Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Structural confidence and template coverage of selected proteins from Flavobacterium azizsancarii

**Cem Boyraz**

23050288@stu.omu.edu.tr

Ondokuz Mayıs University

**Research Article**

**Keywords:** Flavobacterium azizsancarii, AlphaFold 3, structural bioinformatics, enzyme annotation, predicted aligned error, template coverage

**Posted Date:** September 15th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11019004/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

The Antarctic bacterium *Flavobacterium azizsancarii* provides a public genomic resource for investigating microbial enzyme diversity, but structural confidence does not establish the biochemical functions of individual proteins. This study examined 25 annotated proteins selected from the draft genome of strain AC, including ten candidates associated with phosphorus or carbohydrate metabolism. Coding sequences were verified against both the genome and corresponding RefSeq protein records. Five AlphaFold Server diffusion samples were analysed for each protein, giving 125 predicted structures. Residue confidence, predicted aligned error, coordinate agreement and archived template mappings were evaluated together. Median protein-level mean alpha-carbon confidence was 95.19, and 23 of 25 representative models had means of at least 90. Median predicted template modelling score was 0.94. The median global alpha-carbon root mean square deviation across 250 within-protein sample pairs was 0.54 Å, falling to 0.16 Å after independently aligning positions with confidence of at least 70 in both samples. A 1146-residue mutase-family protein contained confidently modelled cobalamin-binding, guanosine triphosphatase and mutase regions. Archived mappings to the known isobutyryl-coenzyme A mutase fusion structure 4XC6 covered 93.37% of its sequence at 60.65% identity, supporting a template-informed family assignment. A helicase and an FK506-binding peptidyl-prolyl isomerase showed greater regional uncertainty despite confident local segments. The analysis identifies tractable structural hypotheses without establishing new folds, substrate specificity or environmental performance.


# Introduction

Public bacterial genomes contain many annotated enzymes for which the corresponding proteins have not been characterised directly. Structure prediction can help identify which parts of these proteins are sufficiently resolved for comparison with established enzyme families. The biological value of a prediction depends on the distinction between a plausible fold, a supported family assignment and a demonstrated reaction. These levels of evidence are especially relevant when proteins are selected because their annotations suggest environmental or biotechnological applications.

*Flavobacterium azizsancarii* strain AC was described from Antarctic lake water collected on Ardley Island. Its original characterisation included genome analysis, psychrotolerance and hydrolytic phenotypes at the strain level (Otur et al. 2023a, b). These observations motivate the investigation of individual gene products, but they do not identify which protein accounts for a phenotype. The publicly available assembly also provides a basis for reproducible secondary analysis without resequencing. A recent preprint investigated intergenic small open reading frames in this strain using a genus-wide composition-matched null framework (C. Boyraz, Research Square preprint, 2026, https://doi.org/10.21203/rs.3.rs-10397597/v1). The present work addresses a different sequence class: annotated protein-coding products with verified translations and existing family assignments.

The genus provides a useful biological context for examining phosphorus and carbohydrate metabolism. Comparative analysis of terrestrial flavobacteria linked niche adaptation to plant glycan metabolism

(Kolton et al. 2013). Studies of plant-associated Bacteroidetes identified specialised organic-phosphorus acquisition strategies (Lidbury et al. 2021), and biochemical work established phosphate-insensitive phosphatase activity for PafA from *Flavobacterium johnsoniae* (Lidbury et al. 2022). Such findings explain why phosphatases and carbohydrate-active proteins merit investigation. They cannot establish that the Antarctic AC isolate is plant associated, that its homologues share identical reaction preferences, or that it improves plant growth.

AlphaFold 3 extends sequence-guided structure prediction to biomolecular interactions and provides confidence estimates alongside predicted coordinates (Abramson et al. 2024). Local predicted confidence and predicted aligned error describe complementary properties of a model. High local confidence can coexist with uncertainty in the relative placement of distant regions, while agreement among generated structures can reflect common sequence and template evidence (Jumper et al. 2021; Abramson et al. 2024). Consequently, a protein-wide average should be interpreted together with regional confidence, model-to-model coordinate variation and the templates actually supplied to the prediction system.

Here, we evaluate a rationale-selected panel of 25 AC proteins spanning phosphorus turnover, carbohydrate processing, lipid and pigment metabolism, and selected cofactor or protein-maintenance functions. The objective is to determine where the archived models support structural interpretation and where additional evidence remains necessary. Particular attention is given to a multidomain mutase-family protein and to proteins with discordant local and global confidence. Template coverage is quantified explicitly so that an existing structural precedent is not mistaken for an independently discovered architecture.

## Materials and methods

# Genome and annotation provenance

The analysis used RefSeq assembly GCF_027980145.1, assembly name ASM2798014v1, and its corresponding genomic FASTA and GenBank annotation files, retrieved on 12 September 2026. The assembly is a draft comprising 167 contigs and 5,836,236 bases. The downloaded annotation checksum was checked against the NCBI manifest, and all contig sequences were compared with the FASTA. Existing RefSeq annotations were analysed; a new genome assembly or an independent run of the Prokaryotic Genome Annotation Pipeline was not performed. RefSeq family-based product names were treated as computational annotations (Li et al. 2021).

The annotation contained 5017 gene features, 4948 coding-sequence features, 4897 protein-identifier-bearing coding sequences and 51 pseudogene coding-sequence records. Candidate selection excluded records marked as partial or pseudogenic. Each selected coding sequence was extracted in its annotated orientation and translated with its recorded genetic code, using coding-sequence validation that accounts for valid initiation codons. Every translation was compared with the annotated amino-acid

sequence and the corresponding retrieved RefSeq protein record. Sequence digests and accession-to-locus mappings were retained. Complete coding-sequence annotation was not interpreted as evidence of a closed genome or experimentally verified start sites.

## Candidate selection and domain evidence

The panel was chosen qualitatively before examining AlphaFold confidence. Ten positions were allocated to proteins with potential relevance to phosphorus or plant-polysaccharide processing: phytase, PafA, xylanase, xylosidase, mannosidase, glucosidase, amylase, galactosidase, arabinofuranosidase and rhamnogalacturonan acetylesterase candidates. Fifteen further positions sampled distinct annotated functions and domain architectures. Selection considered annotation specificity, conserved-region support, biochemical diversity and the interpretability of a structural question. The panel was not selected randomly, ranked by essentiality or designed to estimate the prevalence of a property across the complete proteome.

Conserved Domain Database (CDD) region annotations were obtained from the retrieved NCBI protein records (Wang et al. 2023). Their saved coordinates were used without claiming a newly executed domain search. Overlapping footprints were retained as annotation intervals rather than forced into an exclusive structural partition. Genomic neighbourhood information helped document candidate context, but no pathway-completeness or operon analysis was used to establish function. Full sequences, selection rationales, genomic coordinates and CDD identifiers are supplied in Online Resource 2.

## Structure prediction and archived inputs

Each protein was submitted separately to AlphaFold Server on 12 September 2026, using one copy of the full amino-acid sequence and structure-template use enabled. No sequence truncation, ligand, metal, cofactor, nucleic acid or additional protein chain was supplied. Each of the 25 jobs contained one recorded random seed and returned five diffusion samples, as described for the server workflow (EMBL-EBI n.d.). These 125 structures are prediction samples, not independent experiments or biological replicates.

All original output archives were preserved, including coordinates, summary and full confidence files, input requests, multiple sequence alignments (MSAs), template files, mappings and terms of use. Exact seeds and the software strings embedded in the model files are recorded in the supplementary provenance table. The value 3 in the request's version field identifies the input schema and was not used as a complete software-build identifier. Saved ranking scores were used to identify the highest-ranked representative; where scores were tied at the available precision, the lowest model index was retained. This reproduced model 0 as the representative in all 25 jobs.

## Confidence and coordinate agreement

Predicted local distance difference test confidence, abbreviated pLDDT, was read from alpha-carbon (Cα) atom entries in the mmCIF B-factor field. This field contains prediction confidence, not an experimental displacement parameter. Protein means reported here are arithmetic averages across one Cα value per residue. They differ from the server's all-atom global average because atom-rich residues receive no additional weight. Residues were grouped into confidence bins of below 50, 50 to below 70, 70 to below 90, and at least 90.

The predicted template modelling score (pTM) and the server clash flag were read from each summary file. Interface pTM was null for the single-chain jobs and was not analysed. Predicted aligned error (PAE) was read from the full token-pair matrix, retaining its direction: entry i,j describes uncertainty in the position of residue j when the structure is aligned using the frame of residue i. Whole-matrix means included diagonal entries. Confidence-restricted means used only pairs for which both residues had Cα pLDDT of at least 70 in the representative model. Domain-level summaries used saved CDD intervals; symmetric between-domain PAE averaged both directional blocks.

Corresponding Cα coordinates were superposed by least-squares rotation with reflection exclusion (Kabsch 1976). All ten unique pairs among the five samples were analysed for each protein, yielding 250 comparisons. Global root mean square deviation (RMSD) used all positions. A second RMSD calculation independently refitted only positions with pLDDT of at least 70 in both members of a pair. Consequently, the restricted RMSD measures agreement in the selected subset and is not a coordinate error against experiment. Summaries are descriptive; no hypothesis tests treated residue pairs or diffusion samples as independent biological observations.

## Alignment and template analysis

MSA records were counted after removing lowercase insertion characters and alignment-format punctuation. Query records were included in reported sequence counts. Unique normalised sequences and overlaps between the server-labelled paired and unpaired files were calculated to avoid interpreting their sum as independent homologous support. These counts are not an effective sequence number and do not estimate phylogenetic independence. For single-chain jobs, the label paired does not demonstrate a biological interaction partner.

Archived query-to-template index mappings were matched to template residue identities. Template identity was calculated as identical amino acids divided by evaluable mapped positions, with coverage expressed relative to full query length. File count and distinct PDB-entry count were reported separately. When multiple mappings referred to one PDB entry, their query coverage was combined as a union. This calculation describes the server's selected templates; it is not an exhaustive similarity search against all known protein structures. No template-ablation, independent structure-prediction benchmark or experimentally determined AC protein structure was available.

# Reproducibility and assistance

Sequence and structure parsing used Python 3.12 and Biopython 1.88 (Cock et al. 2009); numerical analysis used NumPy 2.5.3. Publication plots were generated from numeric output using Matplotlib 3.11.2. An independent audit reread the archived files, retranslated the selected coding sequences and recomputed the principal confidence and coordinate statistics. The scripts, input checksums, full output archives and machine-readable tables are included in the accompanying resources.

OpenAI Codex assisted candidate triage, code generation and execution, output inspection, literature retrieval and manuscript drafting. This assistance included substantive content generation and was not limited to language correction. The numerical results were computed from the archived data. The author is responsible for scientific interpretation, verification of cited evidence and approval of the final manuscript.

## Results

# A verified panel with broadly high local confidence

All 25 selected coding sequences matched their genomic translations and retrieved protein records. The proteins spanned 154–1146 residues, with a median length of 387 and 11,096 residues in total. The ten phosphorus/carbohydrate candidates and fifteen additional candidates are listed in Table 1, preserving the distinction between source annotation and inferred research interest. All five samples per target contained the expected residue sequence, and archive checksums matched the retained manifest.

Across representative models, median mean Cα pLDDT was 95.19, with a range of 87.83–97.63. Twenty-three proteins had a mean of at least 90, and all 25 had a mean above 80. Median pTM was 0.94, ranging from 0.80 to 0.97. Pooling residues across the representative models, 90.50% had pLDDT of at least 90, 96.40% at least 70, and 1.91% below 50. Figure 1 shows the heterogeneity obscured by these averages. These percentages describe the selected panel and cannot be extrapolated to the full AC proteome.

The glycoside hydrolase family 1 (GH1) glucosidase had the largest protein mean, 97.63, followed by homogentisate dioxygenase at 97.45 and the type III polyketide synthase at 97.18. The helicase and FK506-binding protein (FKBP)-type isomerase had the two lowest means, 87.83 and 88.45. No sample carried the server's substantial-clash flag. This absence is a limited server diagnostic and does not replace a stereochemical validation against an experimental reference.

Table 1

Selected proteins and representative-model confidence. Numbers 1–10 denote the phosphorus/carbohydrate subgroup; 11–25 are the additional candidates. Short functional descriptions summarise the annotations and do not imply experimental characterisation. Locus and domain metadata are in Online Resource 1; genomic coordinates are in Online Resource 2

| No | RefSeq protein | Annotation summary | aa | Mean Cα pLDDT | pTM |
|---|---|---|---|---|---|
| 01 | WP_271336311.1 | Phytase | 352 | 93.88 | 0.91 |
| 02 | WP_271334492.1 | Alkaline phosphatase PafA | 539 | 95.92 | 0.95 |
| 03 | WP_271336476.1 | Endo-1,4-beta-xylanase | 372 | 92.97 | 0.90 |
| 04 | WP_271336457.1 | Xylan 1,4-beta-xylosidase | 864 | 95.19 | 0.91 |
| 05 | WP_271334993.1 | Beta-mannosidase | 662 | 95.76 | 0.94 |
| 06 | WP_271338273.1 | GH1-family beta-glucosidase | 450 | 97.63 | 0.96 |
| 07 | WP_271334890.1 | Alpha-amylase | 480 | 92.83 | 0.91 |
| 08 | WP_271337138.1 | Alpha-galactosidase | 734 | 96.08 | 0.94 |
| 09 | WP_271335125.1 | Alpha-N-arabinofuranosidase | 515 | 95.94 | 0.94 |
| 10 | WP_271336193.1 | Rhamnogalacturonan acetylesterase | 247 | 95.49 | 0.92 |
| 11 | WP_271335871.1 | Mutase-family fusion candidate | 1146 | 96.75 | 0.97 |
| 12 | WP_271334236.1 | Type III polyketide synthase | 350 | 97.18 | 0.96 |
| 13 | WP_271335165.1 | Phytoene desaturase family | 488 | 94.66 | 0.95 |
| 14 | WP_271335166.1 | Phytoene/squalene synthase family | 279 | 91.79 | 0.93 |
| 15 | WP_271334113.1 | Lycopene cyclase family | 398 | 94.57 | 0.94 |
| 16 | WP_271336501.1 | Mpo1-family dioxygenase | 154 | 92.02 | 0.90 |
| 17 | WP_271336700.1 | Acyl-ACP desaturase | 330 | 96.24 | 0.95 |
| 18 | WP_271335889.1 | Cryptochrome/photolyase family | 467 | 96.61 | 0.95 |
| 19 | WP_271337168.1 | Acyl-ACP thioesterase | 248 | 95.98 | 0.94 |
| 20 | WP_271334241.1 | Lipase family | 367 | 93.86 | 0.92 |
| 21 | WP_271337214.1 | 4-Hydroxyphenylpyruvate dioxygenase | 386 | 93.30 | 0.95 |
| 22 | WP_271337215.1 | Homogentisate 1,2-dioxygenase | 387 | 97.45 | 0.96 |
| 23 | WP_271337586.1 | 4,5-DOPA-extradiol dioxygenase | 276 | 95.15 | 0.94 |
| 24 | WP_271335540.1 | DEAD/DEAH-box helicase | 450 | 87.83 | 0.80 |

| No | RefSeq protein | Annotation summary | aa | Mean Cα pLDDT | pTM |
|---|---|---|---|---|---|
| 25 | WP_271334413.1 | FKBP-type isomerase | 155 | 88.45 | 0.82 |

# Sample agreement and regional uncertainty

The median global Cα RMSD across the 250 within-protein sample pairs was 0.54 Å, with a range of 0.10–6.85 Å. When each pair was independently aligned using shared positions with pLDDT of at least 70, the median decreased to 0.16 Å, with a range of 0.06–5.05 Å. This difference indicates that uncertain positions and their influence on global alignment account for much of the observed coordinate dispersion. It does not show that the remaining coordinates have sub-ångström experimental accuracy.

The median of the per-protein whole-matrix mean PAE values was 3.45 Å, with individual means ranging from 2.37 to 10.32 Å. For the phytase candidate, restricting the calculation to high-confidence residue pairs changed mean PAE from 5.01 to 2.26 Å. The corresponding values were 7.51 and 1.93 Å for the FKBP-type isomerase. The helicase retained a larger restricted mean of 5.03 Å, compared with 10.32 Å for the complete chain. Table 2 and Fig. 2 illustrate how a high-confidence core can be accompanied by uncertain extensions or more extensive interregional ambiguity.

The FKBP-family CDD footprint at residues 55–153 had mean pLDDT 98.40 and within-region PAE 1.62 Å, despite the lower whole-protein scores. The broad helicase-associated SrmB footprint at residues 1–428 had mean pLDDT 88.87 and within-region PAE 8.22 Å. Its annotation did not provide justified boundaries for two separately quantified domains, so such a partition was not imposed. These two examples therefore support different levels of interpretability: a sharply resolved annotated core in the isomerase and broader positional uncertainty in the helicase.

Table 2
Selected examples of complete-chain and confidence-restricted PAE. Restricted means use only pairs of residues with Cα pLDDT at least 70 in the representative model. Maximum RMSD is to the representative using all Cα positions. Each row describes one protein rather than independent replicate measurements

| No | Protein | Mean PAE Å | Restricted PAE Å | Residues ≥ 70% | Max RMSD Å |
|---|---|---|---|---|---|
| 01 | Phytase | 5.01 | 2.26 | 93.75 | 3.32 |
| 11 | Mutase-family fusion candidate | 3.27 | 3.08 | 99.30 | 0.55 |
| 24 | DEAD/DEAH-box helicase | 10.32 | 5.03 | 87.11 | 6.61 |
| 25 | FKBP-type isomerase | 7.51 | 1.93 | 84.52 | 4.34 |

# Domain evidence for the phosphorus and carbohydrate candidates

All ten candidates in the phosphorus/carbohydrate subgroup had representative protein means between 92.83 and 97.63 and pTM values between 0.90 and 0.96. The phytase Phy footprint, residues 25–351, had mean pLDDT 97.80 and mean within-region PAE 2.16 Å. The PafA-associated AP-SPAP region, residues 21–533, had corresponding values of 98.15 and 2.29 Å. These observations locate confident family-associated regions in both phosphatase candidates; they do not establish phosphate release rates or substrate selectivity.

The 864-residue xylosidase candidate contained several annotated footprints, including BglX, Glyco_hydro_3_C, PA14 and an Fn3-like interval. Their relationships were retained as overlapping annotation evidence. The broad Glyco_hydro_3_C interval had mean pLDDT 94.45 but mean within-interval PAE 5.63 Å, compared with 2.28 Å for BglX and 1.78 Å for the Fn3-like interval. Thus, the high protein-wide confidence does not imply uniform certainty over every annotated region. The galactosidase candidate likewise showed high confidence across three saved glycosidase-associated intervals, while the analysis did not establish oligomerisation or the contribution of each region to substrate range. All annotation-interval statistics are provided in Online Resource 1.

# A mutase-family protein with a known structural precedent

The 1146-residue protein NJT12_RS10585, annotated as a methylmalonyl-CoA mutase family protein, had mean Cα pLDDT 96.75, pTM 0.97 and mean PAE 3.27 Å. Three non-overlapping CDD footprints described a cobalamin-binding region at residues 14–137, an ArgK-family GTPase region at 163–461, and a mutase-associated region at 521–1069. Mean local confidence within these regions was 97.90, 96.46 and 97.56, respectively. Their within-region PAE means were 1.35, 2.44 and 2.31 Å.

Symmetric between-region PAE was 2.69 Å for the cobalamin-binding/GTPase pair, 2.71 Å for the cobalamin-binding/mutase pair and 3.68 Å for the GTPase/mutase pair. A lower-confidence interval at residues 487–493 lay between the latter two footprints. Figure 3 presents the regional confidence and PAE measurements. These values support an internally consistent prediction for the supplied full-length chain, while leaving the biological role of flexible connecting segments unresolved.

Template analysis substantially qualifies the interpretation of this architecture. The four archived template files represented only two distinct PDB entries, 4XC6 and 2XIJ. The two non-overlapping query mappings to 4XC6 covered 1070 of 1146 residues, or 93.37%, at 60.65% identity over mapped positions. PDB 4XC6 is an experimentally determined IcmF structure (Jost et al. 2015). The combined evidence is therefore compatible with an IcmF-like family architecture, but does not constitute discovery of a new fusion or a template-independent fold. The exact acyl-CoA substrate of the AC protein remains unresolved.

# Template and alignment support across the complete panel

Every job contained four archived template files, giving 100 files from 90 distinct PDB entries across the cohort. Multiple files occasionally represented different mapped regions of the same entry. Across

proteins, the median maximum identity to an individual template mapping was 37.25%, ranging from 22.64 to 74.02%. The median maximum individual-mapping coverage was 92.25%, ranging from 36.51 to 98.86%. Maximum identity and maximum coverage need not belong to the same mapping and were not combined into a synthetic similarity score.

Unpaired MSA files contained 1256–19,559 records including the query, with a median of 15,498. Substantial overlap with the server-labelled paired files meant that a sum of raw file counts would overstate distinct sequence support. The large counts provide a record of available alignment material but do not measure effective evolutionary diversity. Several proteins with confident models had partial template coverage, including the lipase-family candidate, whose largest individual template mapping covered 36.51% of the query. Confidence outside a mapped segment may reflect additional sequence or learned structural information, and requires separate examination rather than an assumption of complete template coverage.

## Discussion

This analysis produces an accession-linked account of which regions in a defined AC protein panel can be interpreted structurally. Its main finding is the coexistence of broadly high local confidence with meaningful differences in regional uncertainty and template coverage. A list ordered only by mean pLDDT would conceal the confident FKBP-family core, the more extensive ambiguity in the helicase and the strong existing structural precedent for the mutase-family fusion. The combined analysis therefore changes the interpretation of the predictions without assigning unmeasured biological activities.

The phosphorus and carbohydrate subgroup offers tractable starting points for biochemical study. PafA is particularly interpretable because the source annotation and the saved conserved-region footprint describe a family with experimentally established phosphorus metabolism in other flavobacteria (Lidbury et al. 2021, 2022). The AC model resolves most of the associated region at high local confidence. This supports comparison of family architecture and provides a basis for subsequent residue-level analysis, but phosphate insensitivity and the substrate range measured for *F. johnsoniae* cannot be transferred automatically. The phytase and glycan-processing candidates are similarly plausible annotations whose application value depends on direct biochemical evidence.

The carbohydrate candidates also illustrate why annotation intervals and structural domains should not be treated as interchangeable. In the xylosidase, overlapping CDD footprints capture family relationships at different scales. A broad interval can encompass subregions whose relative positions are less certain than their local geometry. Preserving the original footprints makes that limitation visible. An exclusive domain decomposition inferred only from annotation names would introduce boundaries that were not established by the present analysis. Functional conclusions about substrate channels or domain cooperation would require an additional structural comparison with experimentally characterised homologues.

The mutase-family protein provides the clearest example of the importance of template provenance. Its cobalamin-binding, GTPase and mutase regions are jointly compatible with known IcmF biology. Structural studies have already established a fusion between a radical B12 enzyme and its G-protein chaperone (Jost et al. 2015), and later work investigated chaperone-mediated cofactor maturation in a related characterised system (Vaccaro et al. 2023). The extensive 4XC6 mapping means that the AC prediction should be interpreted against this existing framework. The unresolved questions concern protein-specific biochemistry and state-dependent organisation, rather than whether such a domain combination can exist.

The monomer-only design imposes an additional constraint on that interpretation. The prediction request did not include cobalamin, nucleotide, metal ions or a second chain. A coherent model can therefore describe a plausible polypeptide arrangement without establishing the cofactor-loaded catalytic state, biological oligomer or sequence of conformational transitions. The five samples from one seed cannot substitute for a dynamics experiment. Their close agreement is useful for identifying stable output features, but does not establish that the protein occupies a single conformation in solution.

For the helicase and FKBP-type isomerase, lower whole-chain confidence has different practical implications. The isomerase contains a very confident CDD footprint, suggesting that uncertainty is strongly associated with the remainder of the chain. The helicase retains appreciable PAE even after restricting to locally confident residues. Both cases caution against selecting or discarding candidates solely by a protein mean. Low-confidence regions may reflect flexibility, incomplete biological context or modelling limitations; they are not automatically dispensable sequence. No deletions, mutational effects or construct performance were predicted here.

Several limitations delimit the contribution. The 25 proteins were selected for specific biological questions and are not an unbiased sample of the proteome. Their functions mostly originate in homology-based annotations, and CDD agreement may reuse related sequence evidence. Template use was enabled throughout, and the saved mappings do not exhaust the contents of structural databases. No enzyme assay, expression measurement, independent geometry validation, template-ablation analysis or alternative prediction system was included. Accordingly, this work cannot establish novel reactions, cold activity of individual enzymes, essentiality, plant-growth effects or new folds. The evidence supports a reproducible structural survey and a set of specific questions for further study.

# Conclusions

The verified AC panel contains many confidently modelled protein regions, but local confidence, relative-position confidence and template support lead to distinct interpretations. The phosphorus/carbohydrate candidates provide annotated enzyme regions suitable for more detailed comparison, while the FKBP-type isomerase and helicase exemplify different patterns of uncertainty. The multidomain mutase-family protein is consistent with a known IcmF-like architecture whose structural precedent is directly represented among the input templates. Preserving these distinctions makes the models useful as

traceable hypotheses and avoids assigning biochemical or ecological properties that were not measured.

## Declarations

# Competing interests

The author declares no competing interests.

# Ethics declarations

This study reanalysed public microbial sequence data and computational protein predictions. It did not involve human participants, identifiable human research data, animal experiments or newly collected biological specimens.

# Funding

This research received no external funding or material support.

# Author Contribution

C.B. conceived and designed the study, defined its biological scope, performed and interpreted the computational analyses, developed the analysis code, prepared all figures and tables, and wrote and reviewed the manuscript.

# Data Availability

The source genome and annotation are publicly available from NCBI under GCF_027980145.1 and WGS project JAMZNK01. Online Resource 1 contains the analysed protein, model, coordinate-comparison, annotation-region and template tables. Online Resource 2 contains the 25 original AlphaFold output archives, verified sequences, provenance records and scripts used for this analysis. These resources accompany this manuscript package; no new public repository accession has been assigned to the derived data. The original AlphaFold archive terms are preserved with the prediction outputs.

## References

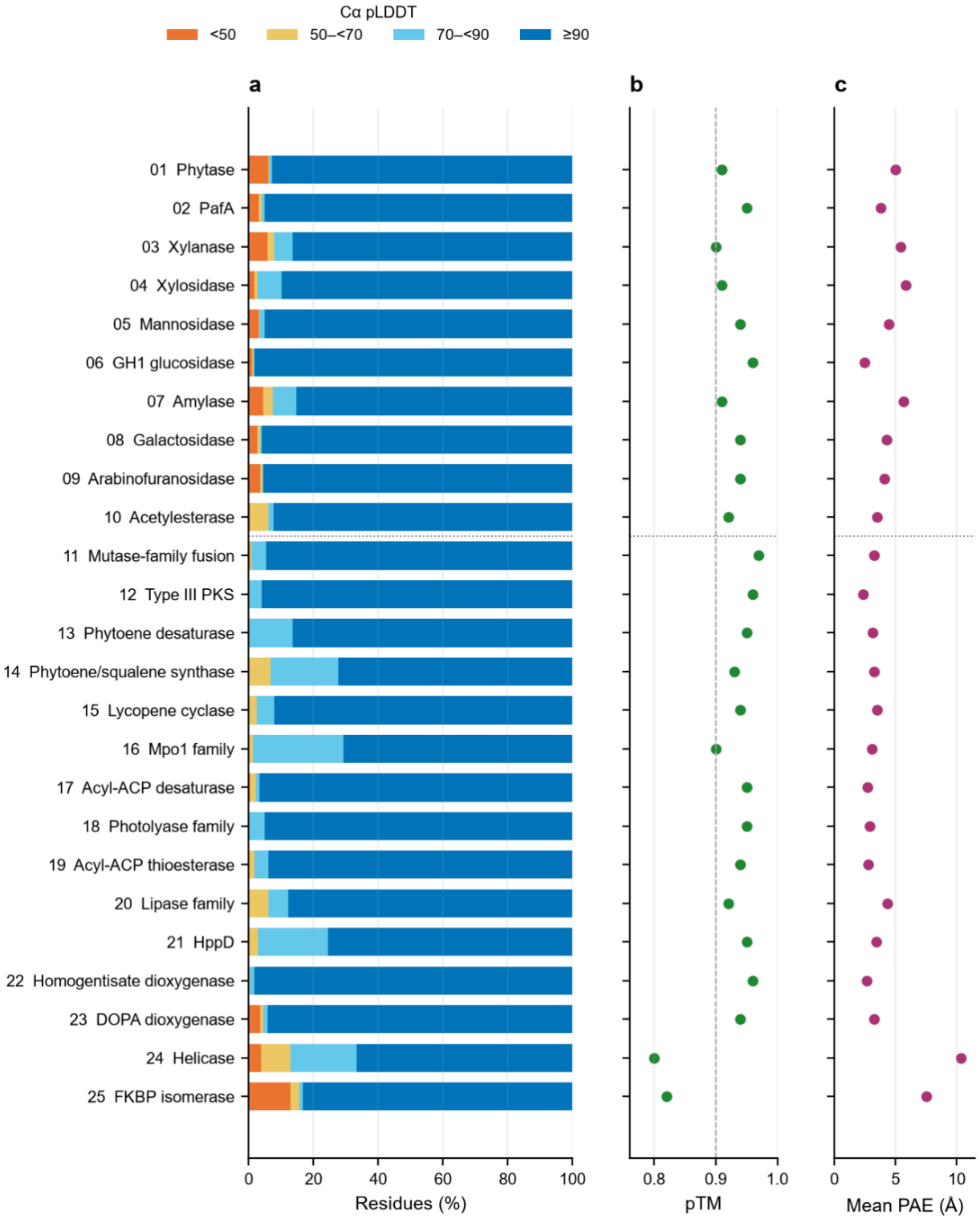
1. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493–500.

https://doi.org/10.1038/s41586-024-07487-w

2. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163
3. EMBL-EBI (n d.) Interpreting results from AlphaFold Server. AlphaFold online training. https://www.ebi.ac.uk/training/online/courses/alphafold/alphafold-3-and-alphafold-server/alphafold-server-your-gateway-to-alphafold-3/interpreting-results-from-alphafold-server/. Accessed 12 September 2026
4. Jost M, Cracan V, Hubbard PA, Banerjee R, Drennan CL (2015) Visualization of a radical B12 enzyme with its G-protein chaperone. Proceedings of the National Academy of Sciences USA 112:2419–2424. https://doi.org/10.1073/pnas.1419582112
5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
6. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallogr Sect A 32:922–923. https://doi.org/10.1107/S0567739476001873
7. Kolton M, Sela N, Elad Y, Cytryn E (2013) Comparative genomic analysis indicates that niche adaptation of terrestrial Flavobacteria is strongly linked to plant glycan metabolism. PLoS ONE 8:e76704. https://doi.org/10.1371/journal.pone.0076704
8. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A et al (2021) RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105
9. Lidbury IDEA, Borsetto C, Murphy ARJ, Bottrill A, Jones AME, Bending GD, Hammond JP, Chen Y, Wellington EMH, Scanlan DJ (2021) Niche-adaptation in plant-associated Bacteroidetes favours specialisation in organic phosphorus mineralisation. ISME J 15:1040–1055. https://doi.org/10.1038/s41396-020-00829-2
10. Lidbury IDEA, Scanlan DJ, Murphy ARJ, Christie-Oleza JA, Aguilo-Ferretjans MM, Hitchcock A, Daniell TJ (2022) A widely distributed phosphate-insensitive phosphatase presents a route for rapid organophosphorus remineralization in the biosphere. Proceedings of the National Academy of Sciences USA 119:e2118122119. https://doi.org/10.1073/pnas.2118122119
11. Otur Ç, Okay S, Kurt-Kızıldoğan A (2023a) Whole genome analysis of *Flavobacterium aziz-sancarii* sp. nov., isolated from Ardley Island (Antarctica), revealed a rich resistome and bioremediation potential. Chemosphere 313:137511. https://doi.org/10.1016/j.chemosphere.2022.137511
12. Otur Ç, Okay S, Kurt-Kızıldoğan A (2023b) Corrigendum to Whole genome analysis of *Flavobacterium aziz-sancarii* sp. nov., isolated from Ardley Island (Antarctica), revealed a rich resistome and bioremediation potential [Chemosphere 313 (2023) 1–11/137511]. Chemosphere 324:138379. https://doi.org/10.1016/j.chemosphere.2023.138379

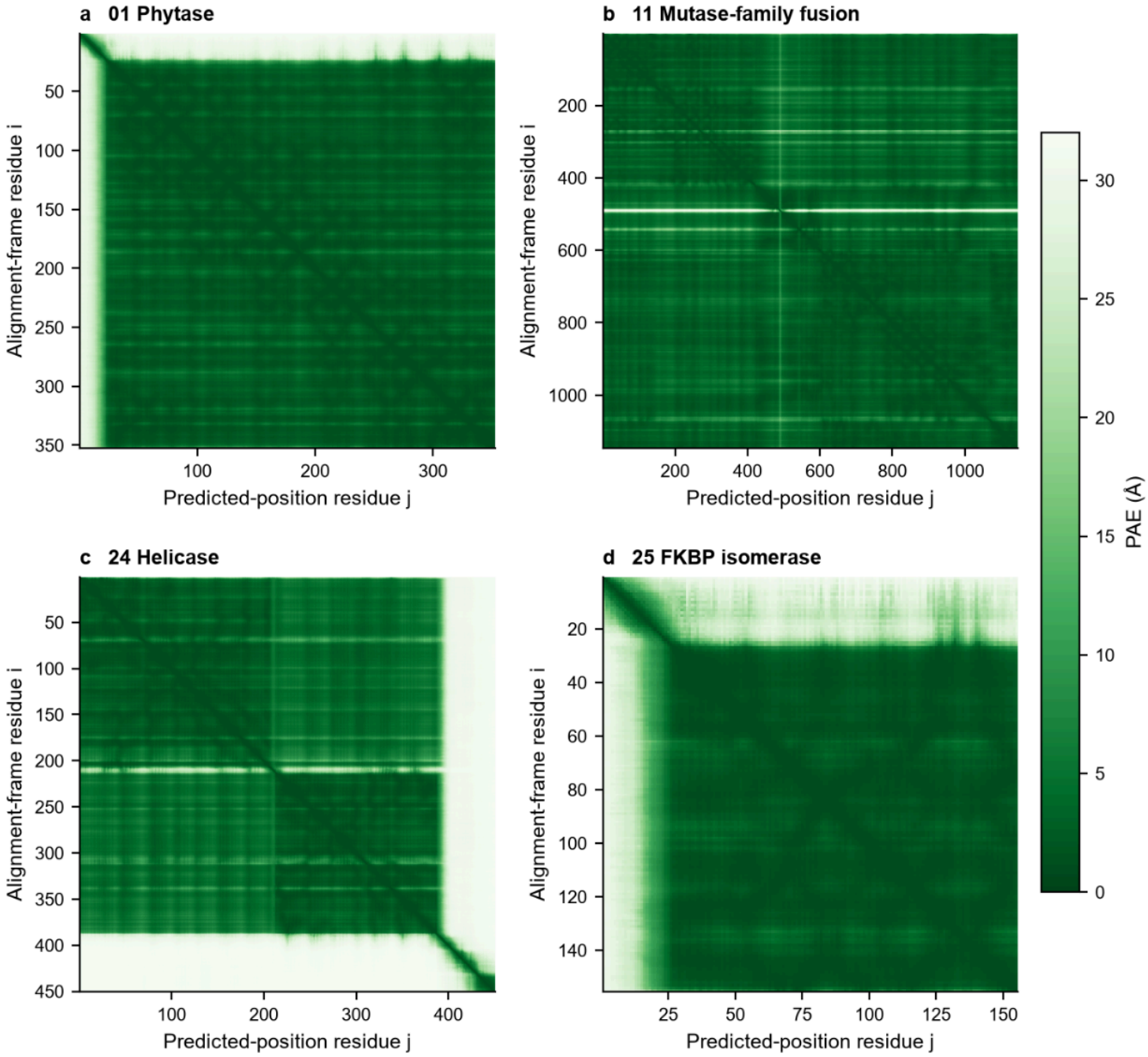
13. Vaccaro FA, Faber DA, Andree GA, Born DA, Kang G, Fonseca DR, Jost M, Drennan CL (2023) Structural insight into G-protein chaperone-mediated maturation of a bacterial adenosylcobalamin-dependent mutase. J Biol Chem 299:105109. https://doi.org/10.1016/j.jbc.2023.105109
14. Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A (2023) The conserved domain database in 2023. Nucleic Acids Res 51:D384–D388. https://doi.org/10.1093/nar/gkac1096

# Figures




## Figure 1

Representative-model confidence across the 25-protein panel. a Proportions of residues in the four Cα pLDDT bins. b Whole-chain pTM. c Mean PAE over the complete directional residue-pair matrix, including its diagonal. Numbers correspond to Table 1; the dotted line separates the ten phosphorus/carbohydrate candidates from the fifteen additional candidates. The representative is model 0, selected by archived rank with the lowest-index tie rule




## Figure 2

Regional positional uncertainty in four representative models. a Phytase, b mutase-family fusion, c helicase and d FKBP-type isomerase. In each matrix, row i specifies the alignment frame and column j the residue whose position is evaluated. Axes use one-based residue numbering and a common 0–32 Å scale. Dark green denotes low PAE and pale regions higher PAE. The maps quantify predicted uncertainty and do not establish experimentally observed motion

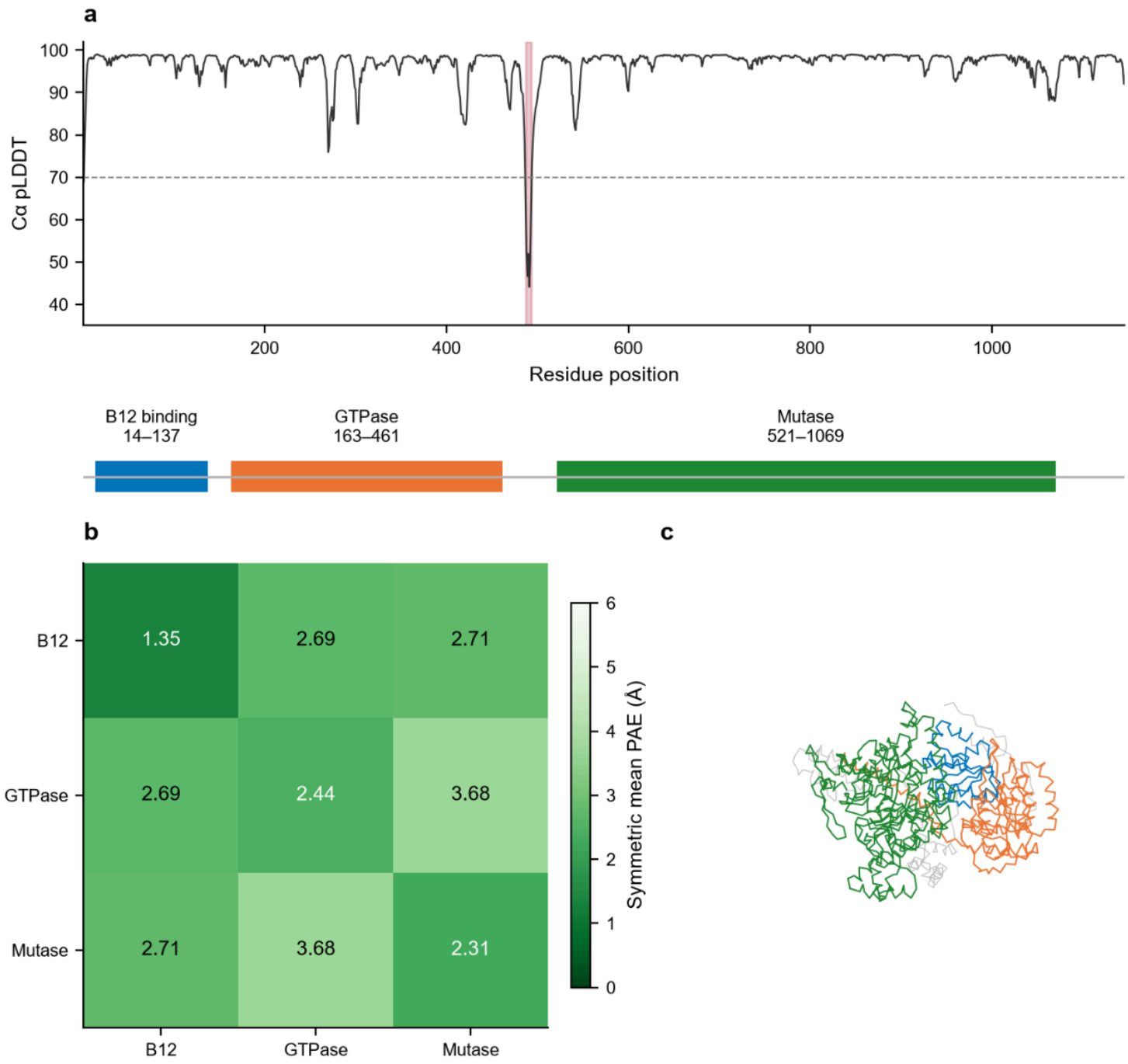



**Figure 3**

Confidence and annotated regions of NJT12_RS10585. a Residue-level Cα pLDDT, with the 70 threshold dashed and residues 487–493 shaded; the track shows the three non-overlapping CDD footprints. b Symmetric mean PAE within and between those footprints; each off-diagonal value averages the two directional blocks. c The actual Cα coordinate trace of representative model 0, coloured by the same footprints; unassigned residues are grey. The orientation is a principal-axis view and the trace is not an experimental structure. Archived mappings to PDB 4XC6 jointly cover 93.37% of the sequence

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1tables.zip
- OnlineResource2reproducibility.zip