Ομάδα υλοποίησης: Ιωάννης Νικολόπουλος – sdi2100292
Ιωάννης Πετράκης – sdi1900155

Protein Embedding ANN Search

Αναφορά Πειραματικής Σύγκρισης & Βιολογικής Αξιολόγησης

Ορισμός remote homologs:

Μια πρωτεΐνη-στόχος Q και ένα hit H θεωρούνται *remote homologs* όταν:

1. Η ομοιότητα ακολουθίας είναι χαμηλή (Twilight Zone)
    - π.χ. BLAST %identity < 30% (και συχνά μέτρια/χαμηλή κάλυψη ή/και μη-ισχυρό E-value), δηλαδή το BLAST δεν δίνει “καθαρή” ένδειξη κοντινής ομολογίας.
2. Στον embedding χώρο είναι κοντά
    - το H εμφανίζεται ψηλά στα Top-N μιας embedding-based ANN μεθόδου (μικρό L2, μικρό rank), άρα το μοντέλο “βλέπει” ομοιότητα που δεν φαίνεται εύκολα στην ακολουθία.
3. Υπάρχει ανεξάρτητη βιολογική/δομική ένδειξη ομολογίας από annotations
    - κοινή domain αρχιτεκτονική (ίδια/πολύ παρόμοια Pfam ή InterPro, ίδια διάταξη βασικών domains), και/ή
    - συμβατή λειτουργία (παρόμοια protein name/function, ίδια κατηγορία ενζύμου), και ιδανικά
    - συμφωνία σε GO terms (MF/BP/CC) ή/και EC αριθμούς (όταν υπάρχουν).

Operational rule:

Remote homolog candidate = (Top-N από embeddings) Λ (BLAST %id < 30%)
Remote homolog supported = candidate Λ (ισχυρή συμφωνία σε domains/λειτουργία από UniProt: κοινά Pfam/InterPro και συμβατή λειτουργική περιγραφή/GO/EC)

Τι δεν είναι remote homolog (πιθανό false positive):

- μικρό L2 αλλά χωρίς κοινά / συμβατά domains και με ασύμβατη λειτουργία στις UniProt σημειώσεις (π.χ. μοιράζονται μόνο πολύ γενικά keywords όπως “ATP-binding”).

3.1 Ποσοτική σύγκριση (Recall@50 έναντι BLAST Top-50, QPS)

Για τα ίδια queries, συγκρίνονται οι μέθοδοι ANN ως προς: (α) Recall@50 έναντι των κορυφαίων hits του BLAST, (β) ταχύτητα αναζήτησης (Queries per Second, QPS) και (γ) μέσο χρόνο ανά query. Οι τιμές προκύπτουν από 112 runs.

Πρωτόκολλο αξιολόγησης

- Ground truth: για κάθε query ορίζουμε ως “relevant” τα BLAST Top-50 hits (μετά από φιλτράρισμα E-value <= 1e-3).
- Recall@50: για κάθε query υπολογίζουμε

  Recall@50 = |ANN_Top50 ∩ BLAST_Top50| / 50,

  και αναφέρουμε τον μέσο όρο (Avg Recall@50) σε όλα τα queries.

- QPS: Queries Per Second (όσο μεγαλύτερο τόσο ταχύτερη η μέθοδος). Ο χρόνος ανά query προκύπτει ως ms/query = 1000 / QPS.

Best run ανά μέθοδο (μέγιστο Avg Recall@50)

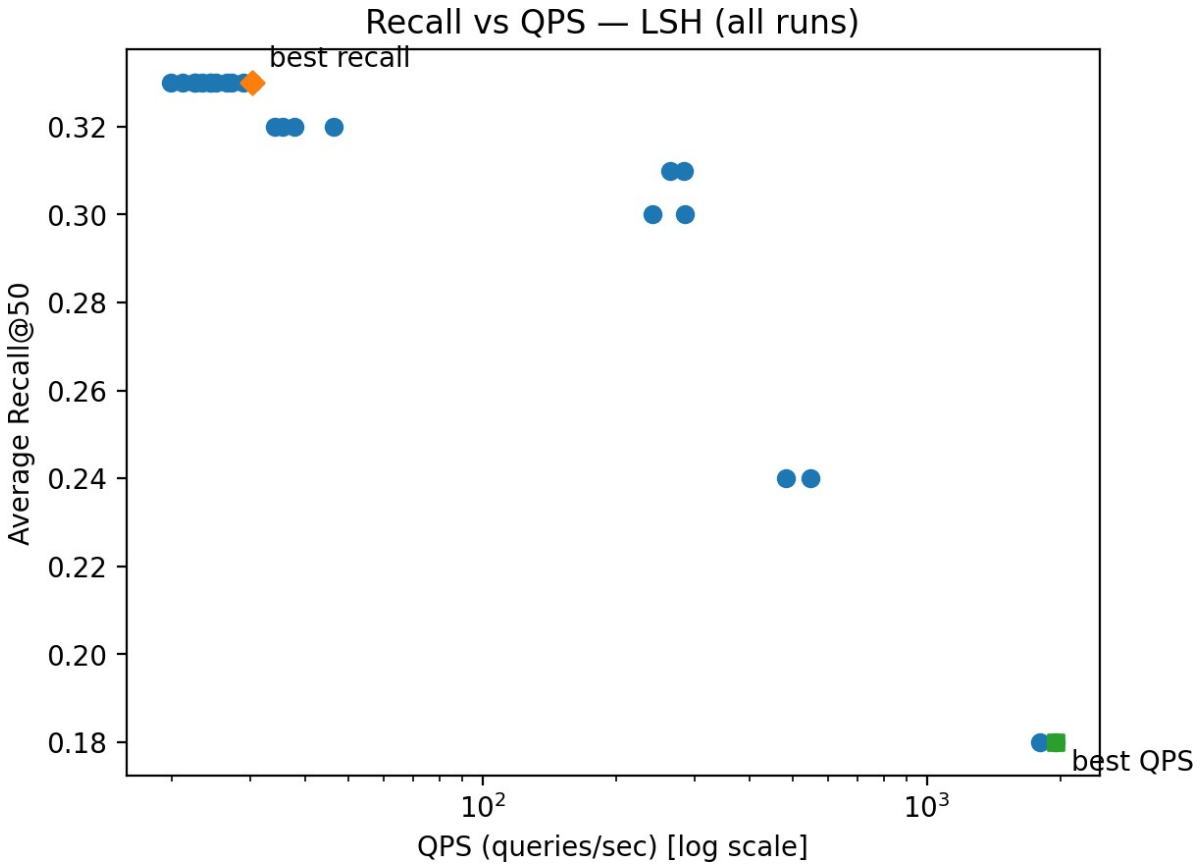
Μέθοδος	Avg Recall@50	Avg QPS	ms/query	Καλύτερες παράμετροι (σύνοψη)
neural	0.35	1648.2	0.607	neural_T=10, neural_m=1024, neural_epochs=5, train_size=50000
ivfflat	0.33	2145.8	0.466	train_size=50000, nlist=1024, nprobe=20
lsh	0.33	30.4	32.888	lsh_k=20, lsh_L=20, lsh_w=25.0
ivfpq	0.32	2839.0	0.352	train_size=50000, nlist=1024, nprobe=5, pq_m=8
hypercube	0.21	3290.6	0.304	cube_k=12, cube_M=2000, cube_probes=10, cube_w=4.0

Best run ανά μέθοδο (μέγιστο Avg QPS)

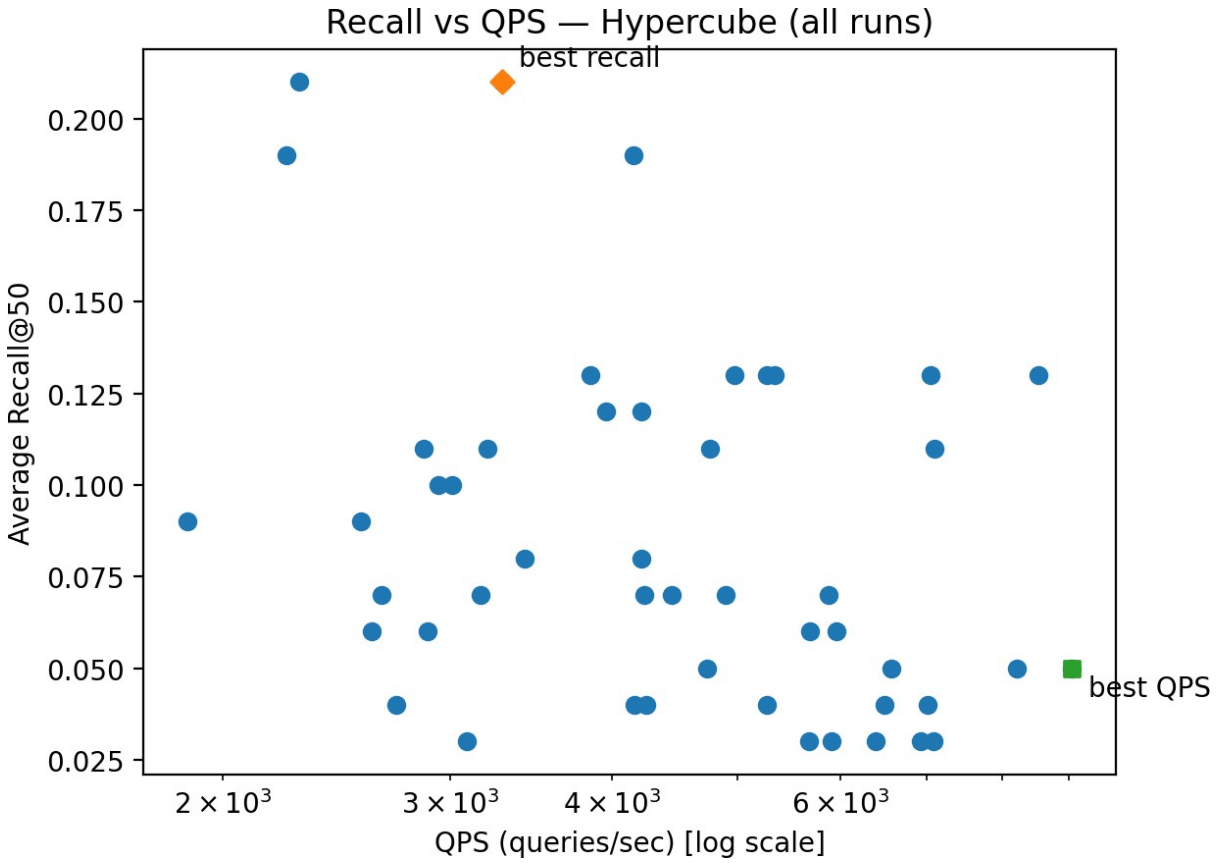
Μέθοδος	Avg QPS	Avg Recall@50	ms/query	Παράμετροι (σύνοψη)
hypercube	9058.0	0.05	0.110	cube_k=14, cube_M=1000, cube_probes=5, cube_w=4.0
ivfflat	4357.6	0.32	0.229	train_size=50000, nlist=1024, nprobe=5
ivfpq	3332.8	0.21	0.300	train_size=50000, nlist=1024, nprobe=5, pq_m=8
neural	2394.5	0.28	0.418	neural_T=5, neural_m=2048, neural_epochs=5, train_size=50000
lsh	1945.9	0.18	0.514	lsh_k=30, lsh_L=20, lsh_w=10.0

Σχέση ταχύτητας / ακρίβειας:

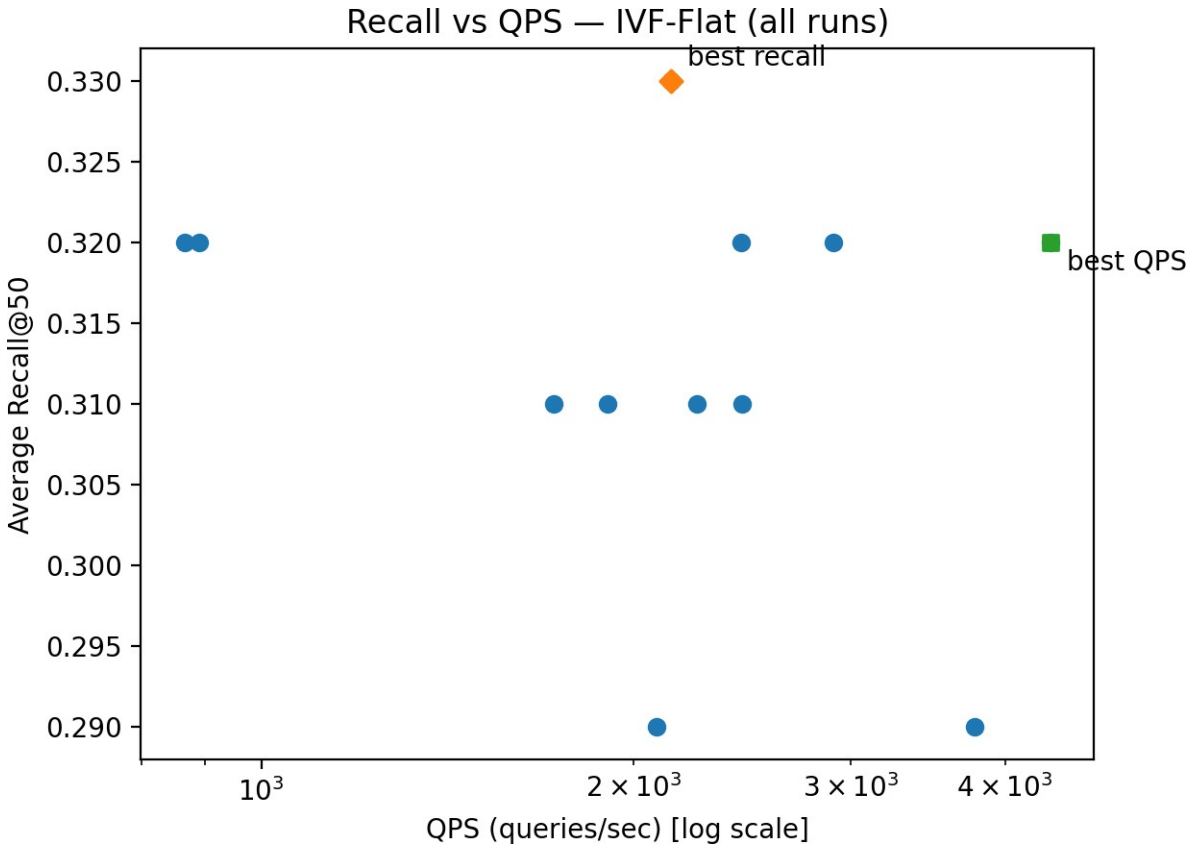
LSH:



Hypercube:

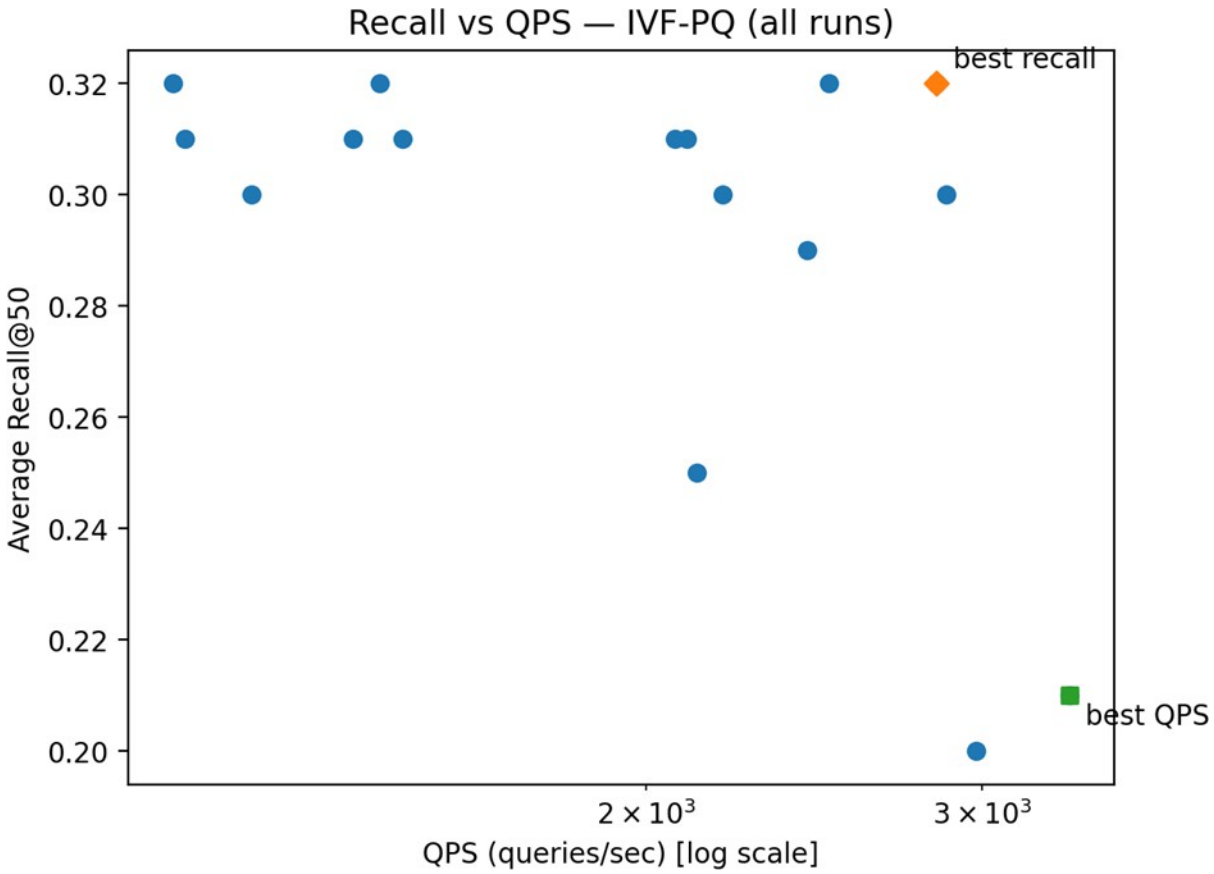


IVFFlat:



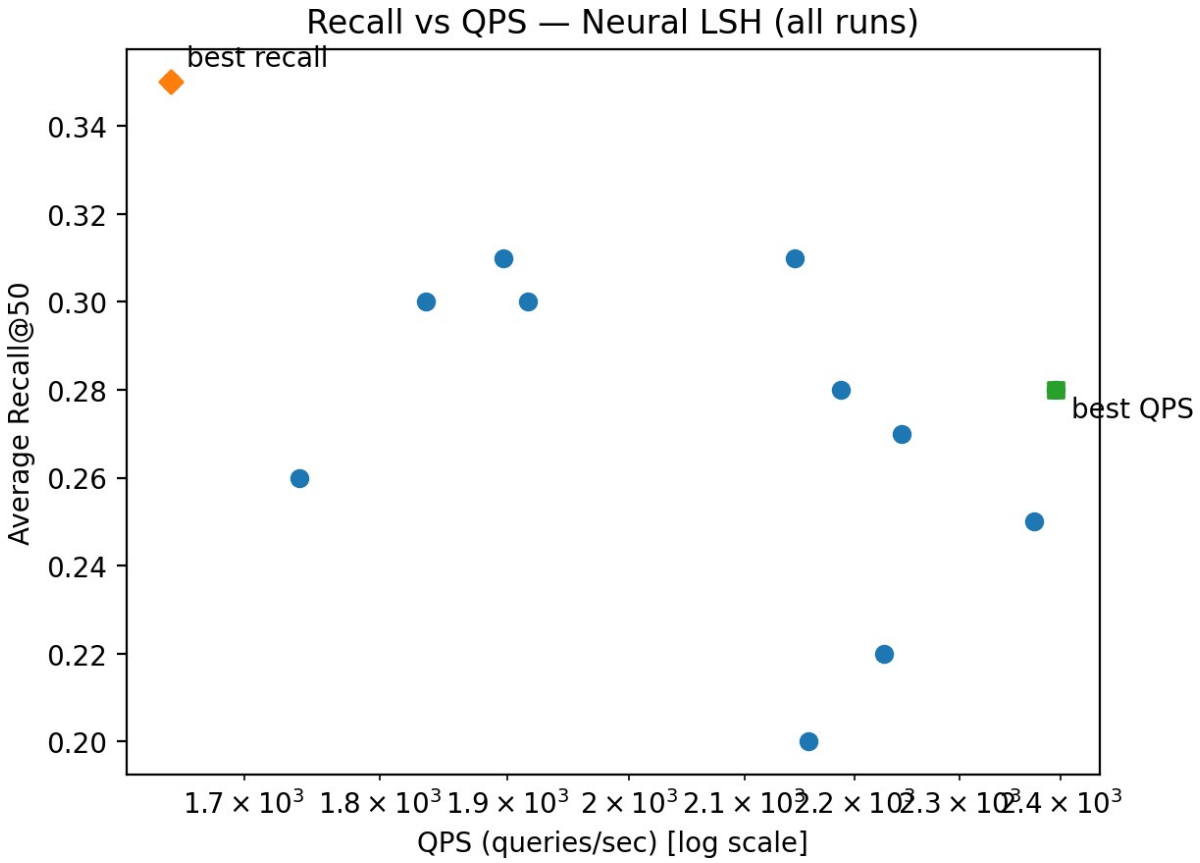
[4]

IVFPQ:

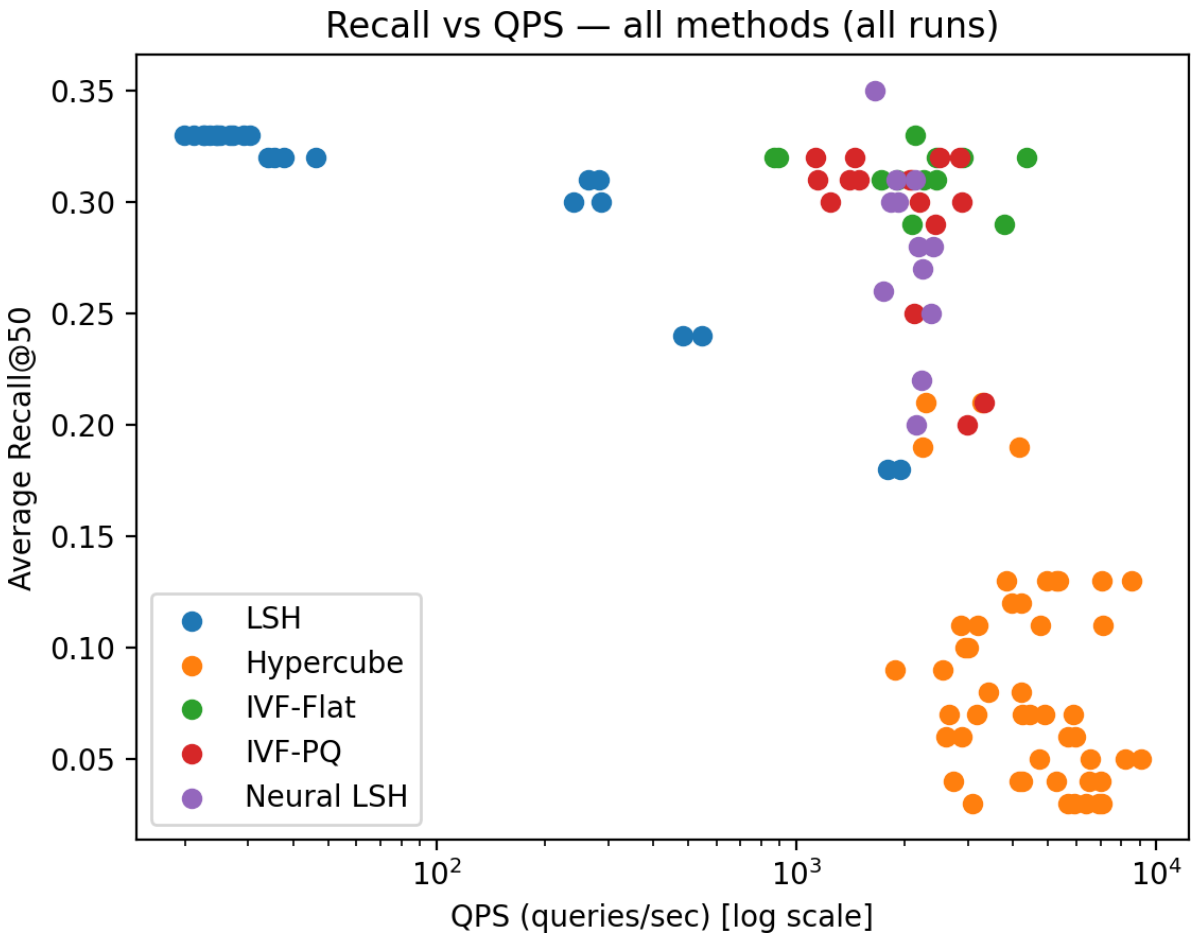

Recall vs QPS — IVF-PQ (all runs)
best recall
best QPS
Average Recall@50
0.32
0.30
0.28
0.26
0.24
0.22
0.20
2×10^3
3×10^3
QPS (queries/sec) [log scale]

[5]

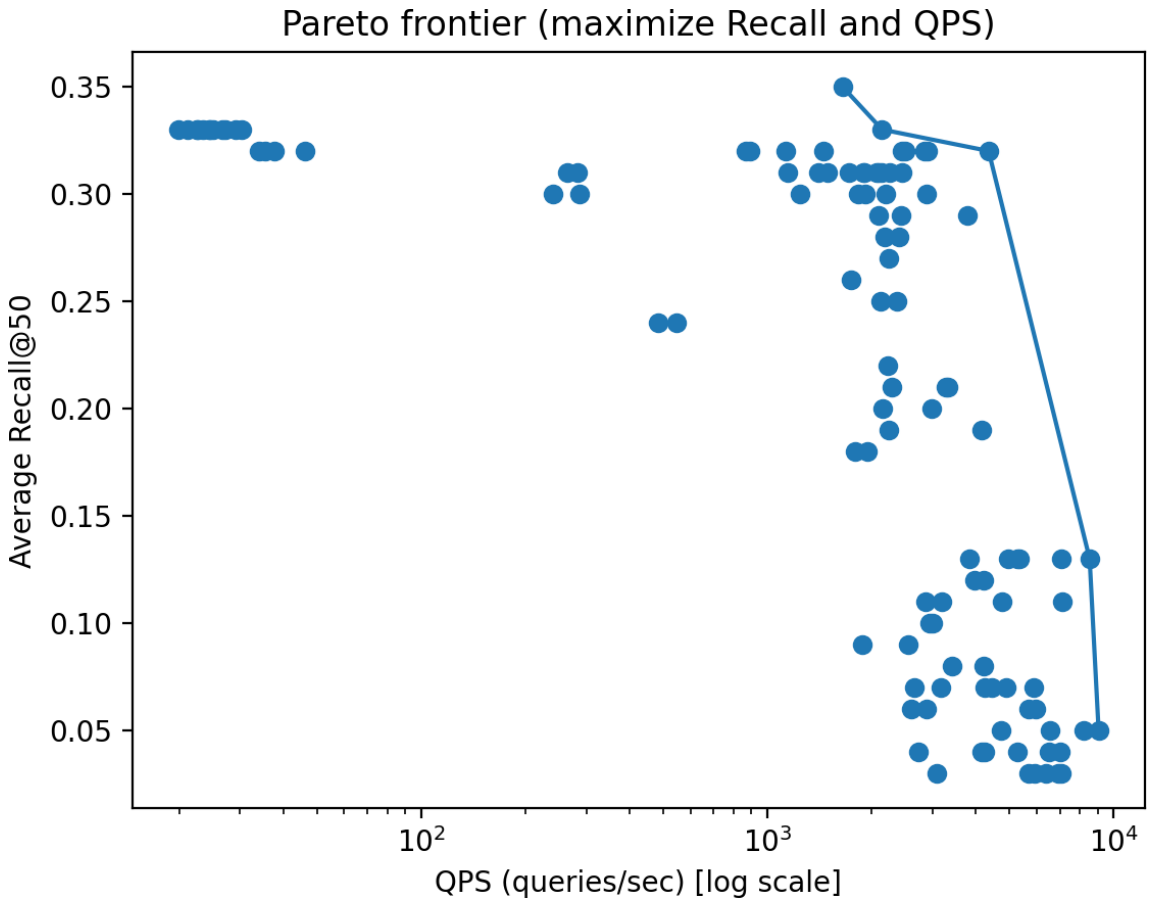
Neural LSH:


Recall vs QPS — Neural LSH (all runs)
best recall
best QPS
Average Recall@50
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
1.7×10^3
1.8×10^3
1.9×10^3
2×10^3
2.1×10^3
2.2×10^3
2.3×10^3
2.4×10^3
QPS (queries/sec) [log scale]

All runs:


Recall vs QPS — all methods (all runs)
Average Recall@50
0.35
0.30
0.25
0.20
0.15
0.10
0.05
LSH
Hypercube
IVF-Flat
IVF-PQ
Neural LSH
10^2
10^3
10^4
QPS (queries/sec) [log scale]

Pareto frontier:



Ενδεικτικά σημεία Pareto:

Μέθοδος	Avg Recall@50	Avg QPS	ms / query	Παράμετροι
neural	0.35	1648.2	0.607	neural_m=1024, neural_T=10, neural_epochs=5, train_size=50000
ivfflat	0.33	2145.8	0.466	nlist=1024, nprobe=20, train_size=50000
ivfflat	0.32	4357.6	0.229	nlist=1024, nprobe=5, train_size=50000
hypercube	0.13	8532.7	0.117	cube_k=12, cube_M=1000, cube_w=4.0, cube_probes=5
hypercube	0.05	9058.0	0.110	cube_k=14, cube_M=1000, cube_w=4.0, cube_probes=5

3.2 Βιολογική αξιολόγηση (remote homologs)

Εξετάζονται περιπτώσεις όπου το BLAST δίνει χαμηλό identity (< 30%) (Twilight Zone), αλλά οι embedding-based μέθοδοι επιστρέφουν κοντινούς γείτονες (μικρό L2, υψηλό rank). Για κάθε ζεύγος γίνεται έλεγχος των UniProt / SwissProt annotations (λειτουργία, GO, Pfam / InterPro, keywords) ώστε να εκτιμηθεί αν πρόκειται για πιθανή ομολογία (remote homolog) ή για πιθανό false positive.

Σύνοψη επιλεγμένων παραδειγμάτων (5 ζεύγη)

Query	Hit	Μέθοδος	Rank	L2	BLAST %id	Pfam overlap	Εκτίμηση
A0A009HL96	Q5HI09	Neural LSH	#1	0.6392	26.13 %	PF00072;PF00486	Remote homolog (υποστηρίζεται)
A0A009HN45	A7MSB2	Neural LSH	#1	0.6813	28.63 %	PF00176;PF00271	Remote homolog (υποστηρίζεται)
A0A009IB02	Q9SQI8	Neural LSH	#3	2.0060	28.80 %	PF00198;PF00364;PF02817	Remote homolog (υποστηρίζεται)
A0A010Q3W2	Q9S7I8	Neural LSH	#4	1.2982	22.01 %	—	Πιθανό false positive
A0A009HPM0	Q07YK6	Neural LSH	#8	1.4613	26.89 %	—	Πιθανό false positive

Παράδειγμα 1: A0A009HL96 -> Q5HI09

Κριτήρια επιλογής:

- Embedding hit: Neural LSH, rank #1, L2 = 0.6392
- BLAST %identity = 26.13% (< 30%), in BLAST Top-50: Yes
- Evidence overlap: Pfam overlap = PF00072;PF00486; InterPro overlap count = 6

UniProt annotations (Query: A0A009HL96)

- Protein: Response regulator
- Organism: Acinetobacter baumannii (strain 1295743) (TaxID 1310613)
- EC: nan
- GO (MF): GO:0000156 phosphorelay response regulator activity; GO:0000976 transcription cis-regulatory region binding
- GO (BP): GO:0006355 regulation of DNA-templated transcription
- GO (CC): GO:0005829 cytosol; GO:0032993 protein-DNA complex
- Pfam: PF00072;PF00486
- InterPro: IPR001789; IPR001867; IPR011006; IPR016032; IPR036388; IPR039420
- Keywords: DNA-binding; Phosphoprotein; Transcription; Transcription regulation; Two-component regulatory system

UniProt annotations (Hit: Q5HI09)

- Protein: Response regulator protein GraR
- Organism: Staphylococcus aureus (strain COL) (TaxID 93062)
- EC: nan
- GO (MF): GO:0000156 phosphorelay response regulator activity; GO:0000976 transcription cis-regulatory region binding
- GO (BP): GO:0006355 regulation of DNA-templated transcription; GO:0046677 response to antibiotic
- GO (CC): GO:0005829 cytosol; GO:0032993 protein-DNA complex
- Pfam: PF00072;PF00486
- InterPro: IPR001789; IPR001867; IPR011006; IPR016032; IPR036388; IPR039420
- Keywords: Activator; Antibiotic resistance; Cytoplasm; DNA-binding; Phosphoprotein; Repressor; Transcription; Transcription regulation (+2 more)

Συμπέρασμα: Remote homolog (υποστηρίζεται) - Τα κοινά domains / InterPro και η παρόμοια λειτουργική περιγραφή (GO) υποστηρίζουν πιθανή ομολογία παρά το χαμηλό identity.

Παράδειγμα 2: A0A009HN45 -> A7MSB2

Κριτήρια επιλογής:

- Embedding hit: Neural LSH, rank #1, L2 = 0.6813
- BLAST % identity = 28.63% (< 30%), in BLAST Top-50: Yes
- Evidence overlap: Pfam overlap = PF00176;PF00271; InterPro overlap count = 6

UniProt annotations (Query: A0A009HN45)

- Protein: DEAD/DEAH box helicase family protein
- Organism: Acinetobacter baumannii (strain 1295743) (TaxID 1310613)
- EC: nan
- GO (MF): GO:0004386 helicase activity; GO:0005524 ATP binding; GO:0016787 hydrolase activity
- GO (BP): —
- GO (CC): —
- Pfam: PF00176;PF00271
- InterPro: IPR000330; IPR001650; IPR014001; IPR027417; IPR038718; IPR049730 (+1 more)
- Keywords: ATP-binding; Coiled coil; Helicase; Hydrolase; Nucleotide-binding

UniProt annotations (Hit: A7MSB2)

- Protein: RNA polymerase-associated protein RapA
- Organism: Vibrio campbellii (strain ATCC BAA-1116) (TaxID 2902295)
- EC: 3.6.4.-
- GO (MF): GO:0003677 DNA binding; GO:0004386 helicase activity; GO:0005524 ATP binding; GO:0016817 hydrolase activity, acting on acid anhydrides
- GO (BP): GO:0006355 regulation of DNA-templated transcription
- GO (CC): —
- Pfam: PF00176;PF00271;PF12137;PF18337;PF18339
- InterPro: IPR000330; IPR001650; IPR014001; IPR022737; IPR023949; IPR027417 (+4 more)

- Keywords: ATP-binding; Activator; DNA-binding; Helicase; Hydrolase; Nucleotide-binding; Transcription; Transcription regulation

Συμπέρασμα: Remote homolog (υποστηρίζεται) - Τα κοινά domains / InterPro και η παρόμοια λειτουργική περιγραφή (GO) υποστηρίζουν πιθανή ομολογία παρά το χαμηλό identity.

Παράδειγμα 3: A0A009IB02 -> Q9SQI8

Κριτήρια επιλογής:

- Embedding hit: Neural LSH, rank #3, L2 = 2.0060
- BLAST %identity = 28.80% (< 30%), in BLAST Top-50: Yes
- Evidence overlap: Pfam overlap = PF00198;PF00364;PF02817; InterPro overlap count = 7

UniProt annotations (Query: A0A009IB02)

- Protein: Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex
- Organism: Acinetobacter baumannii (strain 1295743) (TaxID 1310613)
- EC: 2.3.1.-
- GO (MF): GO:0004742 dihydrolipoyllysine-residue acetyltransferase activity; GO:0031405 lipoic acid binding
- GO (BP): GO:0006086 pyruvate decarboxylation to acetyl-CoA
- GO (CC): GO:0005737 cytoplasm
- Pfam: PF00198;PF00364;PF02817
- InterPro: IPR000089; IPR001078; IPR003016; IPR004167; IPR011053; IPR023213 (+2 more)
- Keywords: Acyltransferase; Lipoyl; Transferase

UniProt annotations (Hit: Q9SQI8)

- Protein: Dihydrolipoyllysine-residue acetyltransferase component 4 of pyruvate dehydrogenase complex, chloroplastic
- Organism: Arabidopsis thaliana (TaxID 3702)
- EC: 2.3.1.12
- GO (MF): GO:0004742 dihydrolipoyllysine-residue acetyltransferase activity
- GO (BP): GO:0006086 pyruvate decarboxylation to acetyl-CoA; GO:0006096 glycolytic process
- GO (CC): GO:0005829 cytosol; GO:0009507 chloroplast; GO:0009534 chloroplast thylakoid; GO:0009570 chloroplast stroma (+3 more)
- Pfam: PF00198;PF00364;PF02817
- InterPro: IPR000089; IPR001078; IPR003016; IPR004167; IPR011053; IPR023213 (+2 more)
- Keywords: Acyltransferase; Chloroplast; Glycolysis; Lipoyl; Plastid; Reference proteome; Transferase; Transit peptide

Συμπέρασμα: Remote homolog (υποστηρίζεται) - Τα κοινά domains / InterPro, συγγενή ενζυμική λειτουργία (EC Number) και η παρόμοια λειτουργική περιγραφή (GO) υποστηρίζουν πιθανή ομολογία παρά το χαμηλό identity.

Παράδειγμα 4: A0A010Q3W2 -> Q9S7I8

Κριτήρια επιλογής:

- Embedding hit: Neural LSH, rank #4, L2 = 1.2982
- BLAST % identity = 22.01% (< 30%), in BLAST Top-50: No
- Evidence overlap: Pfam overlap = —; InterPro overlap count = 4

UniProt annotations (Query: A0A010Q3W2)

- Protein: WD domain-containing protein
- Organism: Colletotrichum fioriniae PJ7 (TaxID 1445577)
- EC: nan
- GO (MF): —
- GO (BP): GO:0006364 rRNA processing; GO:0045943 positive regulation of transcription by RNA polymerase I
- GO (CC): GO:0005730 nucleolus
- Pfam: PF00400;PF09384
- InterPro: IPR001680; IPR015943; IPR018983; IPR019775; IPR020472; IPR036322
- Keywords: Nucleus; Reference proteome; Repeat; WD repeat; rRNA processing

UniProt annotations (Hit: Q9S7I8)

- Protein: Cell division cycle 20.2, cofactor of APC complex
- Organism: Arabidopsis thaliana (TaxID 3702)
- EC: nan
- GO (MF): GO:0010997 anaphase-promoting complex binding; GO:0019900 kinase binding; GO:0097027 ubiquitin-protein transferase activator activity
- GO (BP): GO:0016567 protein ubiquitination; GO:0051301 cell division
- GO (CC): GO:0005634 nucleus; GO:0033597 mitotic checkpoint complex
- Pfam: PF24807
- InterPro: IPR001680; IPR015943; IPR019775; IPR033010; IPR036322; IPR056150
- Keywords: Alternative splicing; Cell cycle; Cell division; Mitosis; Nucleus; Reference proteome; Repeat; Ubl conjugation pathway (+1 more)

Συμπέρασμα: Πιθανό false positive - Παρότι τα embeddings δίνουν μικρό L2, τα domains / λειτουργία δεν συγκλίνουν επαρκώς με αποτέλεσμα να είναι είτε πιθανό false positive ή να έχει πολύ έμμεση συσχέτιση.

Παράδειγμα 5: A0A009HPM0 -> Q07YK6

Κριτήρια επιλογής:

- Embedding hit: Neural LSH, rank #8, L2 = 1.4613
- BLAST % identity = 26.89 % (< 30%), in BLAST Top-50: Yes
- Evidence overlap: Pfam overlap = —; InterPro overlap count = 3

UniProt annotations (Query: A0A009HPM0)

- Protein: Biotin carboxylase

- Organism: Acinetobacter baumannii (strain 1295743) (TaxID 1310613)
- EC: nan
- GO (MF): GO:0005524 ATP binding; GO:0016874 ligase activity; GO:0046872 metal ion binding
- GO (BP): —
- GO (CC): —
- Pfam: PF00289;PF00364;PF02785;PF02786
- InterPro: IPR000089; IPR001882; IPR005479; IPR005481; IPR005482; IPR011053 (+5 more)
- Keywords: ATP-binding; Biotin; Ligase; Nucleotide-binding; Transit peptide

UniProt annotations (Hit: Q07YK6)

- Protein: Formate-dependent phosphoribosylglycinamide formyltransferase
- Organism: Shewanella frigidimarina (strain NCIMB 400) (TaxID 318167)
- EC: 6.3.1.21
- GO (MF): GO:0000287 magnesium ion binding; GO:0004644 phosphoribosylglycinamide formyltransferase activity; GO:0005524 ATP binding; GO:0043815 phosphoribosylglycinamide formyltransferase 2 activity
- GO (BP): GO:0006189 'de novo' IMP biosynthetic process
- GO (CC): GO:0005829 cytosol
- Pfam: PF02222;PF21244;PF22660
- InterPro: IPR003135; IPR005862; IPR011054; IPR011761; IPR013815; IPR016185 (+2 more)
- Keywords: ATP-binding; Ligase; Magnesium; Metal-binding; Nucleotide-binding; Purine biosynthesis; Reference proteome

Συμπέρασμα: Πιθανό false positive - Παρότι τα embeddings δίνουν μικρό L2, τα domains / λειτουργία δεν συγκλίνουν επαρκώς άρα πιθανό false positive ή πολύ έμμεση συσχέτιση.

Σχόλιο για false positives

Οι δύο τελευταίες περιπτώσεις (ταξινομημένες ως «Πιθανό false positive») έχουν μικρό L2 αλλά μηδενικό overlap σε Pfam domains και περιορισμένη συμφωνία σε InterPro. Σε τέτοιες περιπτώσεις, η εγγύτητα στον embedding χώρο μπορεί να οφείλεται σε γενικά / μη ειδικά χαρακτηριστικά (π.χ. κοινά motifs, παρόμοια μήκη ή κοινές γενικές ιδιότητες όπως ATP-binding), χωρίς να συνεπάγεται κοινή οικογένεια ή λειτουργία.

3.2.2 Χαρακτηριστικά παραδείγματα όπου τα embeddings εντοπίζουν υποψήφιες απομακρυσμένες ομόλογες που δεν εμφανίζονται ψηλά στα αποτελέσματα του BLAST

Ως “In BLAST Top-50: No” εννοούμε ότι το αντίστοιχο hit δεν ανήκει στη λίστα των Top-N_eval = 50 subject IDs που επέστρεψε το BLAST για το ίδιο query (δηλ. δεν βρίσκεται στο blast_set που χρησιμοποιείται για το recall / τη στήλη “In BLAST Top-N?”).

Παρακάτω παρουσιάζονται 3 περιπτώσεις όπου το % identity είναι στο Twilight Zone (< 30%), οι embedding-based μέθοδοι φέρνουν τα ζεύγη κοντά στο χώρο των διανυσμάτων (μικρό L2, υψηλό rank), αλλά δεν εμφανίζονται ψηλά στο BLAST (βρίσκονται εκτός του Top-50):

Query	Hit	Embedding hit (method/rank)	L2	BLAST %id	**In BLAST Top-50?**	Pfam overlap	Εκτίμηση
A0A002	Q82MV1	Neural LSH #8, IVF-PQ #8	1.7931	29.52%	No	PF00005	Remote homolog (υποστηρίζεται)
A0A002	P9WQJ4	Neural LSH #10	1.8925	26.82%	No	PF00005	Remote homolog (υποστηρίζεται)
A0A010Q3W2	O35828	Neural LSH #8	1.3315	24.36%	No	PF00400	Πιθανό false positive

Παράδειγμα 1 - A0A002 -> Q82MV1

- Embedding evidence: *Neural LSH* rank #8 (L2 = 1.7931), και επιπλέον εμφανίζεται και με *IVF-PQ* rank #8.
- BLAST: 29.52 % identity, δεν είναι στο BLAST Top-50.
- UniProt / λειτουργία (σύνοψη):
    - Query (A0A002 / MoeJ5): keywords δείχνουν ATP-binding, membrane, transmembrane helix και GO (MF) περιλαμβάνει ATP binding / ATP hydrolysis / ABC-type transporter activity.
    - Hit (Q82MV1): “Aliphatic sulfonates import ATP-binding protein SsuB” (ATP-binding component ABC transporter).
- Domains / δομική ένδειξη:
    - Pfam: κοινό PF00005 (ABC transporter ATP-binding domain).
    - InterPro overlap: κοινές καταχωρήσεις τύπου ABC ATPase (π.χ. IPR003439 / IPR003593 / IPR027417).
- Συμπέρασμα: Remote homolog (υποστηρίζεται) - η συμφωνία σε ABC ATP-binding domain + ATPase λειτουργία εξηγεί γιατί τα embeddings τα φέρνουν κοντά παρότι το BLAST δεν το αναδεικνύει ψηλά (χαμηλό identity / πιθανή μεγάλη εξελικτική απόσταση).

Παράδειγμα 2 - A0A002 -> P9WQJ4

- Embedding evidence: *Neural LSH* rank #10 (L2 = 1.8925).
- BLAST: 26.82 % identity, δεν είναι στο BLAST Top-50.
- UniProt / λειτουργία (σύνοψη):
    - Hit (P9WQJ4): “Oligopeptide transport ATP-binding protein OppD” (ATP-binding component ABC transporter), με GO (BP) που περιλαμβάνει peptide transport.
- Domains / δομική ένδειξη:
    - Pfam overlap: PF00005 (ABC ATP-binding domain).
    - InterPro overlap: κοινές καταχωρήσεις που αντιστοιχούν σε ABC-type ATPase.
- Συμπέρασμα: Remote homolog (υποστηρίζεται) και εδώ η ομοιότητα φαίνεται να είναι σε επίπεδο ABC ATPase module (κοινός δομικός / λειτουργικός “πυρήνας”), κάτι που τα embeddings συχνά “πιάνουν” καλύτερα από τη στοίχιση ακολουθίας στη Twilight Zone.

Παράδειγμα 3 - A0A010Q3W2 -> O35828

- Embedding evidence: *Neural LSH* rank #8 (L2 = 1.3315).
- BLAST: 24.36 % identity, δεν είναι στο BLAST Top-50.
- UniProt/λειτουργία (σύνοψη):
    - Hit (O35828): “Coronin-7”, GO / keywords σχετίζονται με actin binding, Golgi / vesicle transport, Golgi organization.
- Domains / δομική ένδειξη:
    - Pfam overlap: κοινό PF00400 (WD repeat).
    - InterPro overlap: επίσης WD-repeat υπογραφές.
- Συμπέρασμα: Πιθανό false positive, εδώ η εγγύτητα μπορεί να οφείλεται σε γενική ομοιότητα επαναληπτικών WD motifs (κοινή δομή), χωρίς να υπάρχει πειστική λειτουργική σύγκλιση (nucleolus/rRNA processing vs Golgi/cytoskeleton). Είναι καλό παράδειγμα όπου το embedding “πιάνει” κοινό fold, αλλά αυτό δεν αρκεί για να ισχυριστούμε ομολογία σε επίπεδο οικογένειας / λειτουργίας.

3.3 Περιορισμοί και κατευθύνσεις βελτίωσης

Παρότι το BLAST Top-50 χρησιμοποιείται ως ground truth για τον υπολογισμό του Recall@50, είναι γνωστό ότι σε απομακρυσμένη ομολογία (remote homology) το BLAST μπορεί να μην τοποθετεί τους πραγματικούς ομόλογους ψηλά ή να μην τους επιστρέφει καθόλου. Άρα, χαμηλό Recall@50 δεν συνεπάγεται απαραίτητα βιολογικά λάθος εύρημα και για αυτό το λόγο συμπληρώνουμε την ποσοτική αξιολόγηση με λειτουργική / δομική τεκμηρίωση (UniProt, Pfam, InterPro).

Πιθανές βελτιώσεις (επόμενα βήματα):

- Εναλλακτικό / ισχυρότερο ground truth: χρήση PSI-BLAST ή profile-based μεθόδων (π.χ. HMMER/HH-suite) για πιο αξιόπιστη ανίχνευση remote homologs.
- Reranking: επανακατάταξη των Top-K υποψηφίων από ANN με alignment (BLAST / Smith-Waterman) ή με score που συνδυάζει embedding distance και sequence evidence.
- Φιλτράρισμα με βάση domain architecture: απαίτηση για τουλάχιστον ένα κοινό Pfam domain ή συμβατό InterPro signature για μείωση false positives.
- Calibrated thresholds: μελέτη κατωφλίου στην απόσταση (L2 / cosine) και / ή απαίτηση ελάχιστου coverage όπου είναι διαθέσιμο.
- Embeddings: σύγκριση διαφορετικών pooling στρατηγικών (mean / CLS), ή χρήση μεγαλύτερου ESM2 μοντέλου / fine-tuning σε protein families του ενδιαφέροντος.

Συνολικά, οι embedding-based μέθοδοι είναι ιδιαίτερα χρήσιμες για υποψήφια remote homologs, αλλά χρειάζονται βιολογικά φίλτρα / επαλήθευση ώστε να διαχωρίζονται οι πραγματικές δομικές / λειτουργικές συγγένειες από γενικές ομοιότητες (π.χ. κοινά repeats ή πολύ γενικά motifs).