

Low confidence is not absence: what alignment depth lets structure prediction say about partner proteins of phage reverse transcriptases

Author: Chulmin Pyo, AI3 Discovery. **Correspondence:** discovery@ai3.kr

Publication status: Self-published research manuscript. Not independently human peer reviewed.

Label. Results describe predicted structures of homologues found by sequence search in sequencing-run assemblies, and of partner proteins of repeat-array reverse-transcriptase loci whose sequences we hold: the two paralogous partners of Listeria phage LPJP1 (a locus listed in the original description as ART_11; P2 equals the partner drawn there by length and gap), a 91-chain family of their homologues, and four partners of other loci (two matched to the original figures by length and gap, two from separate accessions of the same lineage whose loci did not match the displayed figure fingerprints). The original structural analysis was of another locus (ART_57, accession unpublished). The work makes no novelty claim and says nothing about enzymatic activity.

Abstract

A low-confidence structure prediction is easily read as "no domain", yet for proteins from uncultured phages the public alignments that predictors rely on are often too shallow to support that reading. A recent preprint described Type I partners of array-associated reverse transcriptases as having two tandem GNAT-like (GCN5-related N-acetyltransferase) folds; in the one partner whose analysis it shows, the second region is a 91-residue C-terminal lobe with weaker similarity to tGCN5 (TM 0.31) (Yoon et al. [1]). We asked what structure prediction can say about such partners once the missing alignment depth is supplied from public data, and whether the answer holds beyond a single locus. The initial verdict rules were fixed in a dated local git record before the predictions; in the last stage the Pfam scan was added after the other readouts had been read. We predicted lone-chain structures with AlphaFold2-multimer v3 and Boltz-2 for the two paralogous partners of Listeria phage LPJP1, a 91-chain family of their homologues, and four partners of other repeat-array loci whose complete sequences we held. Alignments came from the ColabFold server (at most 18 rows) and from the public Logan protein clusters; the models were read with Foldseek, Pfam clan models and a name-free structure-panel test, alongside positive controls, negative controls and a row-shuffled alignment control. Adding homologues raised the AlphaFold2 mean pLDDT of P2 to 80–83, against 35 with shuffled rows, and an ART_69 partner of 594 residues was predicted with confidence over most of its length (88.4–88.8, against 46–49 with shuffled rows), whereas the N-termini of P1 and P2 stayed low-confidence. The three readouts disagreed. Foldseek found at most one GNAT-like region and never two (classed NONE for P2 and for 35 of 91 family chains with AlphaFold2 and 17 with Boltz-2); Pfam clan models hit all 8 GNAT control entries (7 distinct proteins) in the family stage and all 3 in the last stage, but found no clan hit in the targets. The registered two-region criterion of the structure-panel test was met by 56 of 91 family chains and by all 24 models of the 4 additional partners (the three-seed model set recorded after the results were read; shared alignments over 3 lineages, not independent observations). Separately, an exploratory second-region score found two adjacent windows in 55 of the 71 family chains in which a second window could be scored (55 of all 91 chains, 60%). The frozen verdict was UNKNOWN for P1 and P2; for the additional partners no verdict was assigned and all 8 cell labels were unknown. A low-confidence model built on a shallow alignment therefore cannot support "no second domain", while a diverged GNAT-

like fold and a different alpha/beta fold both remain possible. These are predictions for a handful of partners without an enzyme assay, and the locus that the preprint analysed was not available to us, so the results neither support nor contradict its claim.

1. Introduction

Structure prediction is routinely used to assign folds to proteins of unknown function, and the reliability of an assignment depends on the multiple sequence alignment (MSA) given to the predictor. For proteins from uncultured phages and metagenomic assemblies, the public MSA databases can contain few homologues. A low-confidence model may then reflect alignment depth rather than a property of the protein, and a negative ("no second domain") is informative only where the region is predicted with confidence.

The array-associated reverse transcriptases (ART) described by Yoon et al. [1] are coupled with a repeat array and a dedicated partner gene. The Type I partner is described as "about 600 residues with two tandem GNAT-like folds", based on AlphaFold2 co-folds with locus-paired family alignments built from the authors' own metagenomic data. For the one partner analysed with a TM-align operation (the ART_57 partner, about 464 residues; tGCN5 against the chain, then against the remainder), the first region (residues 212-373) matched with TM 0.58 and the remaining C-terminal 91 residues (374-464) with TM 0.31. The table in the preprint's Supp. Fig. 7C calls that segment a "C-terminal lobe" (the structure panel of Supp. Fig. 7B labels it "GNAT-like lobe") without a P-loop or strand beta5, but with 20 atoms within 2 Å of the superposed acetyl-CoA (32 for the first region), whereas its main text speaks of two tandem GNAT-like folds. The source analysed here is the [Anthropic-hosted manuscript](#) linked from the research announcement. The analysis refers to that source copy; we did not establish a peer-review status for it.

We started from the Type I pair of a public genome (*Listeria* phage LPJP1; paralogs P1 and P2). LPJP1 is listed in the preprint as ART_11, and P2 equals the partner drawn there for that locus (Section 3.9), but the preprint's GNAT analysis was done on ART_57, whose accession is unpublished and which we could not obtain. We asked three questions: (i) does a public-MSA pipeline reproduce the description, (ii) can missing alignment depth be supplied from public sequence data, and (iii) what does a family of recovered homologues look like when all members are predicted and searched with the same detector? A fourth question was added after these stages: (iv) do partners of other repeat-array loci, whose complete sequences we hold, behave in the same way? Our structure-search detector calls full-size GNAT-like regions (141-213 aa); a separate lobe analysis (60-140 aa) was also run. Both are described below together with what they cannot exclude. Figure 1 gives an overview of the stages and readouts.

We pre-specified the verdict rule before any prediction (in the local, externally unstamped sense defined in Methods), ran the controls first and recorded every deviation. This work is a worked example, not a general estimate, of how a prediction-based negative can be unmeasured rather than negative.

2. Methods

Pre-registration and review. The decision rule of the first stage (D3) was frozen in a dated, hashed document before any MSA-based prediction (the document itself is withheld; its hash and commit identifier are recorded in the public data supplement, see Data and code availability). Amendment A1 was registered before the controls were scored, and a documented reading, A1.7, after scoring had started; its sentence that all scoring code preceded the first control model was inaccurate and is recorded as such. Later stages (D4-D6) were registered in a second document of the same kind before their analyses. The last stage (D13, partners of other ART loci) was registered in a section of that second document and in amendment A1

before any Logan hit of its targets was read (the first-pass stream had already started, with its hit files unread, as A1 states). Amendment A2, which also withdrew the D3 verdict for these lone chains and ran Foldseek and the structure-panel test on all primary units instead of only the resolved ones, was written after the resolution analysis of the finished units had been read (its code was committed first) and before any Foldseek run. Amendment A3, written after the Foldseek and structure-panel results had been read, covers an exploratory Pfam scan and also records after the fact that the panel used all three seeds of each primary unit rather than the seed-0 model alone as in D8. The two-region criterion was registered; the three-seed model set and its S2 values are post hoc. The registration text of each stage was committed before its outputs were read, with the exceptions listed under Deviations (amendments D8-A2, A3 and D8-D were registered after earlier D8 output had been read, and the code of three D8 steps was committed together with its results); exploratory analyses are labelled as such. "Pre-registered" here means committed to a local, never-pushed git repository with self-generated hashes and no external timestamp. Three read-only reviews (of the D3, D4/D5 and D6 stages) and one of the earlier draft were carried out by separate AI sessions, not by humans; the D13 text, which they did not cover, was checked by three further read-only AI sessions (claims against data, English-Korean fidelity, Korean naturalness) and by scripted consistency checks (not shared); they re-derived verdicts from raw rows and recomputed values from raw model files, and their corrections were applied. All headline numbers are computed from the result files into `NUMBERS.json`.

Sequences. P1 (QXN67954.1, 661 aa), P2 (QXN67952.1, 481 aa) and the cognate reverse transcriptases RT1, RT2 are from the LPJP1 genome (GenBank MZ422438.1). Controls (sequences from UniProt, fixed in the pre-registration: Eis P9WFK7, RimI P0A944, MshD P9WJM7, Orf14 Q8KRB5, DnaK P0A6Y8, UPP P9WFF3, GlmU P0ACC7) were Eis (tandem GNAT; structure 3R1K [24]), RimI (single GNAT; no experimental structure), MshD (tandem GNAT-like; 1OZP, 1P0H, 2C27 [25]) and Orf14 (tandem GNAT; 2WPW [26]); DnaK (4B9Q [38]), UPP (5E38 [39]) and GlmU (1HV9 [30]) served as non-GNAT controls.

Partners of other ART loci (D13). Candidates were the first gene downstream of the reverse transcriptase in loci of our own earlier survey of public assemblies that either matched a locus of the preprint's Supplementary Fig. 3C by a figure fingerprint (RT length, partner length, RT-partner gap and repeat spacings read from the vector PDF; two loci, ART_69 and ART_34) or were separate accessions of one lineage (L1) with a repeat-array call whose loci did not match the displayed figure fingerprints. Absence from the preprint's full locus set was not established. The match to the figure is a fingerprint, not an accession: the RT and partner lengths equal those drawn (counting the stop codon) and the RT-partner gap differs by 3 nt; the original probe also compared repeat spacings, which are not re-derived here. Before registration each partner ORF was checked for completeness: its length had to equal three times the residue count plus the stop codon (a stop codon inside the contig), it had to start with methionine and contain no ambiguous residue; the check and its outcome are generated by code (its outcome is in `numbers/NUMBERS_D13.json`). The pipeline was that of the earlier stages with the following registered changes (D13-A1, A2): a family set per target (Logan first pass only; $E \leq 1e-5$, query coverage ≥ 0.5 , 90% identity de-duplication); lone chains only, seeds 0-2; Foldseek with the amended classifier and the unchanged gates; a descriptive structure-panel test on primary units using the registered two-region criterion (the all-three-seed scope was recorded post hoc in D13-A3); the frozen seed-call rule applied to each lone chain with the target treated as a partner without a known first GNAT. The D3 verdict table is defined on co-fold labels and was not assigned, because the reverse-transcriptase sequences of the loci were not part of the frozen targets. Server alignments were fetched on a laptop (the compute pod cannot reach the server), and the target sequences were thereby sent to the public ColabFold server. A Pfam clan scan (98 HMMs, gathering cut-off, Eis, RimI and Orf14 as positive controls) was added after the Foldseek and panel results had been read and is exploratory (D13-A3).

Structure prediction. We used AlphaFold2-multimer v3 [3,4] through ColabFold 1.6.1 [2] (5 models x 3 seeds, 3 recycles, no templates, no relaxation) and Boltz-2 v2.2.1 [5] (3 recycling steps, 5 diffusion samples x 3 seeds, no kernels) on NVIDIA B300 (Blackwell) GPUs (the Boltz-2 runs of D12: H100 80 GB); the same a3m was given to both engines. A CPU-versus-GPU check of one 911-residue, two-chain control co-fold gave sub-angstrom agreement per chain for the ordered parts; the relative chain orientation differed (exploratory). Chai-1 [7] (MSA-free, ESM embeddings) and Protenix 0.5.5 [8] (MSA-free; it resolved none of the controls and is not reported further) were run on the same chains as exploratory cross-checks. They test predictors without alignments, not predictors with equally deep MSAs.

Server and family alignments. Server alignments came from the ColabFold server (the only external sequence transfer). Each server alignment lists the query twice, so the record counts include both copies. P1 has 6 distinct aligned rows and 6 distinct sequences, including the query (5 homologues); P2 has 17 distinct aligned rows but 16 distinct sequences (15 homologues), because A0A233VSS0 occurs in two differently aligned rows. The Logan [17] 50% protein-cluster representatives (public bucket s3://logan-pub, an SRA re-analysis; the exact object is identified by ETag 357cde7dd7847a6726500702e2e7125c-8087 and size 271,342,209,328 B, reproduced by a later HEAD request; a release name was not recorded) were streamed by anonymous HTTPS range download (3,009,464,616 records). Only records of 300-1000 aa (110,770,355, 4%) were searched with MMseqs2 [10] (single pass, sensitivity 7.5, $E \leq 1e-5$, at most 2000 hits per query; then a second pass with a profile built from the first-pass hits; profile rows of two drifting members were not used), counting homologues with query coverage ≥ 0.5 and 90% identity de-duplication; this length filter excludes fragments and any partner above 1000 aa. The family F (45 sequences) contains hits from the human cluster file and from the profile pass; its expansion through the members of the cluster representatives (public unclustered protein file) gave F2 (89 sequences from 784 member records). Family rows were aligned to the query with MMseqs2 (insertions dropped, rows with $< 25\%$ of the query columns removed) and appended to the server rows; nested subsets give the dose-response. The census MSAs (each chain aligned against the other 91 minus one chains, $E \leq 1e-3$, no server rows) have 1-83 rows (median 77).

Resolution. A slot is a 141-residue window (step 10). In the descriptive rule used for Tier 1 and the census, a chain is resolved when at least half of its slots have mean pLDDT ≥ 70 in at least two of three seeds; for P1 in Tier 1 the slots exclude residues 484-658 (the region that carries the known GNAT-like peak). A partner was counted as "resolvable" when any primary unit met the rule in at least one engine. The frozen D3 rule additionally requires every slot outside the first GNAT to reach a predictor-specific floor and an intra-slot PAE ≤ 10 Å before a negative can be called. The AlphaFold2 floor of that rule (pLDDT 90) rests on sparse control data and is not a validated detection limit.

Detector. GNAT-like regions were detected with Foldseek [9] (213-aa windows, step 50, afd50, pdb100 and CATH50, $E \leq 1e-3$). We used the amended classifier of the pre-registration (A1.1). It accepts targets annotated as GNAT/acetyltransferase of the GNAT clan by name, CATH 3.40.630 or Pfam clan CL0257 accession, and it also accepts targets whose name contains a loose 'N-acetyl' string; this branch causes the GlimU false positive described below, while the calls of the SINGLE chains do not depend on it. The peak rule requires an alignment of 141-213 aa, target coverage ≥ 0.6 , at least 10 distinct targets and peak separation ≥ 100 aa; a lobe analysis used 60-140 aa, coverage ≥ 0.4 and the same target threshold. The detector was gated on controls before any target was read (two peaks in Eis and one in RimI at the control cells that passed; none in the non-GNAT controls). In the depth ladder of one control (Eis), both tested GNAT folds were detected at Neff80 17 and 124 in both predictors. At the two lowest rungs (each Neff80 1), AlphaFold2 detected neither fold; Boltz-2 detected some fold instances, including both folds in some seeds. No rung between Neff80 1 and 17 was tested. In the depth-matched controls (D7) the detector gave TANDEM in all 12 control-predictor-depth calls of Eis, MshD and Orf14 (12 of 12), NONE for DnaK, and

TANDEM also for the non-GNAT acetyltransferase GlmU (a false positive: the classifier accepts targets by their annotation names). The lobe-size limit of the controls is 120 aa, so the detector is not shown able to call a 91-aa lobe.

Operational definitions (for re-computation). *Foldseek*. Release 10-941cd33, `easy-search -e 10 --max-seqs 1000 -s 9.5 --alignment-type 2`; window PDBs were cut from the model chain and renumbered from 1; the three databases were pooled. The name test of the amended classifier (A1.1, 'VA') reads the upper-cased target header only (no taxon names; the frozen original also read taxon names): the regular expression `ACETYLTRANSF|bGNAT\b|GCN5|ACETYLASE|N-ACETYL` (it therefore also matches DEACETYLASE), or the string 3.40.630, or an accession of a Pfam-A family of clan CL0257 (69 families in the clan listing used). In the peak rule the alignment length is the query span ($qend - qstart + 1$) and the target coverage is $(tend - tstart + 1)/\text{target length}$. *Peaks*. The midpoint $(qstart + qend)/2$ of each alignment is binned in 25-aa bins; bins holding at least 25% of the maximal bin count are peaks (adjacent bins merged); a peak is valid when at least 10 distinct targets have their midpoints in it (key: the PDB entry id for pdb100, the database plus the first header token otherwise). The valid peaks of all windows are ordered by support (descending), then by smallest E; a peak is accepted when its centre, $(\text{median } qstart + \text{median } qend)/2$, lies at least 100 aa from every accepted peak. A lobe candidate is a peak with a span of 60-140 aa, target coverage ≥ 0.4 , at least 10 targets and a centre at least 90 aa from the accepted peaks. The first GNAT of a chain is its accepted peak of largest support (for P1, a peak with centre in residues 434-661 is the known peak and one at 433 or below is outside it). *Which model is scored*. For AlphaFold2, the best-ranked model (ColabFold rank by pLDDT) of each seed, called 'rank-1' in this paper; for Boltz-2, `model_0` (highest confidence score) of each seed. The file names `rank_001` in the data supplement rank over all 15 AlphaFold2 models of a run and need not be the best of each seed. *D3 seed call*. The region R is the chain without the first-peak span (rounded outward) and, for P1, without residues 484-658; slots (141 aa, step 10, plus one aligned to the end of each stretch) exist only in stretches of R of at least 141 aa; a chain without a slot is UNRES. A slot is resolved when its mean pLDDT is at least the floor (AlphaFold2 90, Boltz-2 60) and the mean PAE of its slot-by-slot block is at most 10 Å (no PAE: unresolved). A seed call is POS with at least two accepted peaks (for P1 at least one outside the known peak); NEG_RESOLVED when it is not POS, the first GNAT was recovered, every slot is resolved, there is no lobe candidate and the pipeline completed; LOBE when only the lobe candidate prevents NEG_RESOLVED; otherwise UNRES. A cell label (partner, engine) is 'ok' for POS in at least two of three seeds, 'not_found' for NEG_RESOLVED in all three, otherwise 'unknown'. The partner verdict of D3 combines the two engines (ok/ok and ok/unknown give YES, not_found/not_found gives NO, anything else UNKNOWN); for the lone chains of D13 no verdict was assigned. *Census classes*. A chain is resolved when the fraction of residues with pLDDT ≥ 70 is at least 0.5 in at least two of three seeds; a chain is TANDEM when at least two accepted peaks occur in at least two seeds, SINGLE with exactly one, NONE with none (MIXED otherwise; none occurred). *Effective sequence number*. Neff80: rows with at least 50% non-gap columns are kept (insertions removed); each row has weight $1/n$, where n counts the rows (itself included) with identity ≥ 0.8 over the columns that are non-gap in both; the query and all server and family rows are included. *Family de-duplication*: `mmseqs easy-linclust --min-seq-id 0.9 -c 0.8`, trailing stop characters removed, rows identical to the query removed. *Structure panel*. Windows of 150 residues, step 25, the last window aligned to the chain end (a chain of 150 or fewer residues gives one window); TM-align with default options, TM-score normalised by the window; the threshold $t = \text{round}(\text{highest score of the non-GNAT references} + 0.01, 2)$; self and sibling entries are excluded. Non-GNAT references (first chain, first model of each file): 3CLA, 1HV9, 1LXA, 1OCX, 3CHY, 1TIM, 4AKE, 2TRX, 2CMD, 1LDM, 1FLV; long non-GNAT structures of D11: 1GPB, 1PKN, 1AON, 1ALK, 1BGV, 1PGJ, 2ACE, 1E7P, 1FCB; D12 negatives: 3EH0, 1NDB, 1KRR, 1QMG, 1ECF, 1YP2, 1PJQ, 2AEV, 1QOR, 1HZ4, 2ACE, 1PGJ. 'Two regions' means two windows with start positions at least 150 residues apart that both reach t . The control calibration that decides whether a negative may be called at all is reported in the Results; every verdict reported here is UNKNOWN because of unresolved slots, which does not depend on that calibration.

Census. For each of the 91 chains (P1, P2 and the family), one of four classes was assigned per predictor: TANDEM (two peaks in at least two of three seeds), SINGLE (one peak), NONE (no peak in a resolved chain) and UNRESOLVED (not resolved by the rule above).

Exploratory analyses. (a2) Row-shuffled alignment control (D10): the D5-B family alignments with the residues of every family row permuted within the row (gaps, row count and server rows kept; the shuffled alignments and the resulting models are retained in the public data supplement), and the same lone-chain predictions. (a3) Nine long non-GNAT experimental structures scored with the D8 operation (D11); the family Boltz-2 models used by D8 are the seed-0 models and the AlphaFold2 models the rank-1 model over seeds. (a0) Depth-matched controls (D7): the same Logan search and the same pipeline for Eis, MshD, Orf14 (tandem GNAT) and GImU, DnaK (non-GNAT), each with nested random subsets of 45 and 89 homologue rows (the sizes of F and F2), and a Logan search with the N-terminal windows of P1 (residues 1-200) and P2 (residues 1-110). (a1) metapredict 3.0.2 [18-20] disorder prediction of the family chains. (a) TM-align [12] (normalised by tGCN5, 1QSR chain A; seed-0 top model) of tGCN5 against the chain and then against the remainder, with a control panel. (b) Pfam-A 38.2 [13,14] (HMMER 3.3.2 [16]) with the GNAT clan CL0257 (D6); later (D8-B) hmmscan with the 98 HMMs of that clan downloaded from InterPro (gathering thresholds; clan hits only, because the full Pfam-A download had stalled); and then (D8-B2) the same scan against the full Pfam-A (30,134 HMMs, supplied by the operator) on seven distinct GNAT proteins (eight entries, Eis once as its D7 model and once as 3R1K) and twelve distinct non-GNAT proteins (GImU, which is also the PDB reference 1HV9, DnaK and ten further PDB references). (c) The CA RMSD (index-wise Kabsch, 28 residues) of every R/Q-x-x-G-x-G/A segment against tGCN5 residues 125-152 (beta4, P-loop, alpha3), calibrated on control GNAT motif A segments and non-GNAT controls. (d) Overlap of F with the server rows.

Deviations and corrections (all recorded in the project record): the first full-Pfam scan (D8-B2) applied no clan filter because of a silent string-replacement miss in the script and printed non-GNAT hits as clan hits; it was caught from the output and repeated with the corrected script. The D8 amendments A2, A3 and D8-D were registered after earlier D8 output had been read (the second-region score S2 and the AlphaFold2 and window replicates were motivated by it); the code of the S2, chimera and Pfam steps was committed together with their results; and the registration text of D8-A2 misstated the first panel output (it says every scored chain reached the threshold, the result file shows two below it). A probe scan of the same family by another session added no new cluster beyond F2 under the registered rule (D9; of 83 candidates of 300-1000 aa, 66 had no MMseqs2 hit to P1 or P2 at $E \leq 1e-5$ and were not examined further). Stop-codon characters at the ends of the Logan sequences stopped the first launch of the census (removed, rerun). The first member-expansion run located members through the wrong index field (724 of 784 missing) and was discarded and repeated. A result file was scored before the last Foldseek batch had finished and was rescored. An analysis crashed once on a tuple unpacking and a CIF parser read a chain incompletely once (both fixed before any result was read or with identical labels). The first-stage controls did not share the MSA source of the enriched targets, a deviation from the pre-registered single-source rule that was handled by the rule's depth brackets and later addressed by depth-matched controls (D7). An earlier statement that the family was absent from the server database was wrong (Section 3.2). Finally, in the last stage the candidate list drawn up in conversation contained two partners that the completeness check showed to be cut at the contig end; they were excluded before any prediction, and the longest complete partner of the L1 lineage proved to be about 850 residues, not the 674 first assumed.

3. Results

3.1 Public alignments leave the partners unresolved

With the ColabFold server alignments (P1 7 rows, P2 18 rows; Neff80 3 for both), the partner chains were largely low-confidence in both predictors. Co-fold mean pLDDT was 35–38 for P1 with AlphaFold2 and 46–47 with Boltz-2, and 31–32 and 41–42 for P2. The frozen rule returned UNKNOWN for P1 and UNKNOWN for P2 in all 9 sensitivity variants: at most 4 of the 34–36 slots outside the first GNAT reached the resolved floor in any unit and seed, so the absence of a second GNAT could not be asserted. For the control cells that passed (Eis, RimI), the pipeline detects GNAT folds when the chain is predicted with confidence.

3.2 Logan holds many more homologues than the server alignment returned

The server alignments were small. UniRef50 (release 2026_03) returned a single distant sequence (reported by 3 of the four queries), and RefSeq viral proteins and the GenBank phage division returned no homologue besides the two paralogs. The Logan 50% protein-cluster representatives, in contrast, held 43 distinct homologues in the first pass (rule-counted: P2 36, P1 residues 1–483 32, P1 full length 29, P1 residues 484–661 20; family F 45). These are not disjoint from the server rows: 14 of the 45 family members have at least 90% identity (80% coverage) to one of the 25 server rows, none is identical, and 43 have some hit. The server database therefore contained close relatives of many members; the difference is the number and diversity of rows, not the existence of the family. The sequences come from sequencing-run assemblies and may include loci of the original description (Logan is that work's own source; several hit identifiers coincide with partner proteins of array loci recorded elsewhere in this project, unverified).

3.3 Supplying the alignment depth resolves much of the partners

Adding the family rows raised the lone-chain mean pLDDT (Figure 2). For P2 with AlphaFold2 it rose from 38–47 (server rows only) to 80–83 (60 rows), and for P1 from 50–54 to 78–79 (47 rows). The rise was monotone for P2 and for P1 with AlphaFold2 across $k = 8, 16, 32$ and all added rows; for P1 with Boltz-2 it was not monotone ($k = 32$ equal to $k = 16$). Doubling the family (89 sequences) changed little: the P2 means moved to 84.1–84.4 (AlphaFold2), the P1 means fell slightly, and the Boltz-2 lone P1 chain no longer met the resolution rule. Neff80 rose only from 32 to 43 (P1) and from 38 to 47 (P2), so "saturation" here means no further gain at this level of added diversity. By the registered descriptive slot rule, P2 was resolved in both predictors, lone and co-folded, and P1 was borderline (not in the Boltz-2 co-fold; the frozen strict rule was met by no P1 unit).

The MSA-free predictor Chai-1 folded the control proteins (Eis, RimI, DnaK, UPP) but left the partners unresolved (mean pLDDT P1 51.6–52.4, P2 52.0–52.7). This shows that the earlier failure is not specific to the two predictors used first; it does not show that a predictor with an equally deep MSA would fail.

A row-shuffled control (D10, registered before any model was produced; Figure 7) tested whether the rise could be produced by row count or alignment format. It kept the row count, gaps and composition of the family rows but permuted the residues within every family row. The lone-chain mean pLDDT then stayed near or below the server-only level (at most about 5 pLDDT above it; P2 with AlphaFold2 35 against 82 with the real rows; P1 41 against 78; Boltz-2 P2 41 against 77, P1 49 against 69). By the registered rule all four partner-by-engine cells are therefore content-specific (all_content_specific = True): the rise is not produced by row count or alignment format. The control does not show that the real rows carry correct co-evolution, only that they carry information that shuffled rows do not. Because shuffled rows keep composition but share only a low identity with the query (about 7%), the control separates alignment content from row count and format but not conservation from co-variation. The dose-response rests on two chains and three seeds.

3.4 The frozen verdict stays UNKNOWN, and the unresolved regions are the N-termini

With the family alignment the frozen D3 rule still returned UNKNOWN for P1 and UNKNOWN for P2 (the same with the expanded family: UNKNOWN / UNKNOWN) in all sensitivity variants. The main reason was that the N-terminal regions (about the first 200 residues of P1 and 100 of P2; Figure 8 shows the per-residue profile) remained low-confidence in every model; some slots outside the N-terminus also failed the floor, particularly for Boltz-2 P1. In the resolved parts, P1 carried exactly one GNAT-like Foldseek peak, in the C-terminal region, and P2 none, in both predictors and all seeds. The N-termini are not predicted to be intrinsically disordered (metapredict: disordered fraction 0.0 for P1 residues 1-200 and 0.0 for P2 residues 1-100; median over the family 0.0; Spearman correlation with the AlphaFold2 N-terminal pLDDT -0.168). A Logan search with the N-terminal windows as queries returned 1 homologue for the P2 window and none for the P1 window (MSAs of 2 rows for the P2 window and 1 row for the P1 window), and the window models stayed low-confidence with no Foldseek peak. These regions therefore behave like orphan regions: ordered by sequence, but without detectable homologues in the searched resource. More rows could not be obtained, and nothing can be said about their fold.

3.5 No resolved family chain has two accepted GNAT-like Foldseek peaks

Across the 91 family chains (Figure 3), none of 546 chain-engine-seed predictions had more than 1 accepted peak. Among chains of 450-750 aa ($n = 70$), AlphaFold2 gave SINGLE 34, NONE 32, UNRESOLVED 4; Boltz-2 gave SINGLE 34, NONE 17, UNRESOLVED 19 (all lengths: UNRESOLVED 6 and 24). The predictors disagreed on 2 of these chains.

The detector is adequate at this depth for full-size tandem GNATs. At 45 or 89 homologue rows it called the tandem GNAT controls Eis, MshD and Orf14 TANDEM in every call (two of three seeds; MshD at 89 rows with AlphaFold2 had one peak in one seed) and DnaK NONE (Section 2). The absence of TANDEM chains is therefore not explained by a loss of sensitivity to full-size tandem GNATs at this row count. The controls match the family in row count but have 60-93% of residues 1-150 covered, against a much lower coverage in the family's NONE chains, and the detector was not shown able to call the 91-residue second region of the original description. The same controls show that the detector also over-calls (GlmU), so SINGLE chains are acetyltransferase-like Foldseek peaks that the TM-align results below support as GNAT-like without settling it. The statement "no TANDEM" therefore holds among resolved chains (66 of 70 with AlphaFold2, 51 with Boltz-2).

The single peak lies in the C-terminal part (centre at 0.78-0.86 of the chain length over all 100 single calls; median 0.84). P1 is a SINGLE chain (its peak is the region already known to be GNAT-like) and P2 a NONE chain in both predictors. The NONE half is partly unmeasured: with AlphaFold2 the median mean pLDDT of residues 1-150 is 47 in NONE chains against 89 in SINGLE chains, and the median MSA coverage of those residues is 0.17 against 0.37. A lobe analysis (60-140 aa) supported no lobe candidate in two or more seeds in any chain (AlphaFold2 0, Boltz-2 0 chains), but with a controls' lobe-size limit of 120 aa this does not exclude a 91-residue lobe.

3.6 The TM-align operation and the structure-search detector disagree

The TM-align operation of the original description gave a first region with TM 0.57 in SINGLE chains and 0.58 in NONE chains, and a remainder with TM 0.55 in SINGLE chains versus 0.39 in NONE chains (Figure 4). In 29 of 50 SINGLE chains the Foldseek peak lies inside the TM-align "second" span, so the pair is the Foldseek-confirmed C-terminal region plus an upstream region that Foldseek does not call. That upstream region scores TM 0.57-0.58 also in NONE chains, so its similarity to tGCN5 is not specific. The control GNAT proteins score 0.73 (Eis, first region) with a remainder of 0.54, and non-GNAT controls 0.32-0.35 for the first region; the family values lie between them. TM-align against a single reference therefore does not settle a domain count here.

3.7 No family chain has a sequence-profile GNAT signature or a motif-A backbone

No chain had a hit in the Pfam GNAT clan (CL0257) with the gathering threshold (0 chains) or at $E \leq 1e-3$ (0 chains). The first run included no positive control, and the Foldseek-confirmed C-terminal region was also negative. The later control run (Section 3.8) found a clan hit in 8 of 8 GNAT controls, none in non-GNAT controls and still none in the family (0 of 91), so the negative is informative: the family has no Pfam GNAT signature. The full-Pfam repeat (D8-B2, competing families present) gave the same clan counts (8 of 8 GNAT controls, 0 family chains), and 88 of 91 family chains had no Pfam family hit of any kind at the gathering threshold. The remaining chains hit single unrelated families (RVT_1, Rotamase, WD40, PFU), so the family is largely unannotated by sequence profiles.

R/Q-x-x-G-x-G/A segments occurred in 27 chains (31 segments, more than expected from composition), but their backbones do not fit the tGCN5 P-loop segment (CA RMSD 4.8–5.0–9.0 Å; median 5.0 Å), in contrast to the motif A segments of the control GNAT proteins (0.4–2.3 Å). The family values coincide with the lower end of non-GNAT segments (minimum 5.0 Å), which may be the floor of this measure for non-matching backbones. The scan covers only segments that start with R or Q. Two of the ten true GNAT motif-A segments of the controls (RMSD 2.2–2.3 Å, MshD first domain) also lie above the 2 Å line, so the registered cut-off is not a clean separator, and the registered motif-context test failed its own positive controls. The recurrent R-K-G-G-x-A segment noted in the review of the census lies in the region that the structure-panel test flags as GNAT-like. GNAT-likeness of the family is therefore at most a fold-level, distant one without a canonical motif A.

3.8 A name-free structure-panel test finds GNAT-like windows (registered two-region rule; second-region score S2 is exploratory)

The references were five experimental GNAT structures (tGCN5 1QSR [21], Eis 3R1K [24], yeast Gcn5 1YGH [27], serotonin N-acetyltransferase 1CJW [28], 1BO4 [29]; PDB [31,32]) and eleven non-GNAT structures. Every 150-residue window (step 25) of a chain was aligned with TM-align to the GNAT panel; the window score is the TM-score normalised by the window, taken for the best reference, with the query itself and its sibling excluded, and the chain score is the best window score. The threshold $t = 0.46$ was fixed from the references alone (largest non-GNAT chain score 0.45; leave-one-out GNAT scores 0.71–0.75; sensitivity 1.00). Because the family consists of predicted models, the identical operation was applied to the depth-matched predicted controls: Eis, MshD and Orf14 scored 0.69–0.72 and GlmU and DnaK 0.35–0.38, so t transferred unchanged. Of the 88 family chains with a window of mean pLDDT ≥ 70 (Boltz-2, seed 0), 86 reached t , including all 17 Foldseek-NONE chains (minimum score 0.59); the median chain score was 0.62, below the GNAT references' 0.71–0.75. The registered "two regions" rule (applied before the second-region score S2 was introduced) was met by 56 of the 91 chains.

The second-region score S2 is the score of the best window lying at least 150 residues from the best window. It was above the claim cut (0.46, the larger of t and the largest predicted non-GNAT S2) in 55 of the 71 chains with a computable S2 (55 of all 91 = 60%; run-bootstrap 95% interval for the share of all chains 51%–70%). These were 50 of the 50 SINGLE chains and 3 NONE and 2 UNRESOLVED chains. For P1 the windows above t start at residues 276–351 and 451–512, that is, a block upstream of the Foldseek-confirmed C-terminal region. In 71 chains a second window could be computed; the remaining chains lack a second window of mean pLDDT ≥ 70 , so the 55 of 71 (and 55 of all 91, 60%) above is not a prevalence among all chains. In nearly all SINGLE chains (49 of 50 within 200 residues of each other) the two windows are adjacent tiles (a score dip lies between them), so the data do not distinguish two domains from one long GNAT-like region.

The registered robustness criteria were met, although the shares criterion is permissive, and AlphaFold2 and Boltz-2 used the same family alignment, so the replicate is not independent of the alignment. AlphaFold2 rank-1 models gave 62%; windows of 130 and 100 residues gave 68% and 81% (with 13 NONE chains at 100 residues against 3 at 150, so specificity falls as windows shrink). A chimera of tGCN5 and DnaK (800 residues) had S2 0.39, below the cut, but this is satisfied by construction (tGCN5 is shorter than twice the minimum window separation) and is not evidence of specificity.

Two further negative controls were run. As an additional negative control for the S2 rule (D11, registered before it was run), 9 long experimental non-GNAT structures (449-823 aa, nine proteins of different families, two of them Rossmann dehydrogenases) were scored with the same operation and threshold: none reached t (largest chain score 0.43, largest S2 0.43; 0 of 9 above the cut). They are experimental structures, not predicted ones. A further control (D12, registered before any model; Boltz-2 models with ColabFold-server alignments) predicted 12 non-GNAT proteins (203-596 aa), among them the non-GNAT acyltransferases LpxD, carnitine acetyltransferase and galactoside acetyltransferase and nine other alpha/beta enzymes. None reached t (largest chain score 0.45, largest S2 0.41; 3 proteins (203-374 aa) had no window 150 residues or more from their best window, so no S2 exists, and 0 of the rest exceeded the cut), although their models are confident (mean pLDDT 91-97), so more of their windows were scored than for the family. Their alignments are deep server alignments rather than family-depth ones, and one protein (acetohydroxy-acid isomeroreductase) scored 0.45 against $t = 0.46$. Figure 6 shows the scores. The clan-HMM control (Section 3.7) is the sequence-level counterpart: GNAT controls hit, the family does not.

This test measures the similarity of 150-residue windows to a small GNAT panel; windows partly match long references and no domain boundary is inferred. It does not use the Foldseek classifier's name rule, which is why it is not subject to the GlmU false positive, but its own specificity is shown only by the controls above.

3.9 Partners of other ART loci with complete sequences (registered D13)

P2 is the partner drawn for the LPJP1 locus (ART_11) in the preprint's Fig. 2B. The figure gives an RT of 1365 nt, a partner of 1446 nt and a gap of -20 nt (an overlap), and the GenBank record MZ422438.1 has RT2 of 1365 nt, P2 of 1446 nt and the same overlap; the gene of 1986 nt drawn upstream is P1. In our extraction of the vector PDF, Supp. Fig. 3C draws the partners of ART_11, ART_69 and ART_34 in the Type I class, but the text does not say so for each locus, and the GNAT analysis was of another locus (ART_57). We therefore added partners of other loci whose complete sequences we held. Of 6 candidates, 4 passed the completeness check; the other two (473 and 674 residues) were cut at the contig end and were excluded before any prediction (Table 1; numbers/NUMBERS_D13.json, key partner_qc).

Partner	Basis of the locus	Length (aa)	Server rows (distinct)	Family rows	Neff80	AlphaFold2 mean pLDDT	Boltz-2 mean pLDDT	Resolution rule met (AlphaFold2 / Boltz-2)
ART_69 partner	figure fingerprint (Supp. Fig. 3C)	594	6	35	33	88.4–88.8	80.0–80.1	strict / strict
ART_34 partner	figure fingerprint (Supp. Fig. 3C)	890	6	18	13	78.1–81.1	72.2–73.0	strict / strict
L1 partner (ERR10896470)	separate accession; no displayed fingerprint match	852	6	15	13	67.6–68.9	58.0–61.1	RESOLVED_70 / no
L1 partner (SRR26805139)	separate accession; no displayed fingerprint match	850	5	14	9	68.0–69.5	59.0–60.2	RESOLVED_70 / no

Table 1. Partners of other ART loci analysed in D13. Server rows: distinct sequences in the ColabFold alignment (the query counted once). Family rows: Logan homologues after 90% identity de-duplication. Neff80 is that of the alignment with server and family rows. Mean pLDDT is over the whole chain, range over three seeds. "strict": at least half of the 141-residue slots have a mean pLDDT of at least 90 (AlphaFold2) or 60 (Boltz-2) and an intra-slot PAE of at most 10 Å, in at least two of three seeds (registered rule of stage D4); "RESOLVED_70": the descriptive slot rule of Section 2 only; "no": neither. The frozen seed-call rule below is stricter (every slot outside the first peak). No server-only prediction was run for these partners, so the effect of adding rows is shown by the k ladder and the shuffled control.

Server alignments held 5-6 distinct rows, and the Logan search returned 14-35 family rows per partner, fewer than the 45 of the LPJP1 family. With server and family rows together all four partners met the registered first-tier criterion, which counts a partner as resolvable when at least one engine meets the rule; the two L1 partners met it in AlphaFold2 only (Table 1). The ART_69 partner, the one of about the described size, reached an AlphaFold2 mean pLDDT of 88.4–88.8 with 100% of its 141-residue slots above 70 in every seed, and 88% of its residues reached pLDDT 70 in the seed-0 model (Figure 10). The N-terminal part that remained unresolved for P1 and P2 in the tested engines and settings is largely confident here: 18% of the residues before the start of the Foldseek peak are below 70, and none from the peak on. In the three longer partners the first part of the chain stayed much less confident: 37% (ART_34 partner), 57% and 64% (L1 partners) of the residues before the peak are below 70, against 5%, 20% and 11% from the peak on.

Only the ART_69 and ART_34 partners met the strict rule in both engines; the two L1 partners met the descriptive rule in AlphaFold2 only. The gain was content-specific. Replacing the family rows by rows whose residues were permuted within each row (the operation of Section 3.3) lowered the AlphaFold2 mean pLDDT to 37-49 and the Boltz-2 mean pLDDT to 41-57 over the four partners; the real rows exceeded the shuffled rows by at least 24.1 (AlphaFold2) and 12.5 (Boltz-2) pLDDT units. For the ART_69 partner the AlphaFold2 mean pLDDT rose with the number of family rows ($k = 8, 16, 32$ and all: 76.4–79.9, 81.7–85.1, 87.5–88.1, 88.4–88.8; Figure 9).

The three readouts again gave different answers. Foldseek (amended name-based classifier, unchanged gates; it over-calls GImU) accepted exactly one GNAT-like peak in each of the 24 seed calls, in the C-terminal part of the chain (AlphaFold2 seed-0 spans: ART_69 partner 405–557, ART_34 partner 676–833, L1 partners 641–805 and 639–795). The Pfam GNAT clan models (exploratory) found no hit in any partner, whereas the same models found hits in the three positive controls. The structure-panel test ($t = 0.46$) found two non-overlapping windows above t in 24 of 24 models (second-region score S_2 0.57–0.63); the window starts fall in two adjacent groups (ART_69 partner: 201–276 and 376–445), and the second group contains the Foldseek peak (Figure 10). The two-region criterion is the registered D8 rule; the three-seed model set and S_2 values are post hoc (D13-A3). Adjacent tiles do not distinguish two domains from one long GNAT-like region (Section 3.8). The frozen seed-call rule returned UNRES in all 24 seed calls (the part of the chain outside the first peak did not reach the strict floor or the PAE limit), and all 8 cell labels were unknown. No verdict of the D3 table was assigned.

Three of the four partners are longer than the "about 600 residues" of the description, their family sets are small, and the structure-panel test has more windows that can cross the threshold in a longer chain; its specificity was checked on nine long experimental non-GNAT structures (Section 3.8), not on these lengths. The four partners are three lineages (the two L1 partners are closely related), were found as relatives of P1, and their seeds and engines share one alignment per partner, so the 24 of 24 models are not independent observations; their overlap with the 91-chain family was not checked. In the Eis depth ladder, complete fold-instance detection was demonstrated at Neff80 17 and 124; the depth-matched controls used 45 and 89 homologue rows. Of the 4 additional partners, 3 have Neff80 9-13 with 14-18 family rows, inside an untested

depth range. For these partners, exactly one Foldseek peak is not a demonstrated negative for a second full-size region. The result is therefore the same as for the LPJP1 family, with one difference: the unresolved N-termini of P1 and P2 are not a property of every partner of this family.

4. Discussion

The methodological result is that the unresolved partner models were largely a consequence of too few alignment rows, and that a public metagenomic resource held many more homologues than the server alignment returned. A prediction-based "no second domain" made on a low-confidence model would have been uninformative here, and the pre-registered rule correctly returned UNKNOWN. The dose-response in the number of rows (Figure 2) is the evidence for this. Its limits are stated in Section 3.3: two chains, a row-shuffled control that cannot separate sequence conservation from co-variation, N-termini that remained unresolved in the tested engines and settings, and Chai-1 testing predictors without alignments.

On the original question, the data do not reproduce two GNAT-like Foldseek peaks in any resolved family member. In an exploratory second-region analysis, a name-free structure-panel test (Section 3.8) finds two non-overlapping GNAT-like windows in all 50 SINGLE-class chains, mostly as adjacent tiles that do not distinguish two domains from one long region. The calibration references were used to set the threshold and are not independent negative validation. In the additional tests, 9 long experimental non-GNAT structures and 12 predicted non-GNAT proteins stayed below it. These are different control sets, not a count of distinct independent proteins: two proteins occur in both sets as an experimental structure and a prediction. The largest predicted non-GNAT chain score was 0.45 against $t = 0.46$. The lobe that the original description reports as the second region, in turn, is below what our detector was shown able to call. These findings do not refute the description: TM-align, the method used there, returns a second region in the chains that carry the C-terminal peak (although not specifically), the N-termini are low-confidence, and the original alignments were locus-paired family blocks we could not reproduce. The most defensible reading is that any second GNAT-like region would not be a canonical GNAT by sequence profile, motif-A structure or structure search, and that its description as GNAT-like depends on the method used to compare folds.

Partners of other loci whose sequences we hold (Section 3.9) point the same way. The readouts disagree again, now in four more proteins, and the one partner of about the described size (ART_69, 594 residues) is predicted with confidence over most of its length. That shows that an unresolved N-terminal region is not a property of every partner of this family; it does not show why P1 and P2 differ, since the alignment of the ART_69 partner is not deeper (Neff80 33, against 32 for P1 and 38 for P2 with the family rows), and the first part of the three longer partners remains largely low-confidence. Nor does it settle whether the region around the single Foldseek peak is part of a second GNAT: the window set that the structure panel calls GNAT-like extends upstream of the peak into a region that is confidently predicted and has neither a Pfam signature nor a Foldseek peak. Each of the three readouts has its own known limits (a detector shown only on full-size GNATs, clan models that may miss a diverged fold, a window score without a domain boundary), and together they cannot separate a diverged GNAT-like fold from another alpha/beta fold.

Limitations. The chains are sequence-search homologues from sequencing-run assemblies (possible chimeras or contaminants; four of the 45 first-pass members entered on one marginal hit), and they are clustered representatives of one family rather than independent loci, whose RT partners and genomic contexts are unknown. The census MSAs were built from the family itself, and P1 and P2 were not blind to the results of earlier stages. The primary structures are predictions from the two engines tested. Statements about unresolved regions and detector outcomes apply to these models and settings, not to every predictor. AlphaFold 3 was not evaluated and might yield more confident models or different structural readings. That possibility does not invalidate the observed within-predictor alignment comparisons, but it limits claims

about what can be resolved with other predictors. An AlphaFold 3 evaluation would require its own control checks; predictor-specific confidence floors are not assumed to transfer unchanged. The AlphaFold2 resolution floor of the frozen rule is not a validated limit, and part of the "no GNAT" half is unmeasured because the N-terminal regions are poorly predicted. The detector's target-coverage rule can hide a GNAT inside a large multi-domain target, and its peak-separation rule merges tandem domains closer than 100 aa (it merged the two GNATs of MshD in one engine). TM-align used the seed-0 model only. The first-stage controls did not share the MSA source of the enriched targets (the depth-matched controls of D7 did, and exposed the false positive of the name-based classifier). The motif scan covers only segments starting with R or Q, and the full-Pfam scan used gathering thresholds only and sequences read from model files. The structure-panel test uses 150-residue windows against five GNAT references (its specificity falls with smaller windows), and its S2 score was added after the first panel result was read (exploratory, registered before it was computed). Its negative controls are eleven short and nine long experimental structures, twelve predicted non-GNAT proteins (D12) and two predicted proteins (GlmU, DnaK; two depths each), plus one chimera that cannot fail by construction; only three of the twelve predicted negatives of D12 are non-GNAT acyltransferases. The last stage adds its own limits: four partners, lone chains only, three of them outside the 450-750 aa band that we used (the preprint states 'about 600 residues' and draws the ART_34 partner, 890 residues, as Type I), two of them one lineage, family sets of 14-35 rows, loci identified by a figure fingerprint rather than an accession (ART_69, ART_34) or taken from separate accessions whose loci did not match the displayed figure fingerprints (L1; absence from the preprint's full locus set was not established), and a Pfam scan with the clan HMMs only that was added after the other readouts had been read. In the Eis depth ladder, complete fold-instance detection was demonstrated at Neff80 17 and 124; the depth-matched controls used 45 and 89 homologue rows. Of the 4 additional partners, 3 have Neff80 9-13 with 14-18 family rows, inside an untested depth range. For these partners, exactly one Foldseek peak is not a demonstrated negative for a second full-size region. The two partners excluded for truncation show that a partner reading frame must be checked for completeness before it is used. No statement is made about enzymatic activity or acetyl-CoA binding. A direct test needs experiments (an acetyltransferase assay), additional structure-search methods with several GNAT references, or the family alignments of the original description.

5. Declarations

AI assistance. This work was carried out with a scientific-discovery harness developed by AI3 Discovery, operated together with an AI coding assistant (Claude, Anthropic) acting as an autonomous coding and analysis agent under the operator's direction. The harness and the assistant were used to write and run the analysis and workflow scripts, to organise the structure predictions and searches on the compute machines, to apply the decision rules that had been specified in advance, to draft and revise the manuscript, and to run automated review passes, including checks of claims against the result files, English-Korean consistency checks and adversarial reviews by AI agents. These reviews were performed by AI and are not independent human peer review. The harness itself is not shared. The author is responsible for the content of the manuscript. **Competing interests / disclosure.** The work that is tested (Yoon et al. [1]) was authored by employees of Anthropic, and the AI system used here (Claude) is a product of Anthropic. The author has no affiliation or research collaboration with Anthropic or the authors of Yoon et al. The author previously emailed the corresponding authors to request locus accessions and partner sequences; no reply or non-public data was received. The present analyses use public sources. The reviews labelled 'independent' in this work were performed by AI sessions of the same model family and are not independent in the sense of human peer review. **Funding.** None; the work was carried out by a single researcher at AI3 Discovery without external funding. **AlphaFold 3** was not used because its official model parameters had not been obtained for this study. Access to the public parameters is governed by organisation and purpose eligibility conditions separate from the code licence. This is a statement about the access route used when the study was

conducted, not a claim that every commercial access route is unavailable. **Data and code availability.** This revised manuscript and its figures are released through [AI3 Discovery](#). Data for checking and re-deriving the reported results are publicly available at [Zenodo](#), under CC BY 4.0. The data supplement contains protein sequences, input family alignments, predicted models, Foldseek hit tables, per-unit result tables, computed numbers and file-level SHA-256 manifests. File names cited in the text, such as `numbers/NUMBERS_D13.json` and `integrity/SHA256SUMS`, refer to paths inside that supplement; its README describes how to reassemble the split archives and check their hashes. The manuscript PDFs included in the Zenodo record are the 3 October 2026 version and are superseded by the revised manuscript on the AI3 Discovery site. The current revision corrects prose and Figure 1 layout; the deposited research data files are unchanged. The analysis and scoring code, calibrated threshold files, probe models, harness modules and pre-registration texts are not shared. In their place, the data record includes integrity commitments consisting of pre-registration hashes, sizes and local git commit identifiers, together with combined hashes for withheld file categories; these do not establish an externally timestamped registration. Public inputs are identified in Methods, and the number and table files identify the Logan objects by ETag and size. Data enquiries may be sent to discovery@ai3.kr. **References.** Preprints are marked as such in the list; entry 1 is the work that was tested.

1. Yoon PH, Athukoralage JS, Ameisen E, Kauderer-Abrams E, Perry NT, Durrant MG (2026). Autonomous AI agents discover reverse transcriptases with tandem repeat arrays. Manuscript hosted by Anthropic. <https://www-cdn.anthropic.com/22573675ada52a8ca8a97a1a4b4326b2f208a071.pdf>.
2. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022). ColabFold: making protein folding accessible to all. *Nature Methods* 19(6):679-682. doi:10.1038/s41592-022-01488-1.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583-589. doi:10.1038/s41586-021-03819-2.
4. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. (2021). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. doi:10.1101/2021.10.04.463034.
5. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR, et al. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*. doi:10.1101/2025.06.14.659707.
6. Wohlwend J, Corso G, Passaro S, Getz N, Reveiz M, Leidal K, et al. (2024). Boltz-1 Democratizing Biomolecular Interaction Modeling. *bioRxiv*. doi:10.1101/2024.11.19.624167.
7. Chai Discovery, Boitreaud J, Dent J, McPartlon M, Meier J, Reis V, Rogozhnikov A, Wu K (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*. doi:10.1101/2024.10.10.615955.
8. ByteDance AML AI4Science Team, Chen X, Zhang Y, Lu C, Ma W, Guan J, et al. (2025). Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv*. doi:10.1101/2025.01.08.631967.
9. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 42(2):243-246. doi:10.1038/s41587-023-01773-0.
10. Steinegger M, Söding J (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35(11):1026-1028. doi:10.1038/nbt.3988.
11. Steinegger M, Söding J (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 9(1):2542. doi:10.1038/s41467-018-04964-5.
12. Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 33(7):2302-2309. doi:10.1093/nar/gki524.
13. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research* 49(D1):D412-D419. doi:10.1093/nar/gkaa913.

14. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. (2024). The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research* 53(D1):D523-D534. doi:10.1093/nar/gkae997.
15. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. (2024). InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* 53(D1):D444-D456. doi:10.1093/nar/gkae1082.
16. Eddy SR (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology* 7(10):e1002195. doi:10.1371/journal.pcbi.1002195.
17. Chikhi R, Lemane T, Loll-Kripplbeier R, Montoliu-Nerin M, Raffestin B, Camargo AP, et al. (2024). Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. *bioRxiv*. doi:10.1101/2024.07.30.605881.
18. Emenecker RJ, Griffith D, Holehouse AS (2021). Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical Journal* 120(20):4312-4319. doi:10.1016/j.bpj.2021.08.039.
19. Emenecker RJ, Griffith D, Holehouse AS (2022). Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *bioRxiv*. doi:10.1101/2022.06.06.494887.
20. Lotthammer JM, Hernández-García J, Griffith D, Weijers D, Holehouse AS, Emenecker RJ (2024). Metapredict enables accurate disorder prediction across the Tree of Life. *bioRxiv*. doi:10.1101/2024.11.05.622168.
21. Rojas JR, Trievel RC, Zhou J, Mo Y, Li X, Berger SL, Allis CD, Marmorstein R (1999). Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide. *Nature* 401(6748):93-98. doi:10.1038/43487.
22. Vetting MW, de Carvalho LPS, Yu M, Hegde SS, Magnet S, Roderick SL, Blanchard JS (2005). Structure and functions of the GNAT superfamily of acetyltransferases. *Archives of Biochemistry and Biophysics* 433(1):212-226. doi:10.1016/j.abb.2004.09.003.
23. Dyda F, Klein DC, Hickman AB (2000). GCN5-Related N-Acetyltransferases: A Structural Overview. *Annual Review of Biophysics and Biomolecular Structure* 29(1):81-103. doi:10.1146/annurev.biophys.29.1.81.
24. Chen W, Biswas T, Porter VR, Tsodikov OV, Garneau-Tsodikova S (2011). Unusual regioversatility of acetyltransferase Eis, a cause of drug resistance in XDR-TB. *Proceedings of the National Academy of Sciences* 108(24):9804-9808. doi:10.1073/pnas.1105379108.
25. Vetting MW, Roderick SL, Yu M, Blanchard JS (2003). Crystal structure of mycothiol synthase (Rv0819) from *Mycobacterium tuberculosis* shows structural homology to the GNAT family of N-acetyltransferases. *Protein Science* 12(9):1954-1959. doi:10.1110/ps.03153703.
26. Iqbal A, Arunlanantham H, Brown T, Chowdhury R, Clifton IJ, Kershaw NJ, Hewitson KS, McDonough MA, Schofield CJ (2009). Crystallographic and mass spectrometric analyses of a tandem GNAT protein from the clavulanic acid biosynthesis pathway. *Proteins: Structure, Function, and Bioinformatics* 78(6):1398-1407. doi:10.1002/prot.22653.
27. Trievel RC, Rojas JR, Sterner DE, Venkataramani RN, Wang L, Zhou J, Allis CD, Berger SL, Marmorstein R (1999). Crystal structure and mechanism of histone acetylation of the yeast GCN5 transcriptional coactivator. *Proceedings of the National Academy of Sciences* 96(16):8931-8936. doi:10.1073/pnas.96.16.8931.
28. Hickman AB, Nambodiri MAA, Klein DC, Dyda F (1999). The Structural Basis of Ordered Substrate Binding by Serotonin N-Acetyltransferase. *Cell* 97(3):361-369. doi:10.1016/s0092-8674(00)80745-x.
29. Wolf E, Vassilev A, Makino Y, Sali A, Nakatani Y, Burley SK (1998). Crystal Structure of a GCN5-Related N-acetyltransferase. *Cell* 94(4):439-449. doi:10.1016/s0092-8674(00)81585-8.

30. Olsen LR, Roderick SL (2001). Structure of the Escherichia coli GlmU Pyrophosphorylase and Acetyltransferase Active Sites,. *Biochemistry* 40(7):1913-1921. doi:10.1021/bi002503n.
31. Berman HM (2000). The Protein Data Bank. *Nucleic Acids Research* 28(1):235-242. doi:10.1093/nar/28.1.235.
32. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. (2022). RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research* 51(D1):D488-D508. doi:10.1093/nar/gkac1077.
33. The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Adesina A, et al. (2024). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* 53(D1):D609-D617. doi:10.1093/nar/gkae1010.
34. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium (2014). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926-932. doi:10.1093/bioinformatics/btu739.
35. Sayers EW, Cavanaugh M, Frisse L, Pruitt KD, Schneider VA, Underwood BA, Yankie L, Karsch-Mizrachi I (2024). GenBank 2025 update. *Nucleic Acids Research* 53(D1):D56-D61. doi:10.1093/nar/gkae1114.
36. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. (2020). CATH: increased structural coverage of functional space. *Nucleic Acids Research* 49(D1):D266-D273. doi:10.1093/nar/gkaa1079.
37. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. (2023). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* 52(D1):D368-D375. doi:10.1093/nar/gkad1011.
38. Kityk R, Kopp J, Sinning I, Mayer MP (2012). Structure and Dynamics of the ATP-Bound Open Conformation of Hsp70 Chaperones. *Mol.Cell* 48:863-. doi:10.1016/J.MOLCEL.2012.09.023 (PDB 4B9Q; record: RCSB primary citation).
39. Ghode P, Jobichen C, Ramachandran S, Bifani P, Sivaraman J (2015). Structural basis of mapping the spontaneous mutations with 5-fluorouracil in uracil phosphoribosyltransferase from Mycobacterium tuberculosis. *Biochem.Biophys.Res.Commun.* 467:577-582. doi:10.1016/j.bbrc.2015.09.133 (PDB 5E38; record: RCSB primary citation).

Figures

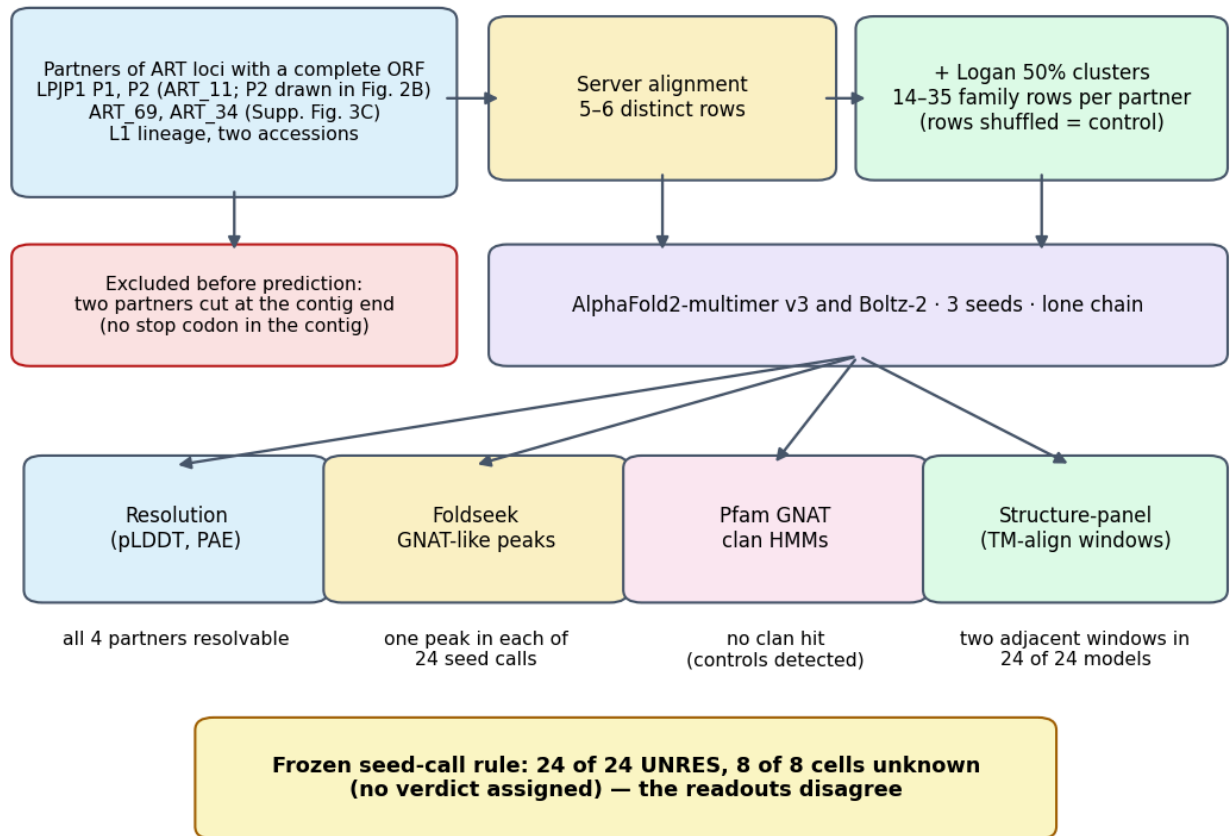


Figure 1. overview of the last stage and its readouts. Partners of ART loci with a complete reading frame are predicted from server alignments and from alignments enriched with Logan homologues (rows shuffled within each row as a control) by two predictors; four readouts (resolution, Foldseek peaks, Pfam clan models, structure-panel windows) are compared; the frozen seed-call rule returns UNRES in every seed call and all cell labels are unknown (no verdict is assigned). The numbers in the boxes are those of Section 3.9.

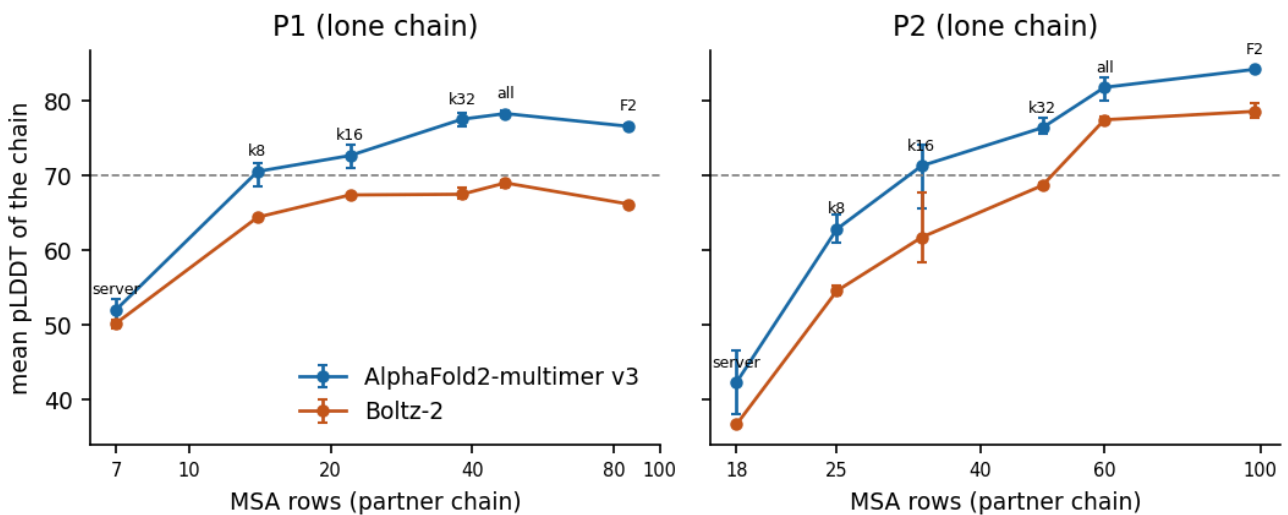


Figure 2. mean pLDDT of the whole lone P1 and P2 chains against the number of MSA rows in the partner alignment (server rows only; server plus $k = 8, 16, 32$ requested family rows, which with the server rows give the plotted row counts; server plus all family rows; expanded family), mean and range over three seeds; the dashed line at 70 is a visual aid (the resolution rule is a slot rule and for P1 excludes residues 484-658).

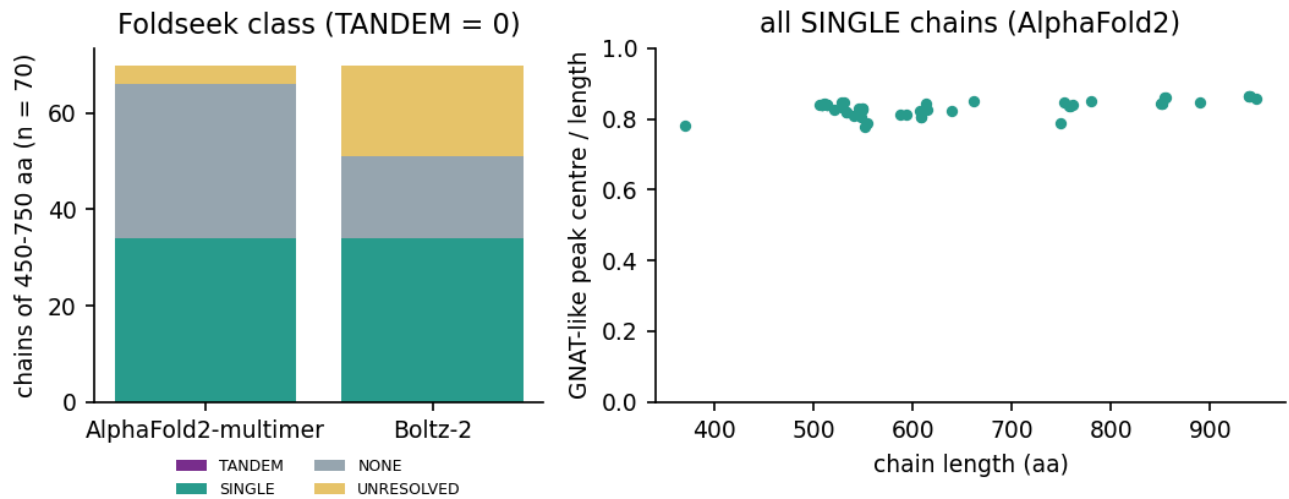


Figure 3. left, Foldseek classes of the chains of 450-750 aa per predictor (TANDEM: none); right, position of the single GNAT-like peak along the chain for all SINGLE chains (AlphaFold2).

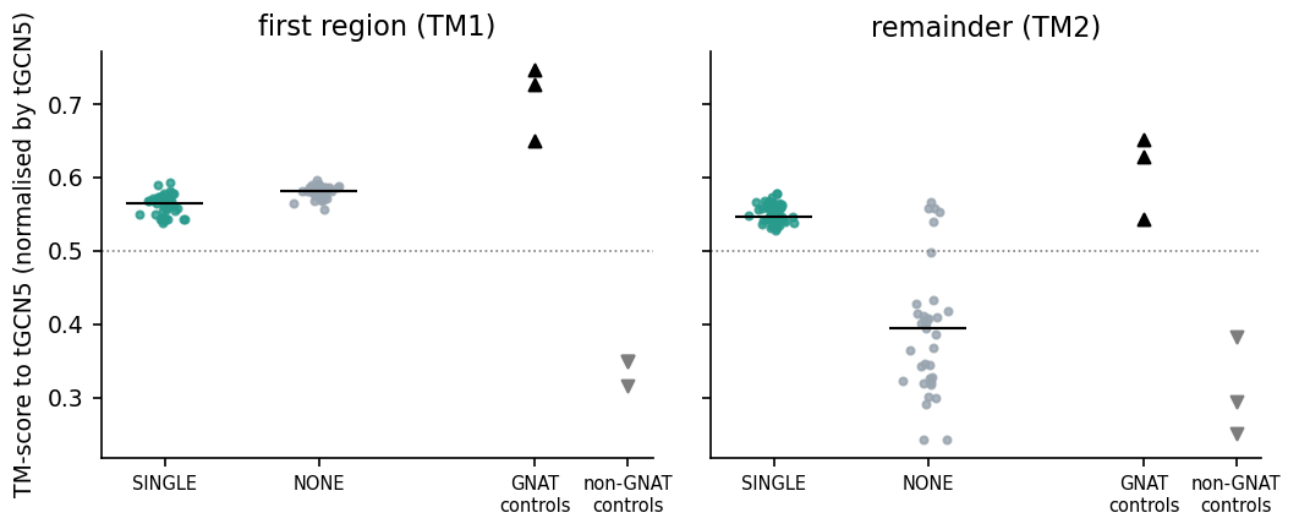


Figure 4. TM-score to tGCN5 of the first region and of the remainder in SINGLE and NONE chains (AlphaFold2, seed 0, all lengths), with control GNAT proteins (Eis, MshD, Orf14; RimI has no remainder) and non-GNAT controls (GlmU, UPP, DnaK); the dotted line at 0.5 is a visual aid.

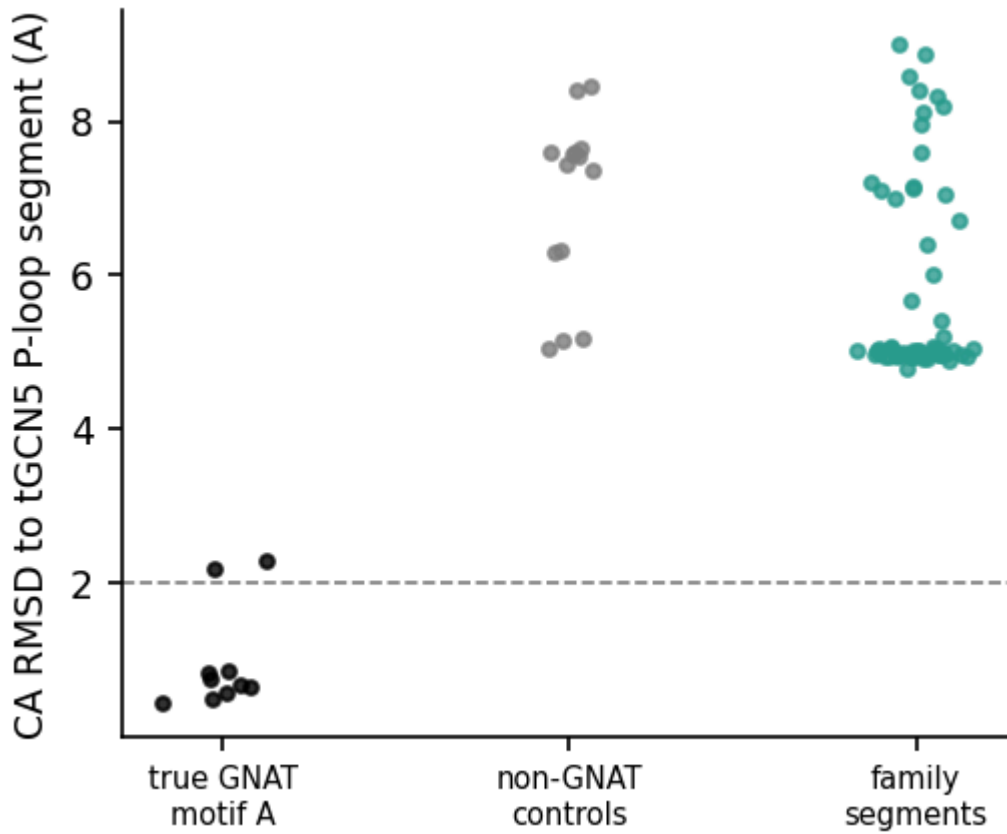


Figure 5. CA RMSD of motif segments to the tGCN5 P-loop segment: true GNAT motif A segments of the controls and R/Q-x-x-G segments of non-GNAT controls (loose pattern) against the family's R/Q-x-x-G-x-G/A segments (strict pattern); the dashed line at 2 Å is the registered cut-off.

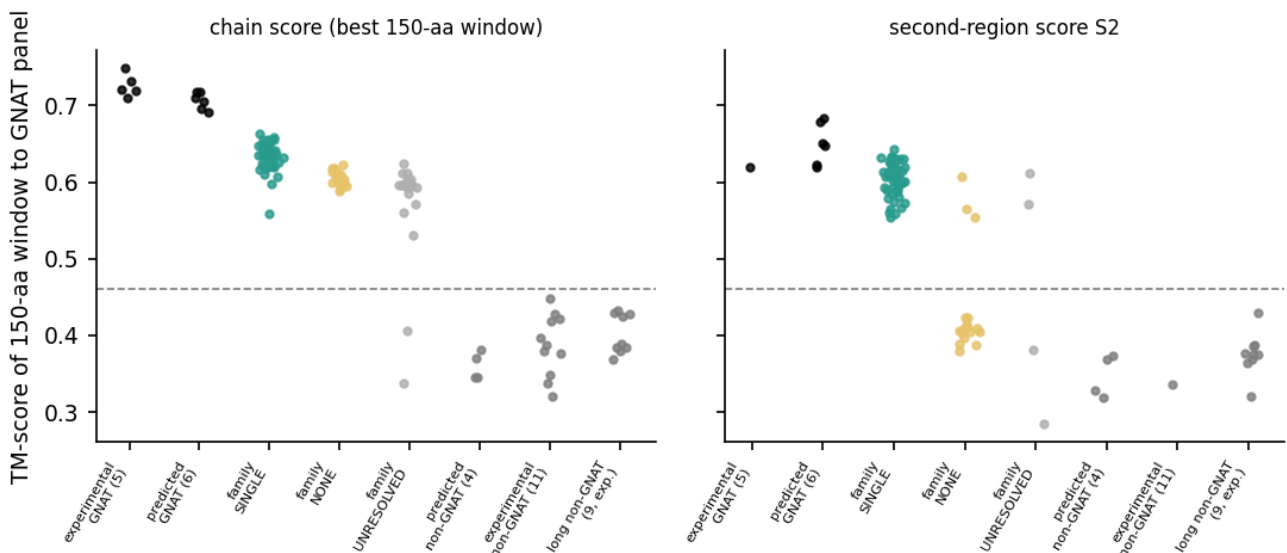


Figure 6. structure-panel scores (TM-score of a 150-residue window to the best of five GNAT references, normalised by the window; dashed line = threshold fixed from experimental references). Left, chain score (best window); right, second-region score S2 (best window at least 150 residues from the best window; absent where no second window exists or, for the short controls, none can exist). Groups: experimental GNAT references (leave-one-out), predicted GNAT controls (Eis, MshD, Orf14 at two depths), the family chains by Foldseek class (Boltz-2 seed-0 models), predicted non-GNAT controls (GlmU, DnaK at two depths), experimental non-GNAT references and nine long experimental non-GNAT structures (D11).

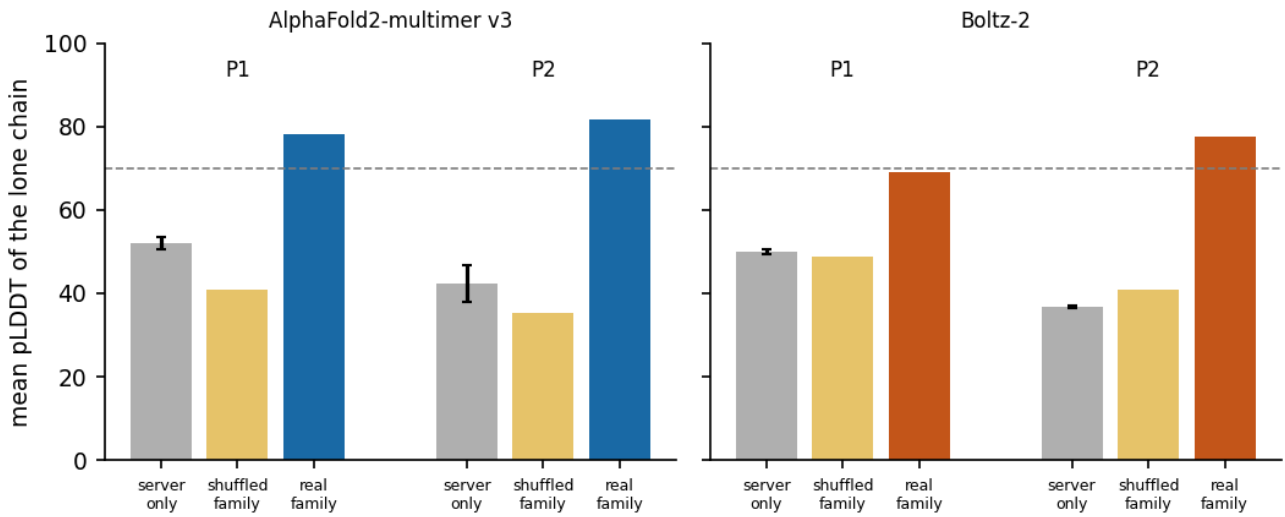


Figure 7. row-shuffled alignment control (D10): mean pLDDT of the lone chain with server rows only (bar: range over seeds), server rows plus family rows with residues permuted within each row, and server rows plus the real family rows (seed means, AlphaFold2 left and Boltz-2 right; dashed line at 70 as a visual aid).

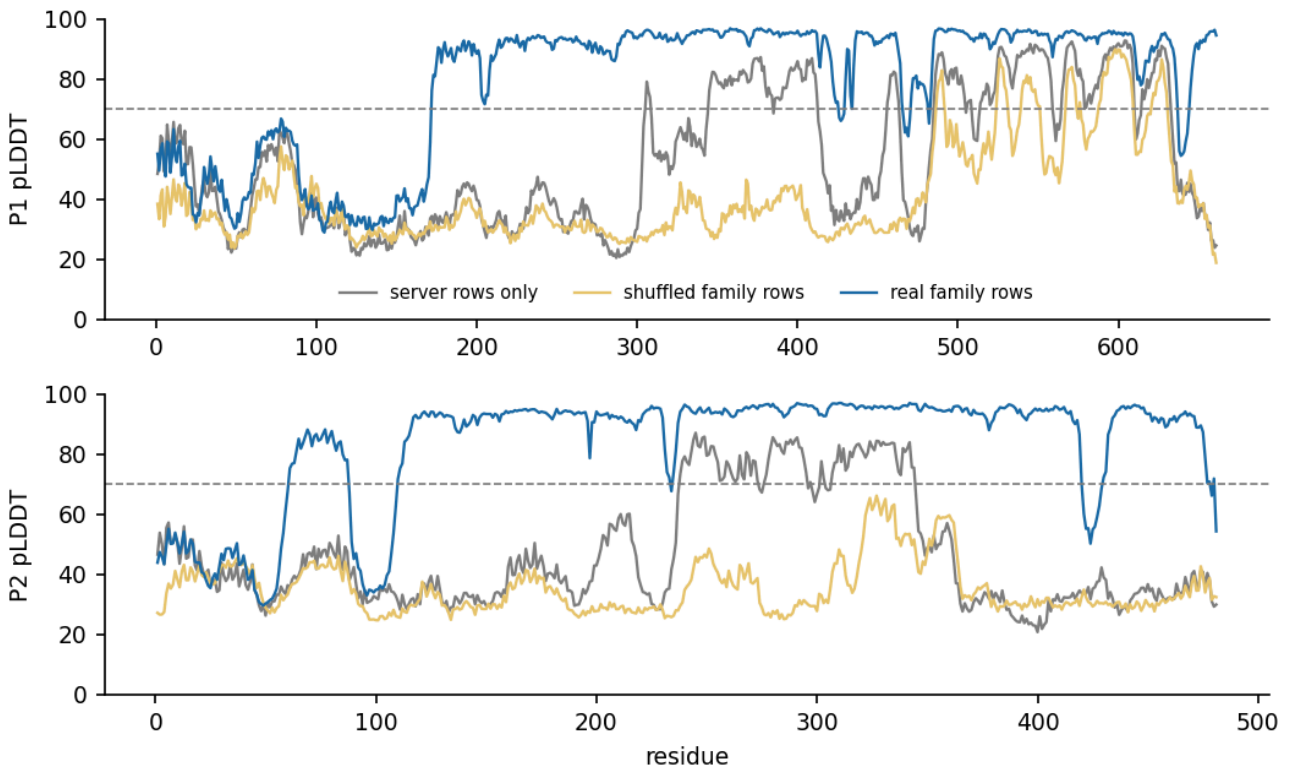


Figure 8. per-residue pLDDT of the lone P1 (top) and P2 (bottom) AlphaFold2 rank-1 model with server rows only, with server rows plus row-shuffled family rows (D10) and with server rows plus the real family rows (D5-B); one seed each (the first file of the harvest; the seed is not matched across curves); the dashed line at 70 is a visual aid. The N-terminal regions stay low with the real family rows and the shuffled rows do not raise the profile.

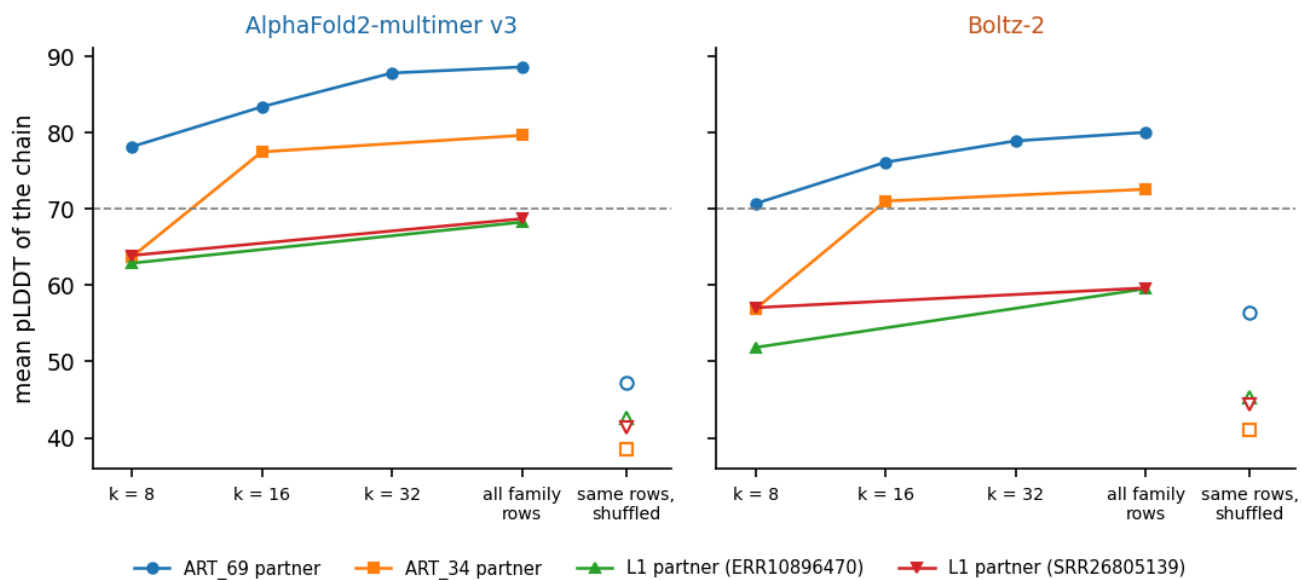


Figure 9. mean pLDDT of the lone chain of the four partners of Section 3.9 against the number of Logan family rows added to the server rows ($k = 8, 16, 32$ and all; a subset larger than the family was not built), and with all family rows shuffled within each row (open symbols); mean of the range over three seeds; AlphaFold2 left, Boltz-2 right; dashed line at 70 as a visual aid.

orange bar: Foldseek GNAT-like peak · green bar: span of the structure-panel windows above the threshold
 grey: AlphaFold2 pLDDT (seed 0, rank 1 model of the server + Logan alignment) · Pfam GNAT clan: no hit in any partner

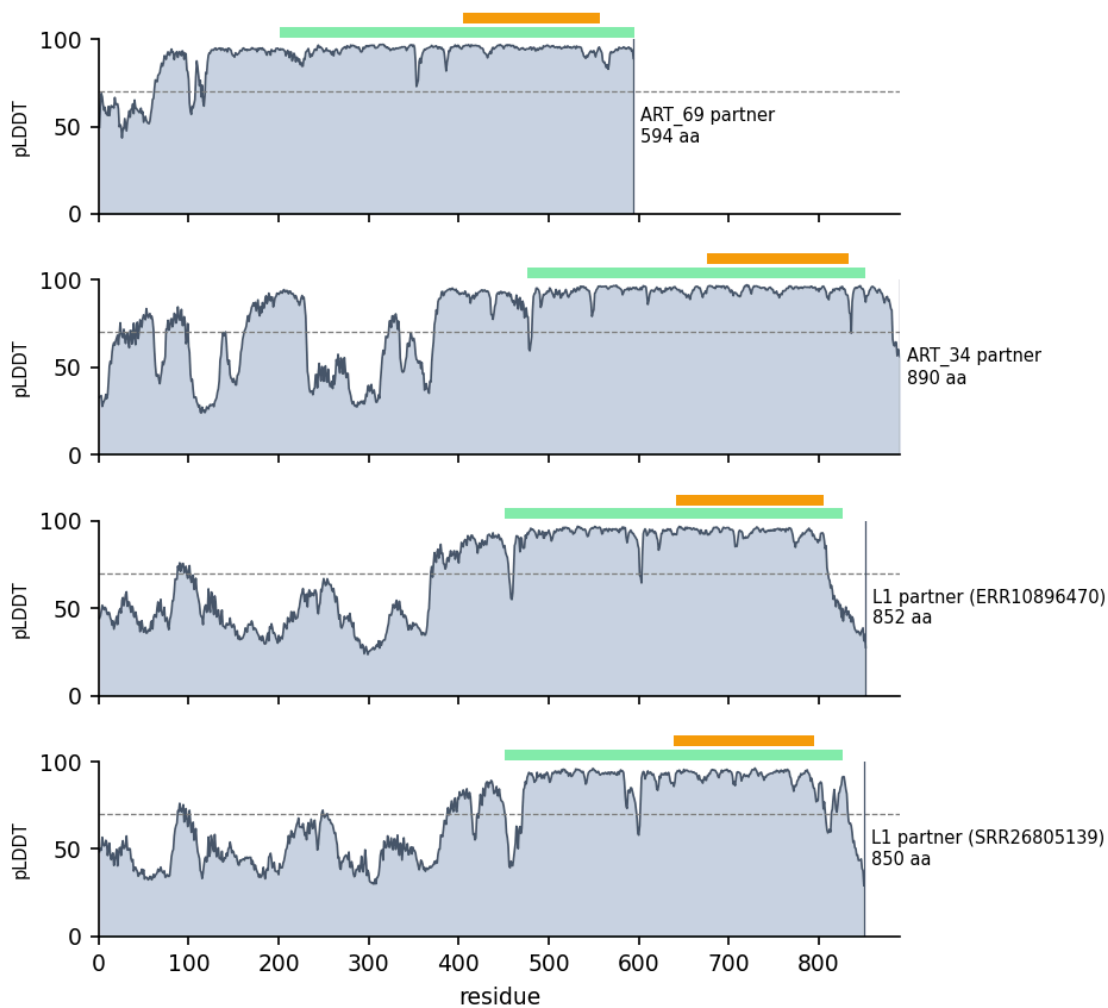


Figure 10. per-residue map of the readouts for the four partners: AlphaFold2 pLDDT of the rank-1 model of seed 0 (grey), Foldseek GNAT-like peak (orange bar), union of the structure-panel windows above the threshold (green bar); the Pfam GNAT clan models found no hit in any partner.