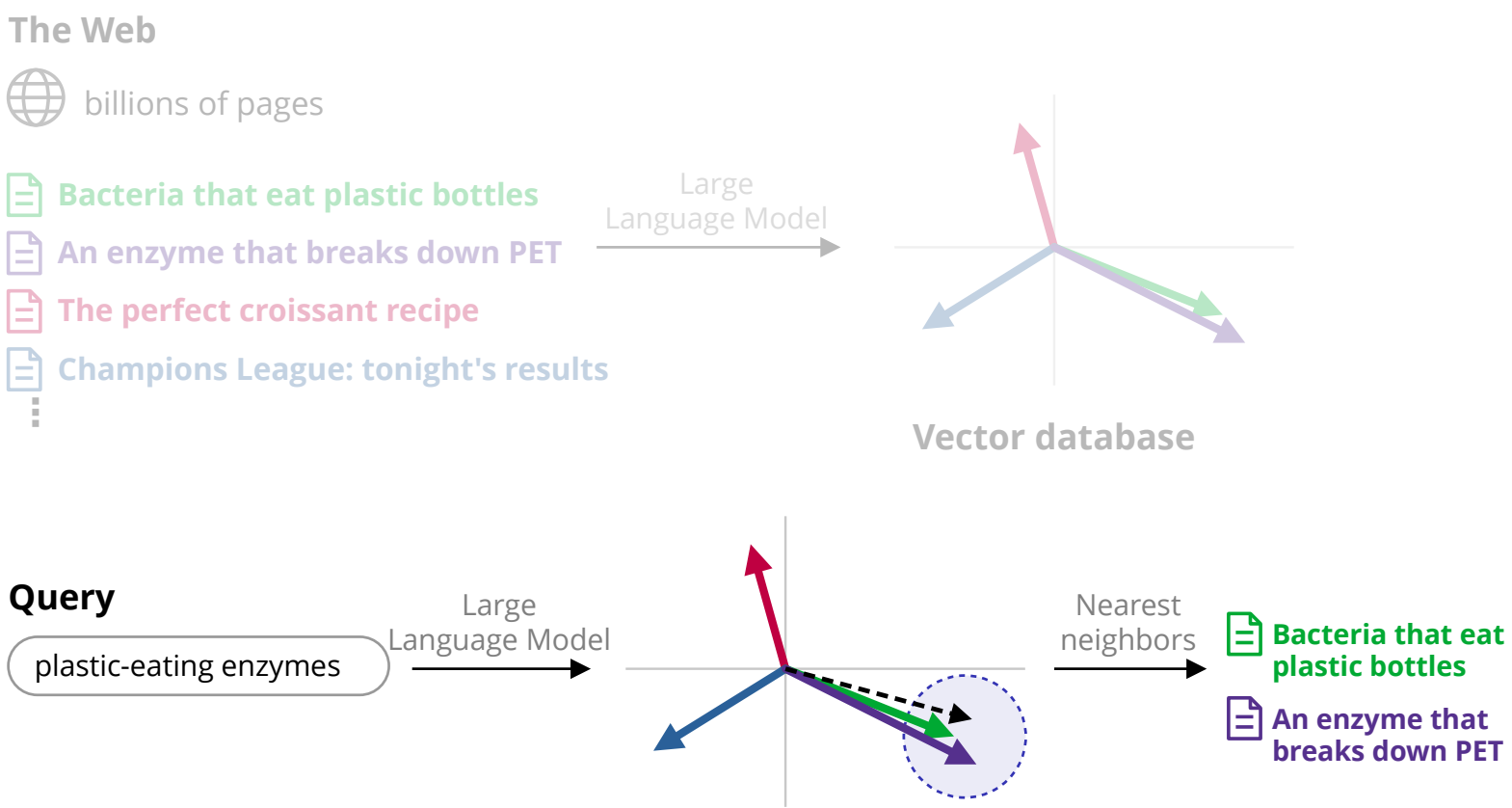
Large Language Model



similar pages, close vectors

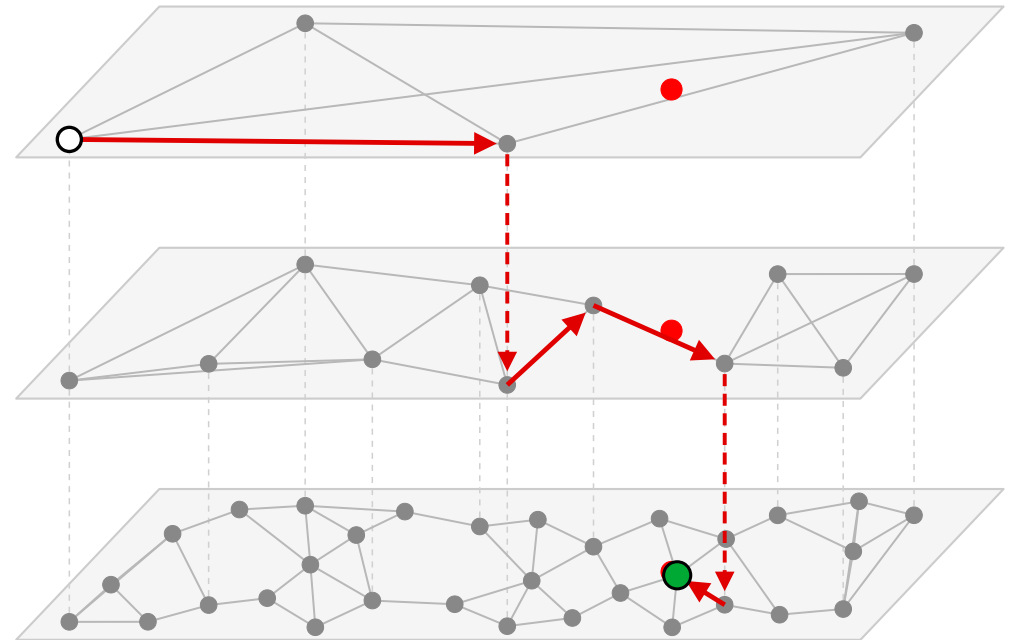
Vector database

How do web search engine work?



How do web search engine work? Fast vector search

Hierarchical Navigable Small World (HNSW) index



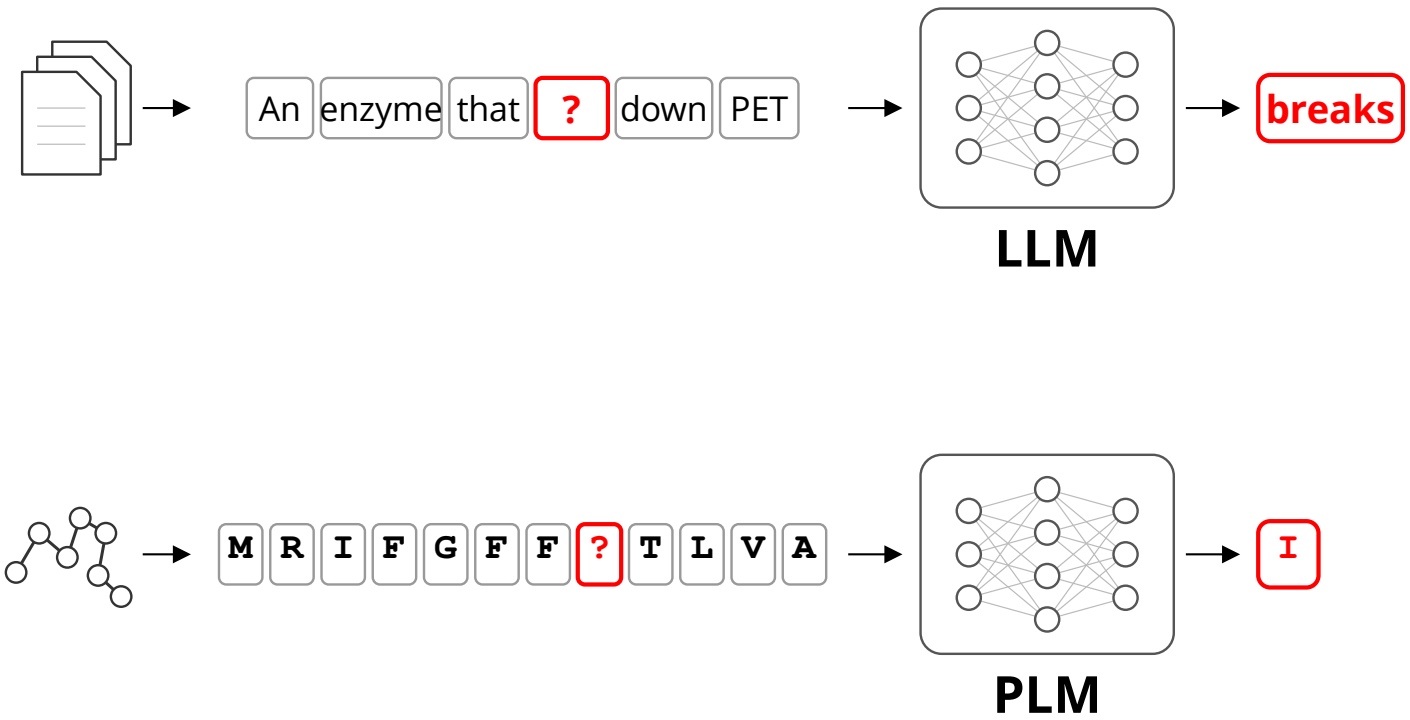
index level 2
few vectors

index level 1
more vectors

index level 0
all vectors

start query nearest page

Protein Language Models (PLMs)



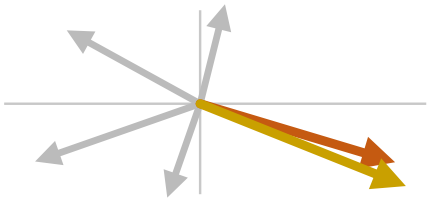
PLMs capture more than sequence similarity

PETase

SRR25547275_40149_2

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

sequence similarity



embedding similarity

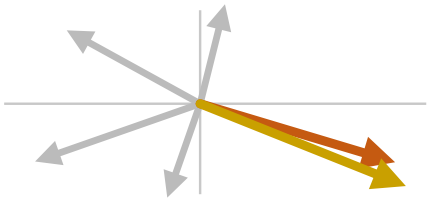
PLMs capture more than sequence similarity

PETase

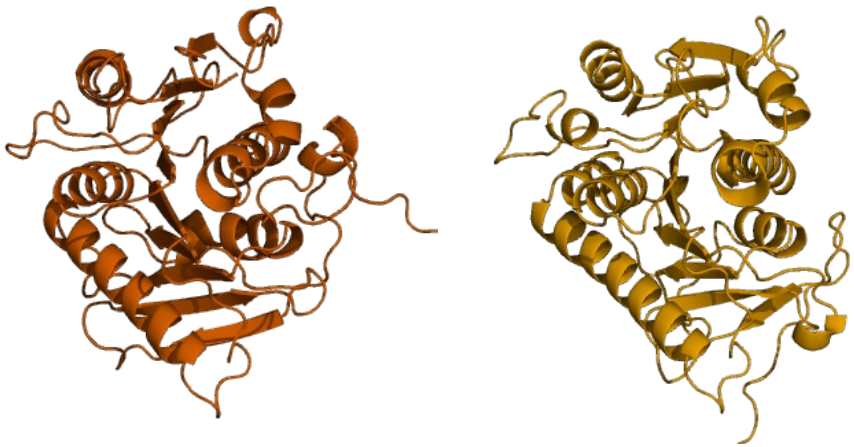
SRR25547275_40149_2

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

sequence similarity

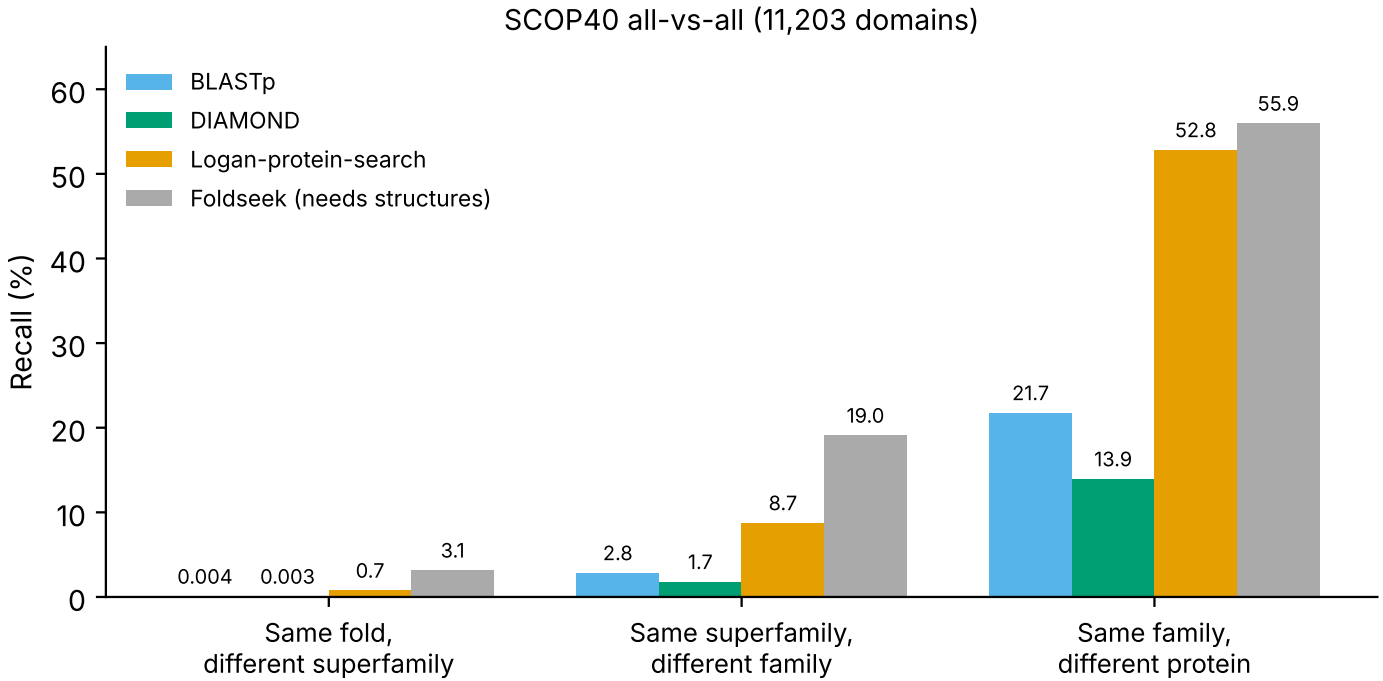


embedding similarity



similar fold

Protein Language Models are more sensitive than alignment-based metrics



- All-vs-all on SCOP40, pairs with known structural relationship
- Foldseek is best, but needs a structure per protein → does not scale to billions

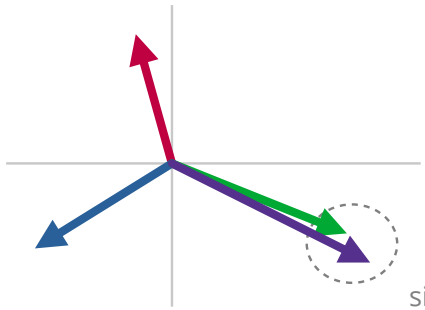
Logan-Protein-Search

Logan-proteins

151 billion proteins, **25 TB**

MRIKGFFITLIAIIFQ
MRIFGFFITLVAIIGQ
MSIYHMKVRTITGKDMTLQP
MTFFLYISPMISILIGFK
⋮

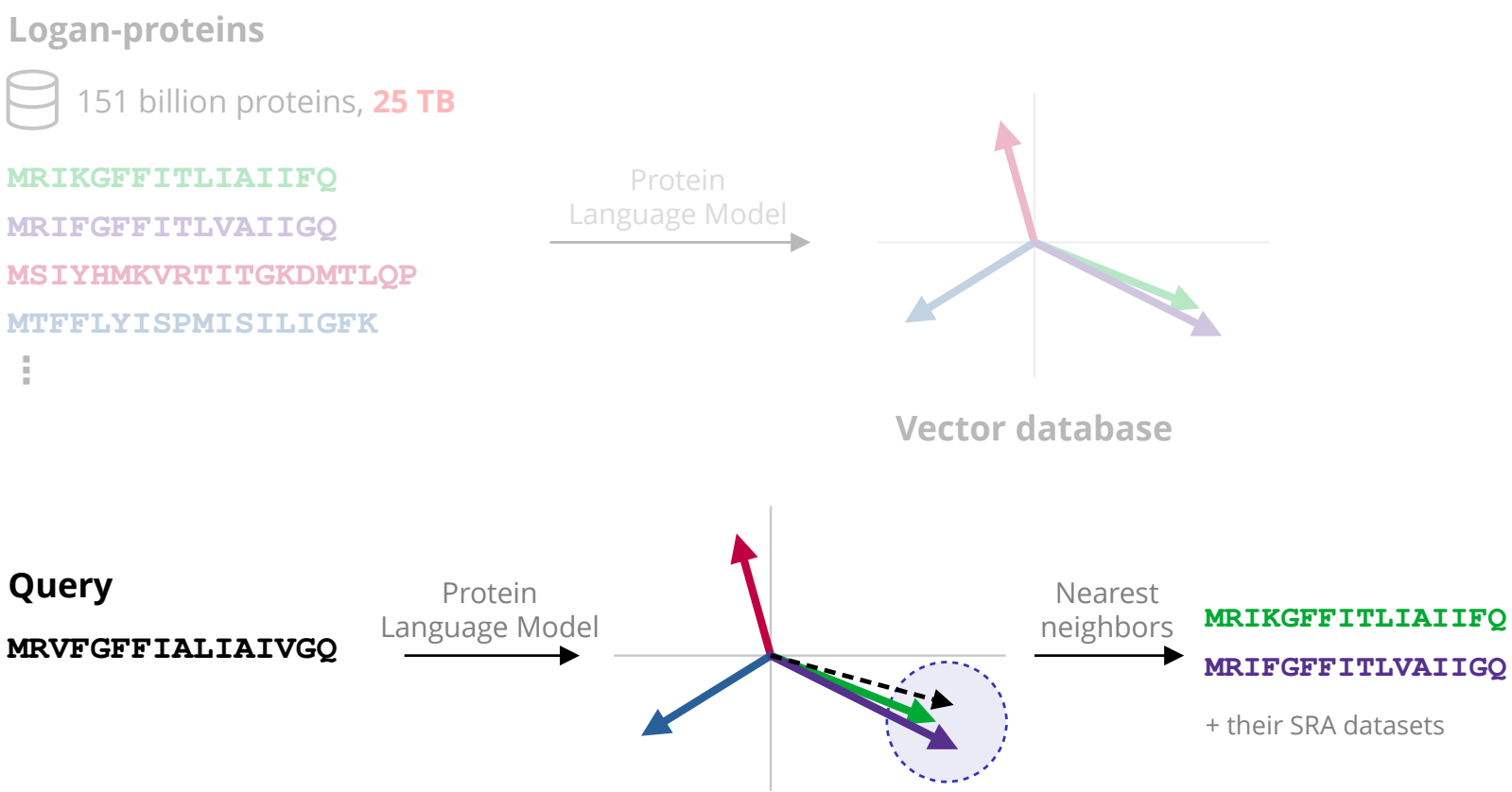
Protein
Language Model



similar proteins,
close vectors

Vector database

Logan-Protein-Search



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?

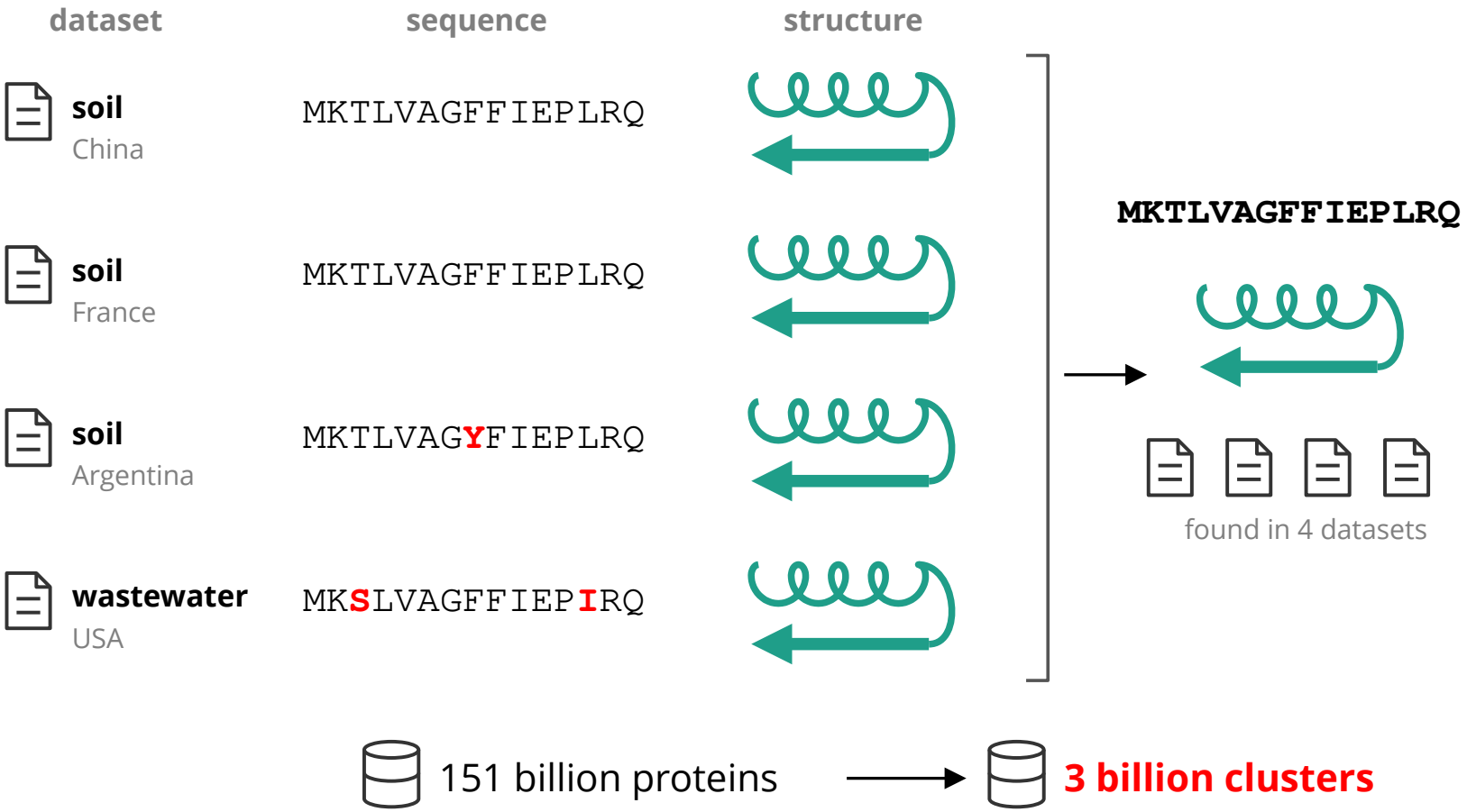
A: 5 days

B: 5 months

C: 5 years

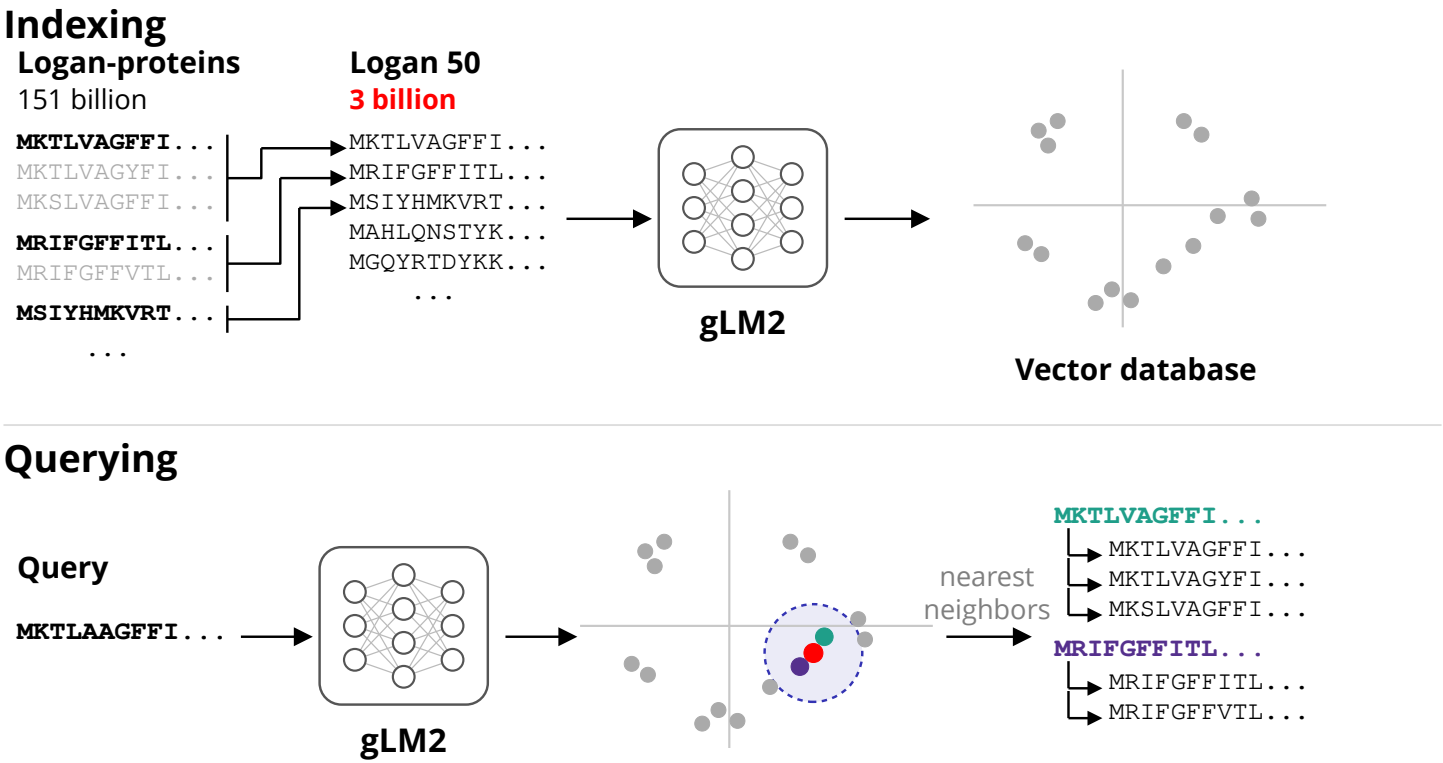
D: 50 years

Clustering the proteins to reduce the problem



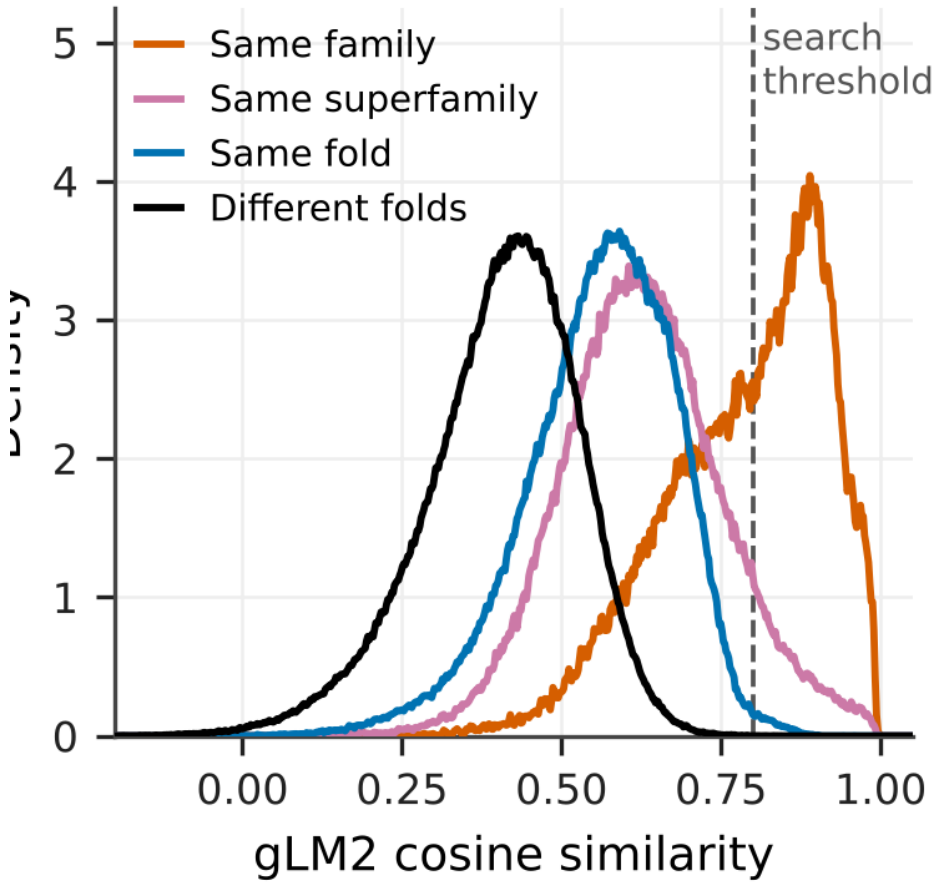
- Index only the clusters in the vector database

Logan-Protein-Search



- gLM2 embeddings (512 dimensions), 3 billion vectors
- Vector database: usearch

Choosing the cosine threshold



Searching for the PETase with Logan-protein-search

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

Searching for the PETase with Logan-protein-search

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

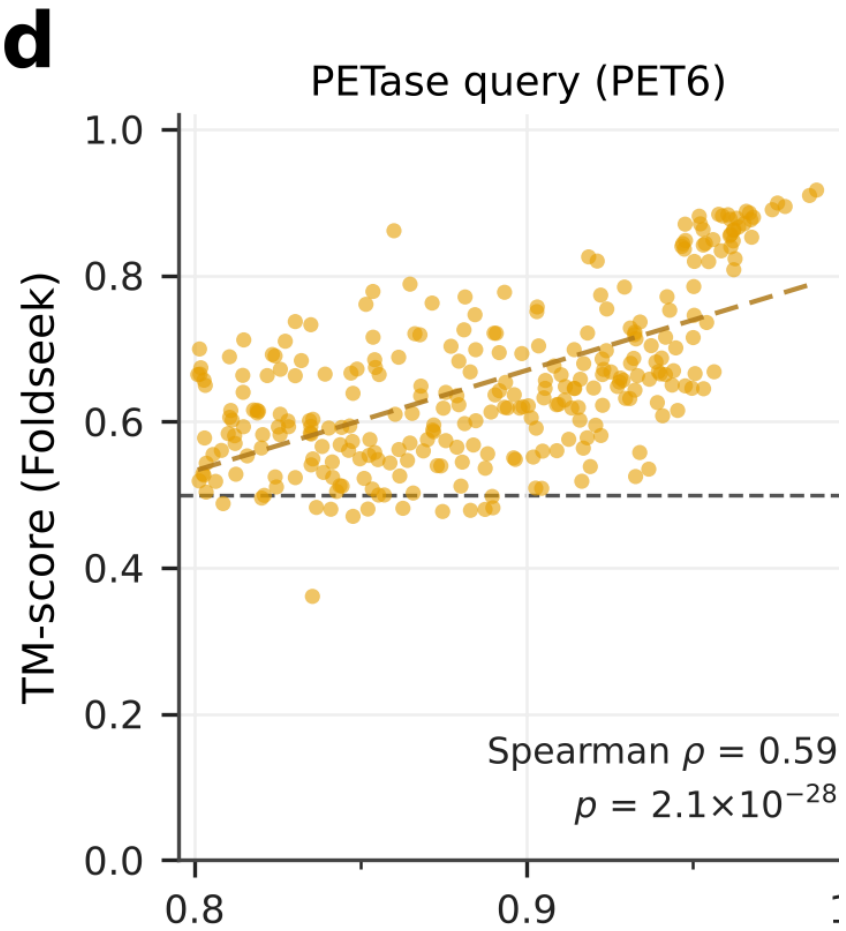
```
Query_10001  150  YHQIDEKEIGIVGYSQGGAGAYNTLEGKDGDK---FKTM-VTVSG---    ------VTESIGEKLHLPVWIYDPSKVTI  213
Query_10002  98   QPEVDANRVALIGWALGGGVVVAEAADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD-DMVRDRPERARSGRSRTV  176
Query_10003  95   EK-TGNPRVGVVGYCAGGGLALMLAAQRP-----------DAVKAVA   PYY---------GLIPWPGAQPDWAAIDA  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
```

Searching for the PETase with Logan-protein-search: cost

Method	Database	CPU.h (PET6)
DIAMOND blastx	Logan contigs (0.9 Pbp)	52,000
BLASTp	nr	6
Logan-protein-search	Logan-proteins (141 B)	5

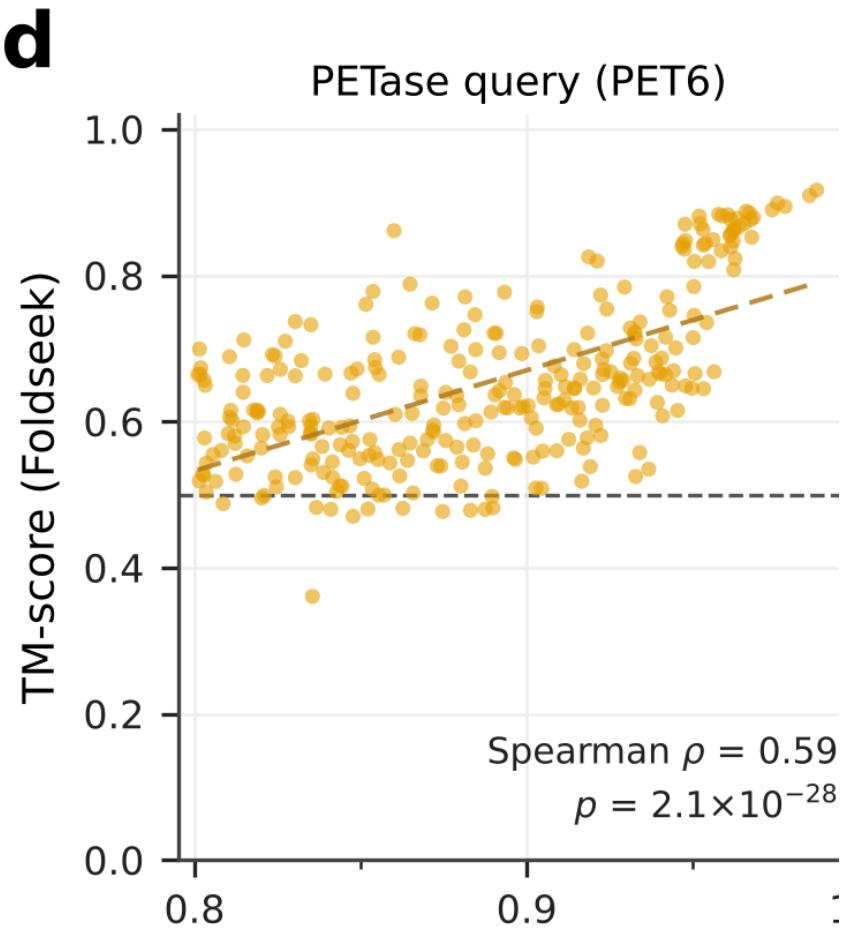
- $\sim 10^4\times$ cheaper than exhaustive alignment

Searching for the PETase with Logan-protein-search: structures



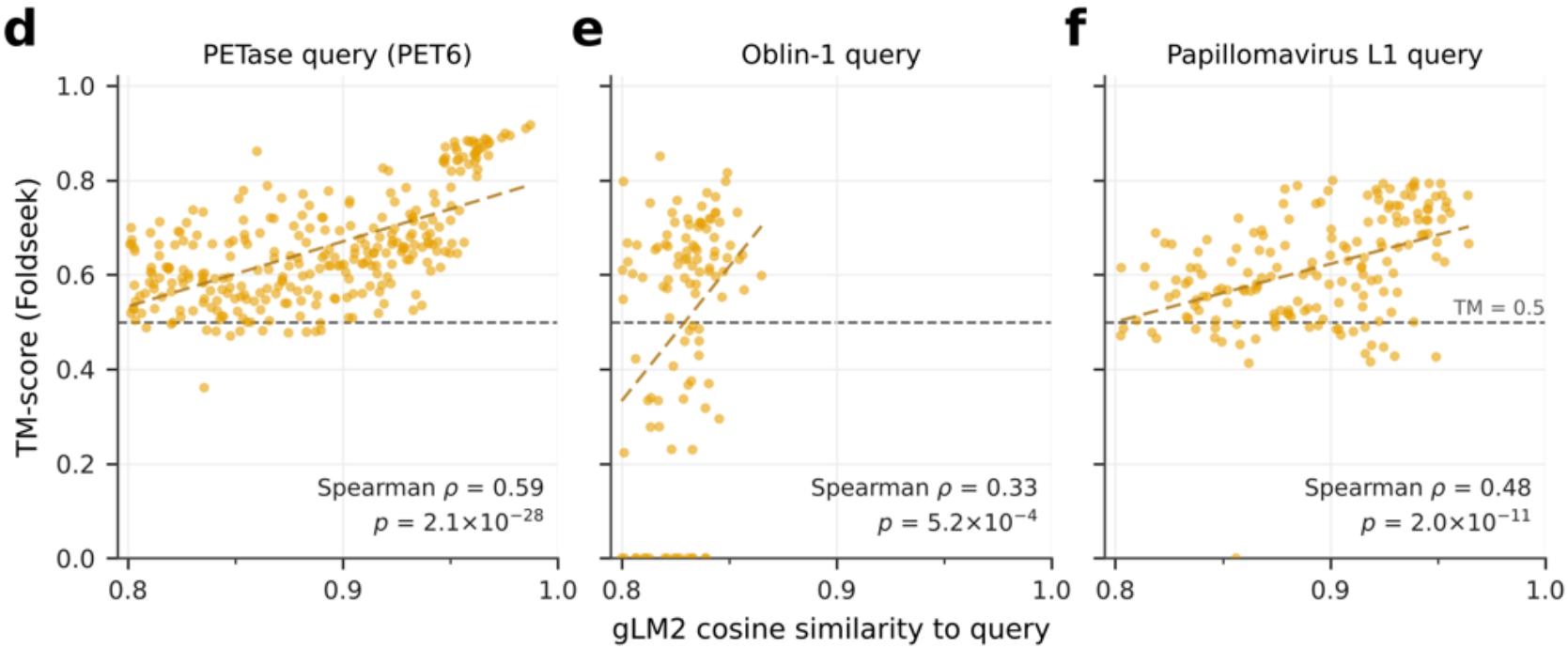
- 300 hits sampled along the ranking, folded with ColabFold
- TM-score vs. PETase query

Searching for the PETase with Logan-protein-search: structures



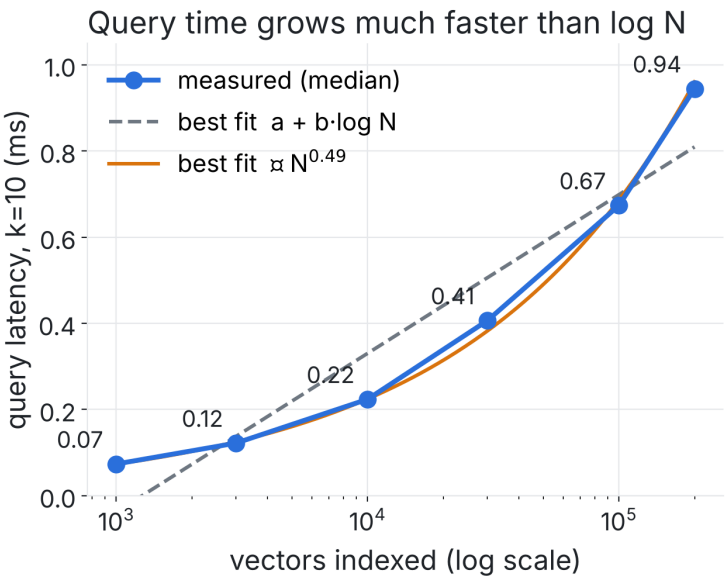
- 300 hits sampled along the ranking, folded with ColabFold
- TM-score vs. PETase query
- Most hits have the same fold (TM $>$ 0.5)
- Embedding similarity correlates with structural similarity

Limits of vector search: dependance on the PLM



- Oblin results not as good at all
- Oblins are very rare → poorly learned by the PLM

Limits of vector search: in practice does not scale that well (yet?)



- usearch: Hierarchical Navigable Small World (HNSW) index
- Speed scales like sqrt(# vectors)
- RAM scales lineary with # vectors $\rightarrow$ does not scale to trillion
- Bottleneck is RAM bandwidth

Conclusion: k-mers vs vectors



k-mers vs vectors - Proteins

- PLMs embed much richer information than k-mers
- In the short-term future we have no big scalability issues
- Simpler datastructure & algorithms
- Strong overlap with other fields of research

k-mers vs vectors - Nucleotide

JOURNAL ARTICLE

Embed-Search-Align: DNA sequence alignment using Transformer models

Pavan Holur ✉, K C Enevoldsen, Shreyas Rajesh, Lajoyce Mboning, Thalia Georgiou, Louis-S Bouchard, Matteo Pellegrini, Vwani Roychowdhury Author Notes

Bioinformatics, Volume 41, Issue 3, March 2025, btaf041, https://doi.org/10.1093/

New Results

Follow this preprint

Anniemap: Vector Search for Viral Short Read Alignment

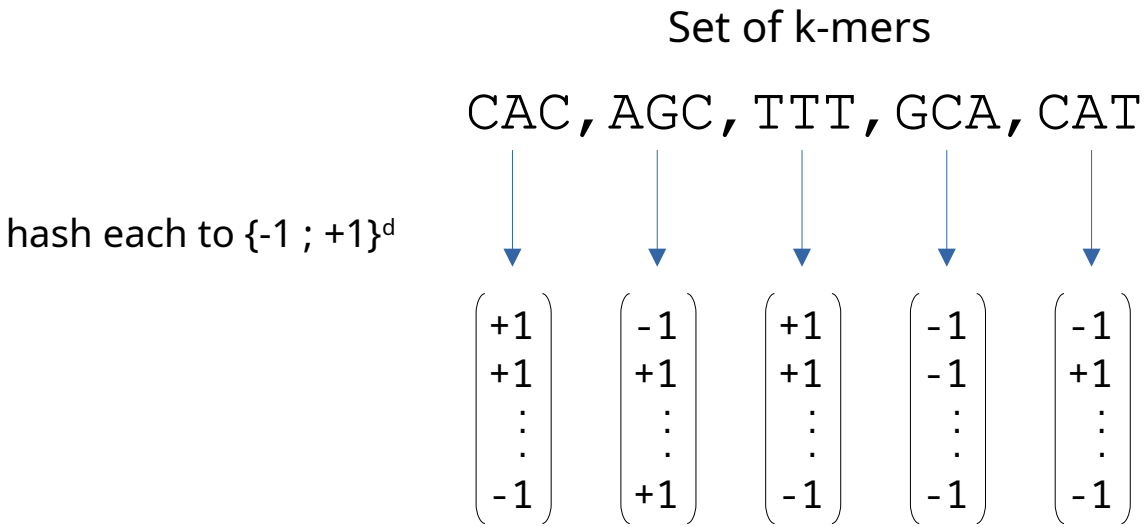
Daniel J. van Zyl, Houriiyah Tegally, Cheryl Baxter, The INFORM Africa research study group, Tulio de Oliveira, Joicymara S. Xavier, Marcel Dunaiski

doi: https://doi.org/10.64898/2026.08.26.747390

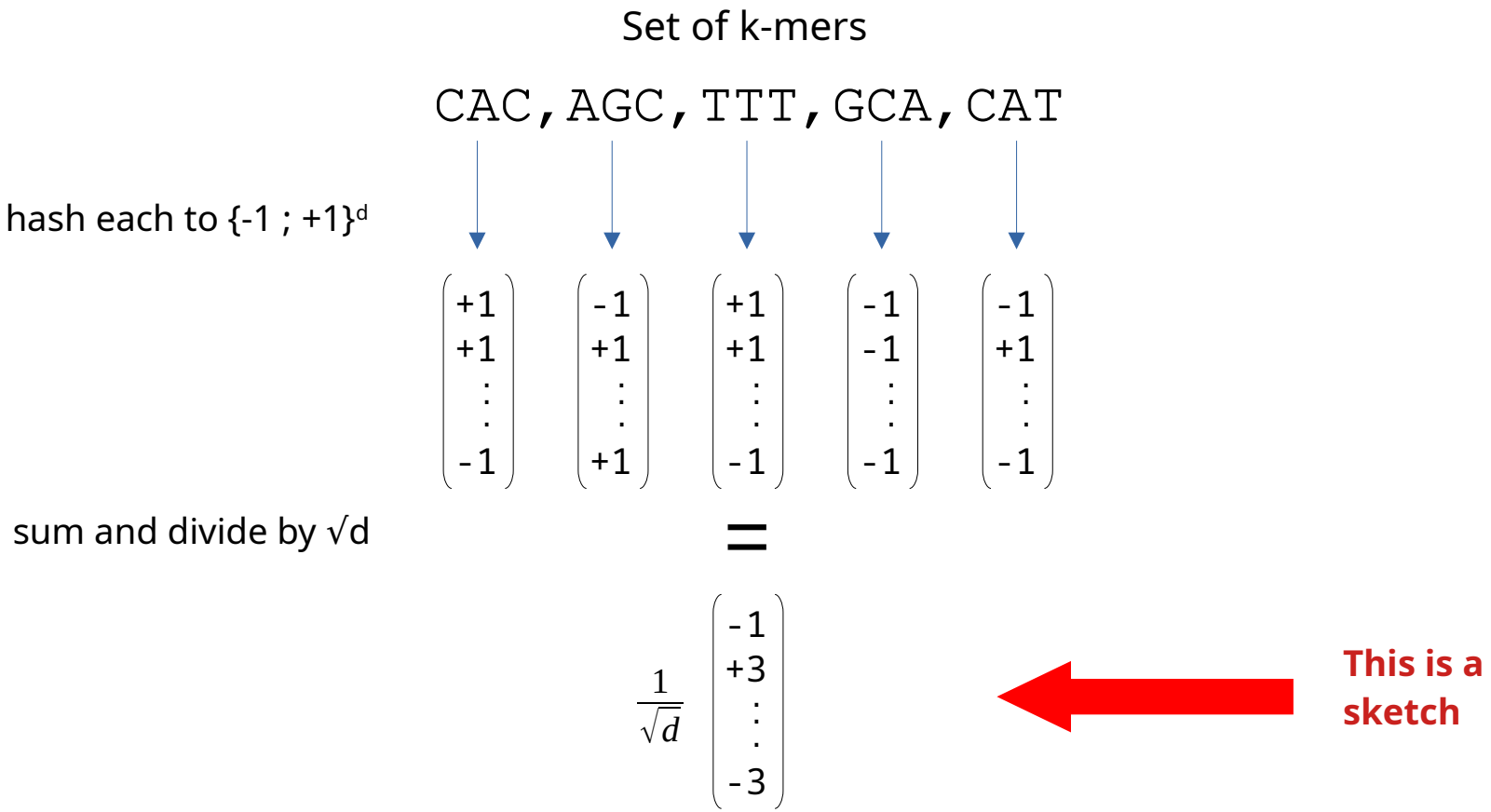
Vectorizing nucleotide sequences

- Genome language models (but not that good), not scalable in the foreseeable future
- Tetramer frequencies
- Random projection (DotHash)

Vectorizing nucleotide sequences: random projection



Vectorizing nucleotide sequences: random projection



k-mers vs vectors - Nucleotide

Arguments against vectors

- Does not really capture more meaning than kmers *yet* - unclear that it will improve quality of mapping
- k-mer datastructures scale better for now

Arguments in favor of vectors

- *Could* capture a more complex signal than k-mers, and not only with GLMs

Take-home messages

- Protein Language Models can be used to encompass complex signal in a simple datastructure

Take-home messages

- Protein Language Models can be used to encompass complex signal in a simple datastructure
- There exist industrial-grade solutions for neighbor indexing and nearest-neighbors query

Take-home messages

- Protein Language Models can be used to encompass complex signal in a simple datastructure
- There exist industrial-grade solutions for neighbor indexing and nearest-neighbors query
- Combining these gives great protein search engines

Take-home messages

- Protein Language Models can be used to encompass complex signal in a simple datastructure
- There exist industrial-grade solutions for neighbor indexing and nearest-neighbors query
- Combining these gives great protein search engines
- Unclear if this could benefit nucleotide mapping

Acknowledgments

- Rayan Chikhi
- Logan team: Téo Lemane, Pierre Peterlongo, Artem Babaian & others
- Galaxy team: Anton Nekrutenko, Björn Grüning, Danielle Callan, Saim Momin, Patrik Smeds, Nate Coraor
- TACC team: Kelsey Beavers
- Wei Shen, Martin Steinegger



Rayan Chikhi

Galaxy PROJECT

TACC
TEXAS ADVANCED COMPUTING CENTER

aws