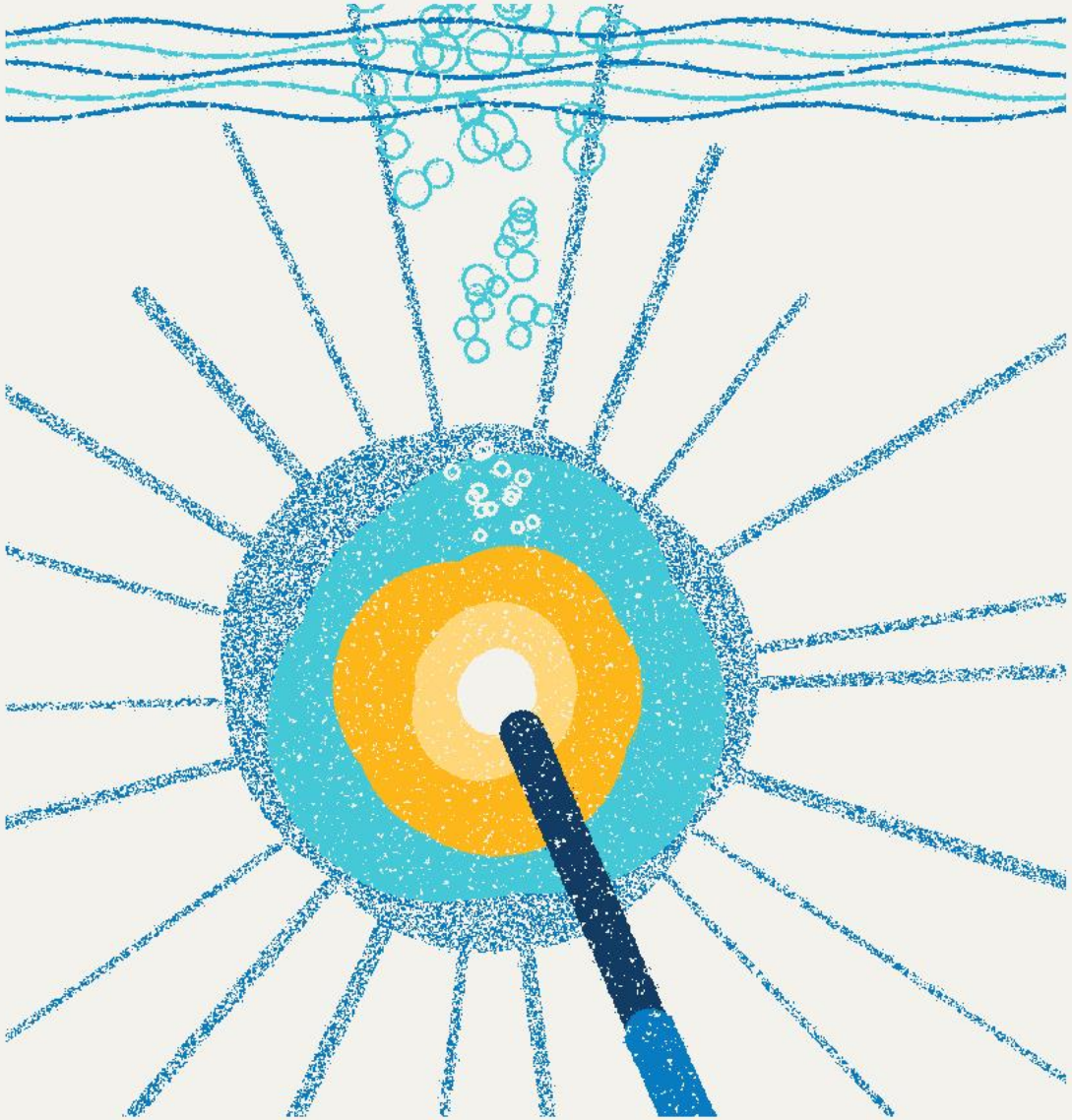
Protein mining

From alignment to embeddings

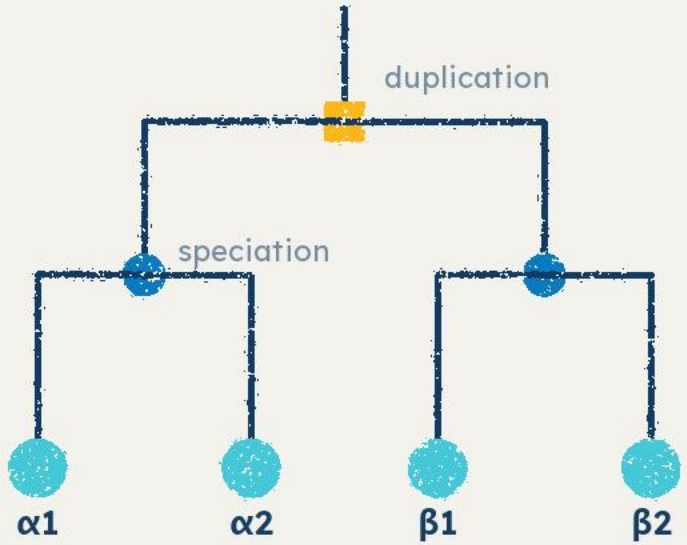
A field guide

Mannat Vikramaditya Jain



MINING

Homology is shared ancestry, and a significant score is the evidence.



α1, α2: orthologs (split at speciation)

α1, β1: paralogs (split at duplication)

> 2.5 billion yr

protein alignments see this far back; DNA only 200–400 million

1 in 3

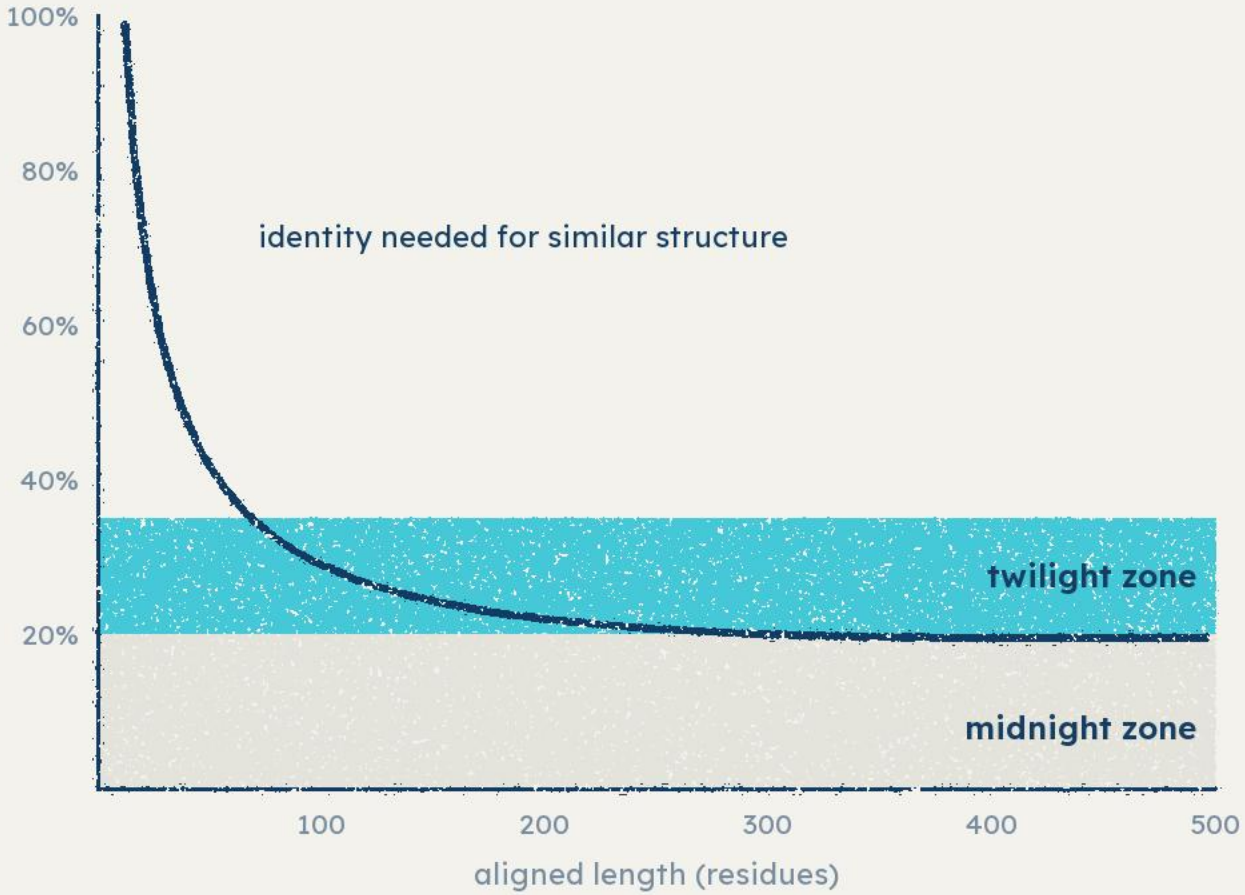
yeast–human homologs missed by a 30%-identity rule

one way

a significant score proves homology; a weak one does not disprove it

Pearson 2013; Rost 1999

Below 30% identity, sequence signal fades and structure persists.



20–35%

twilight zone: identity alone stops telling homologs from chance

8–10%

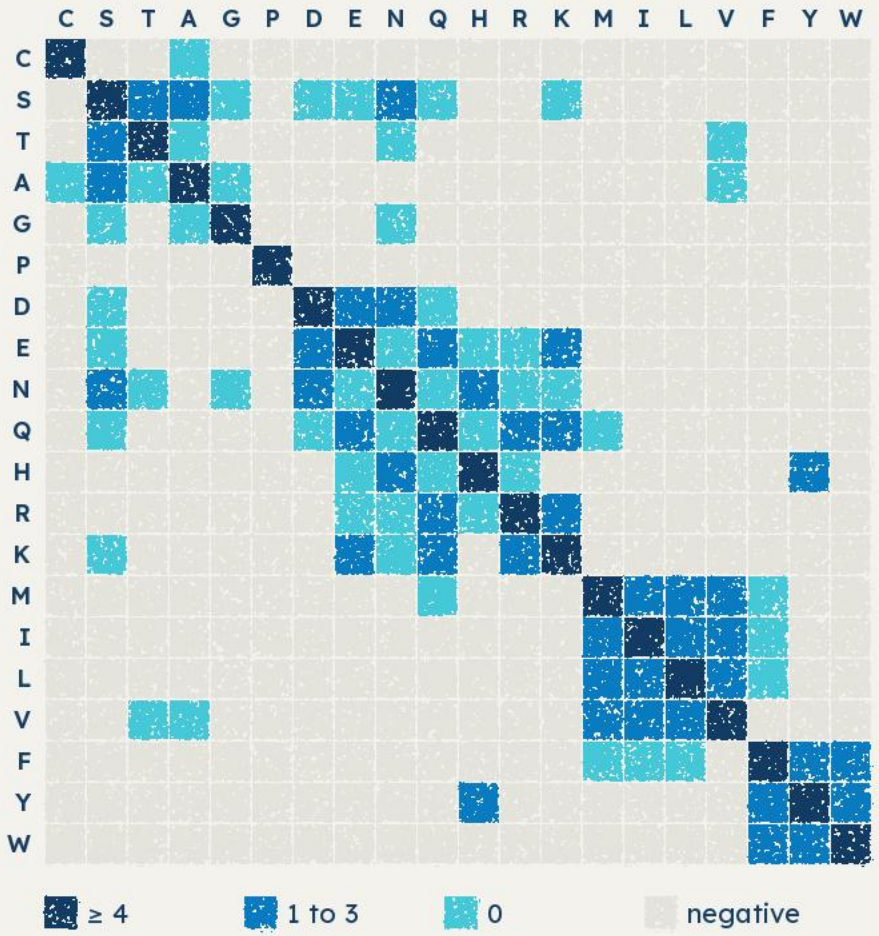
midnight zone: average identity of pairs with similar structures

≈ 1.8 Å

core deviation of homologs at 20% identity: the fold barely moves

Rost 1999; Chothia & Lesk 1986

BLOSUM62 scores each pair by how much more often homologs share it.



s(a, b) = log p(a, b) / f(a) f(b)

How often the pair appears in trusted alignments of homologs, against how often chance would pair them. Built from blocks clustered at 62% identity.

W·W +11, L·L +4

rare residues aligned are more surprising, so they score higher

K·E +1, A·L −1

a statistic about homologs, not chemical similarity

Henikoff & Henikoff 1992; Eddy 2004

Needleman–Wunsch aligns two sequences end to end.

		P	A	W	H	E	A	E
	0	-8	-16	-24	-32	-40	-48	-56
H	**-8**	-2	-10	-18	-16	-24	-32	-40
E	**-16**	-9	-3	-11	-18	-11	-19	-27
A	-24	**-17**	-5	-6	-13	-19	-7	-15
G	-32	**-25**	-13	-7	-8	-15	-15	-9
A	-40	-33	**-21**	-15	-9	-9	-11	-16
W	-48	-41	-29	**-10**	-17	-12	-12	-14
G	-56	-49	-37	-18	**-12**	-19	-12	-14
H	-64	-57	-45	-26	-10	**-12**	-20	-12
E	-72	-65	-53	-34	-18	-5	**-13**	-15
E	-80	-73	-61	-42	-26	-13	-6	**-8**

Global alignment lines up two sequences over their whole length. Every residue is either paired or set against a gap, and every gap costs, including those at the ends.

Each cell holds the best score of aligning the two prefixes. It takes the best of three moves: pair the residues, or open a gap in either one.

Right when two proteins are similar end to end.

score -8: BLOSUM62, gap 8 per residue

Needleman & Wunsch 1970

Smith–Waterman finds the best shared segment, and ignores the rest.

		P	A	W	H	E	A	E
	0	0	0	0	0	0	0	0
H	0	0	0	0	8	0	0	0
E	0	0	0	0	0	13	5	5
A	0	0	4	0	0	5	17	9
G	0	0	0	2	0	0	9	15
A	0	0	4	0	0	0	4	8
W	0	0	0	15	7	0	0	1
G	0	0	0	7	13	5	0	0
H	0	0	0	0	15	13	5	0
E	0	0	0	0	7	20	12	10
E	0	0	0	0	0	12	19	17

Local alignment finds the pair of segments that score best, and leaves the rest unaligned.

The only change is a floor at zero. When a prefix scores badly, the cell resets to 0 and the alignment starts afresh. The answer is the highest cell anywhere, traced back to a zero.

Proteins share domains, not whole lengths, so local alignment is the default for search.

finds AWGHE / AW-HE; the ends are ignored

Smith & Waterman 1981

BLAST skips most of the table and counts chance hits.

1 Seed

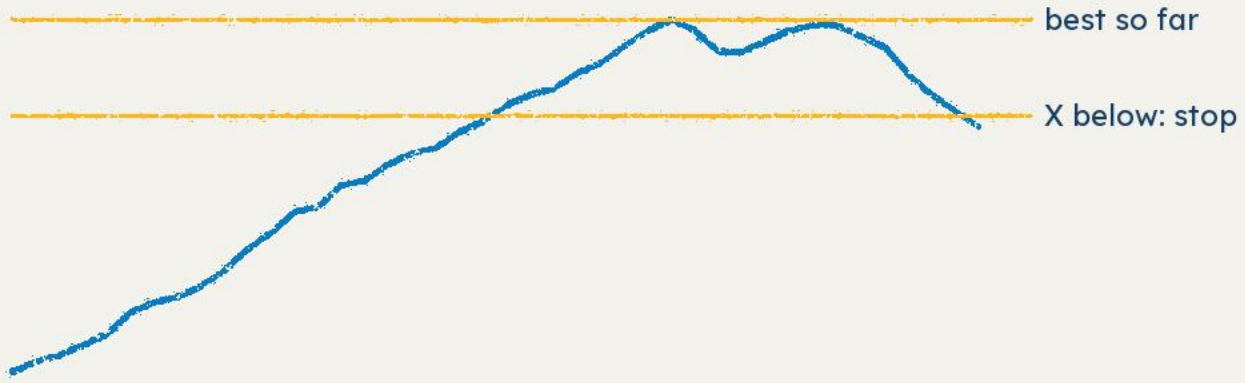
List every 3-letter word that scores at least T against a query word, and find exact copies in the database.

2 Extend

Grow each pair of hits on one diagonal without gaps, and stop once the score falls X below the best so far.

3 Gap

Run a gapped alignment only from strong ungapped hits, about one in fifty database sequences.



running score along one diagonal

$$E = K\,m\,n\,e^{-\lambda S}$$

expected number of chance hits this good; it grows with the database, bit scores do not

Altschul et al. 1990, 1997; Karlin & Altschul 1990

Profiles learn which positions matter.

bits of information per query position



search → add hits with E < 0.01 → rebuild profile → search again

16 rounds

for PSI-BLAST from BRCA1 to converge, finding 7 new BRCT proteins

2 of them

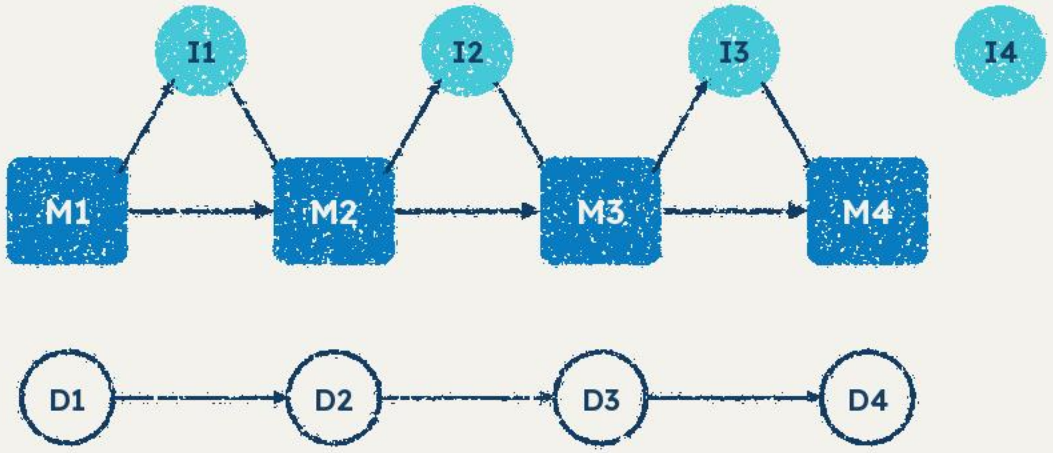
judged likely false positives by the authors themselves

"Once a sequence has been used … low E-values are virtually guaranteed in future iterations."

Wrong inclusions reinforce themselves.

Altschul et al. 1997 (PSI-BLAST)

Profile HMMs make gaps position-specific.



match emits a column's residues; insert adds residues; delete skips a column

0.3%

of hits lost by HMMER3's filter pipeline, at about BLAST speed

30,134

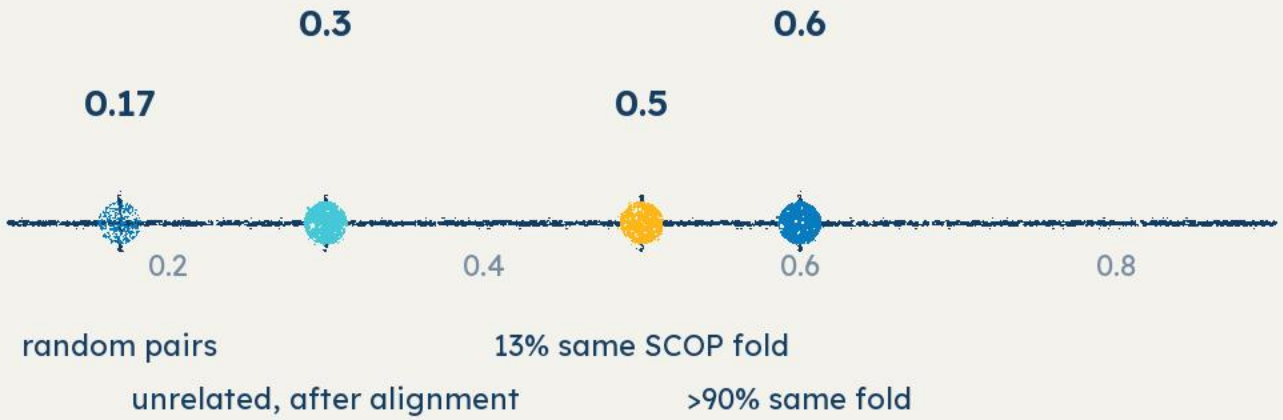
Pfam families (release 38.2), each a seed alignment, an HMM and a cutoff

77.1%

of reference-proteome sequences match at least one Pfam family

Eddy 1998, 2011; Pfam 38.2

Shape outlasts sequence, but TM 0.5 proves little.



TM-score: length-normalised overlap of two superposed structures

3–10×

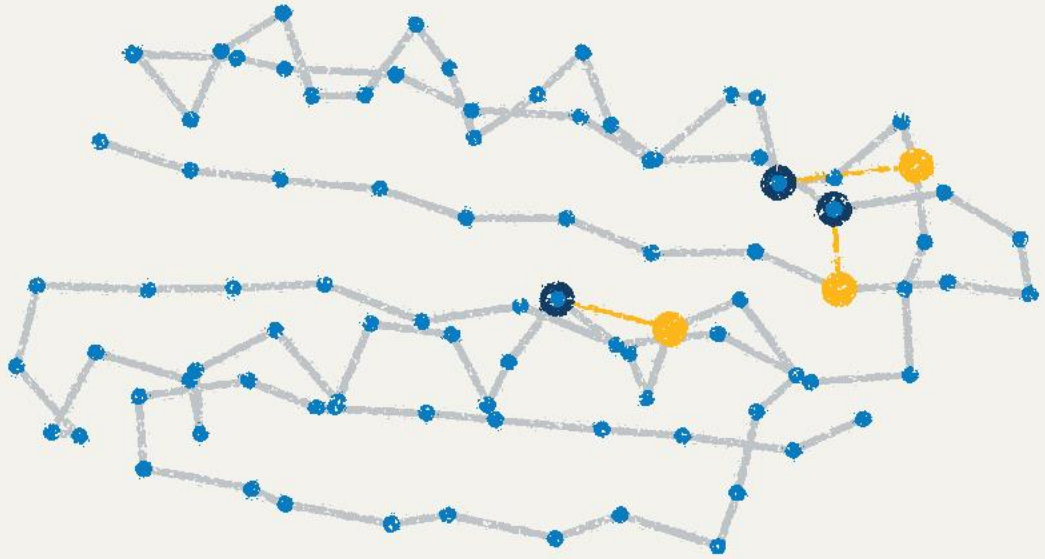
more conserved than sequence: shape decays far more slowly

≈ 1 month

for one TM-align query against 100 million structures, on one core

Zhang & Skolnick 2004, 2005; Xu & Zhang 2010

Foldseek cuts a structure into one letter per residue.



Top7 (PDB 1QYS): each residue (marigold) and its nearest partner in space (navy)

20 letters

each describes a residue's contact with its nearest partner in space

86% / 88%

of Dali's and TM-align's sensitivity on SCOPe40

6 s vs 10 days

one query against AlphaFoldDB v1, Foldseek vs Dali

first 35 residues: amino acid (cyan) and its 3Di letter (marigold), computed with mini3di

D	I	Q	V	Q	V	N	I	D	D	N	G	K	N	F	D	Y	T	Y	T	V	T	T	E	S	E	L	Q	K	V	L	N	E	L	M
d	k	w	k	w	k	w	a	d	a	p	p	d	i	d	i	d	i	d	d	d	d	d	v	v	c	v	v	v	v	l	v	v	v	l

van Kempen et al. 2024 (Foldseek); Kuhlman et al. 2003 (Top7)

Function is written in several vocabularies, and most labels are copied.

EC number

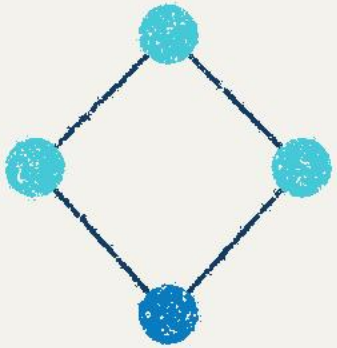
a four-level address for a reaction



6,980 active EC numbers

Gene Ontology

function, process and component, as a graph of terms



38,092 terms in a directed graph

Pfam

domain families, grouped into clans



one protein, several domains

30,134 families, 77.1% of sequences

CATH / Gene3D

structural domains: class, architecture, topology, homology

C class

A architecture

T topology

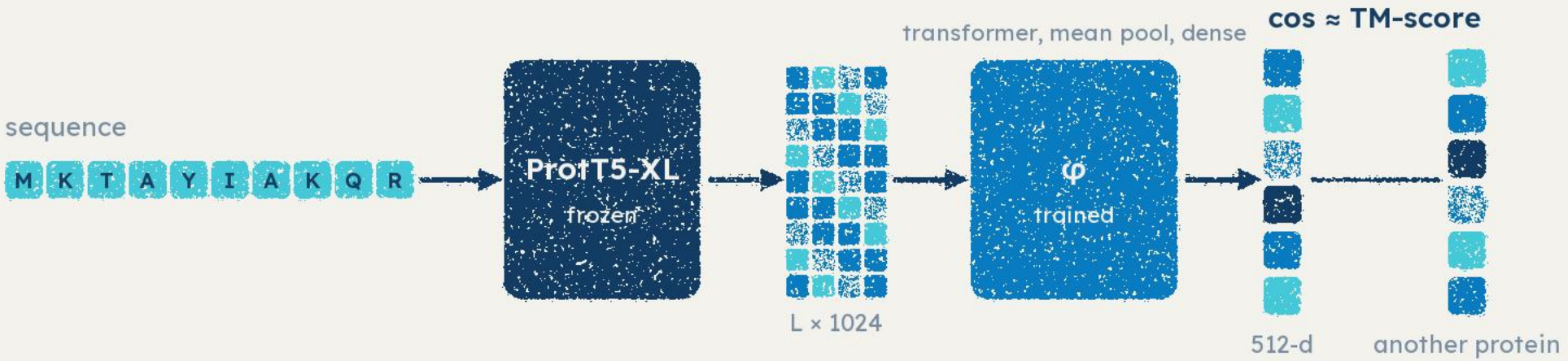
H homologous superfamily

a shared fold is not a shared function

62.5% of GO annotations are electronic (IEA); 7.2% rest on experiments.

GO Consortium Aug 2026; Expasy ENZYME Sep 2026; Pfam 38.2

TM-Vec turns structural search into nearest-neighbour lookup.



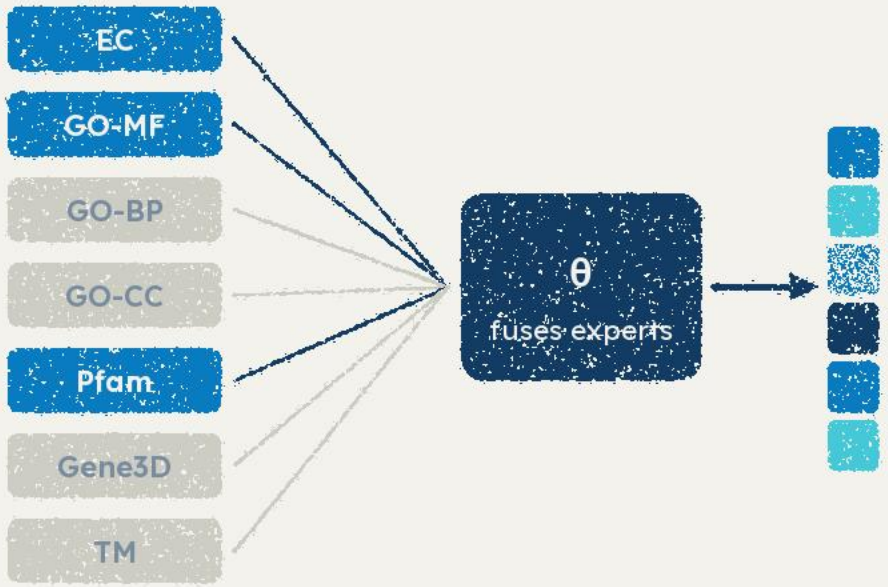
r = 0.97
predicted vs true TM-score, held-out pairs

r = 0.78
on folds never seen in training

≈ 45%
of calls false at a 0.5 cutoff on a novel-fold set (our arithmetic)

Hamamsy et al. 2024 (TM-Vec)

Protein-Vec lets a mask choose what “similar” means.



seven aspect experts; the mask (cobalt) picks which count

Seven experts, one vector.
Each expert is a TM-Vec-style network on frozen ProtT5, trained on one notion of similarity: EC, three GO branches, Pfam, Gene3D or TM-score.

A mask picks which experts count.
A fusion network θ combines the chosen experts into one 512-d vector, so the same two proteins can be close under one mask and far under another.

Trained on triples of proteins.
Each triple is a protein, one with the same label, and one with a different label. Training puts the first pair closer. The different one is picked to be nearly right (EC 2.3.1.199 vs 2.3.1.x), so the space learns fine distinctions. 450 million triples.

Search: encode under your mask, then copy the top hit’s label.

Hamamsy et al. 2023 (preprint)