**Evolutionary divergence and functional investigation of NAMLAA amidases – bacterial proteins essential to cell division.**

Sophie Rose Allen

Supervisor: Dr Mark Williams

MSc Bioinformatics (Part-Time, 2023-2024)

Birkbeck, University of London

# ABSTRACT (200-300 words)

A

B

C

(if I can change the title perhaps):

“The case for splitting NAMLAA domain classification: how a helix insertion in the AmiC protein leads to difference in function across bacterial species”

CONTENTS

[ABSTRACT (200-300 words) 2](#_Toc170025766)

[ACKNOWLEDGEMENTS 4](#_Toc170025767)

[ABBREVIATIONS 4](#_Toc170025768)

[INTRODUCTION 4](#_Toc170025769)

[MATERIALS AND METHODS 7](#_Toc170025770)

[Data Collection 7](#_Toc170025771)

[Identification of candidate AmiC structures 7](#_Toc170025772)

[Generating an AmiC sequence homologue list 8](#_Toc170025773)

[Experimental Structural Analysis 8](#_Toc170025774)

[Initial candidate structural alignment 8](#_Toc170025775)

[Sequence Analysis 9](#_Toc170025776)

[Sequence annotation and filtering 9](#_Toc170025777)

[Multiple Sequence Alignment, validation, and analysis 10](#_Toc170025778)

[Conservation of regions in the final MSA 10](#_Toc170025779)

[Sequence clustering and phylogeny 11](#_Toc170025780)

[Feature identification and Sequence Clustering 11](#_Toc170025781)

[Phylogeny 11](#_Toc170025782)

[Synthetic Structural Analysis 11](#_Toc170025783)

[RESULTS 12](#_Toc170025784)

[EXPERIMENTAL STRUCTURAL ANALYSIS 12](#_Toc170025785)

[Generating a candidate AmiC structure shortlist 12](#_Toc170025786)

[Candidate structural alignment 14](#_Toc170025787)

[SEQUENCE ANALYSIS 15](#_Toc170025788)

[Generating an AmiC sequence homologue list 15](#_Toc170025789)

[Multiple Sequence Alignment and Validation 16](#_Toc170025790)

[Conservation of mobile region across species 18](#_Toc170025791)

[Sequence clustering and phylogeny 19](#_Toc170025792)

[Feature identification and clustering 19](#_Toc170025793)

[Phylogeny 21](#_Toc170025794)

[Sequence clustering validation 22](#_Toc170025795)

[SYNTHETIC STRUCTURE ANALYSIS 22](#_Toc170025796)

[DISCUSSION 24](#_Toc170025797)

[CONCLUSION 28](#_Toc170025798)

[REFERENCES 28](#_Toc170025799)

[APPENDICES 33](#_Toc170025800)

[Appendix A: Longlist of candidate NAMLAA experimental structures identified from searches of the Protein Data Bank. 33](#_Toc170025801)

[Appendix B: Grouped multiple structural alignments of experimental structures with PDBeFOLD 38](#_Toc170025802)

[Appendix C: All validation scores for the multiple sequence alignment 40](#_Toc170025803)

[Appendix D: Full Logo for the final Multiple Sequence Alignment 41](#_Toc170025804)

[Appendix E: Occupancy plots for each insertion region 42](#_Toc170025805)

[Appendix F: PCA scatter plots for components PC1/PC3 and PC2/PC3 43](#_Toc170025806)

[Appendix G: Representative phylogenetic tree 44](#_Toc170025807)

# ACKNOWLEDGEMENTS

Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

On a personal note, I would also like to thank and acknowledge my supervisor Dr Mark Williams for his assistance throughout this project to guide both tool suggestions and the direction of analysis.

# ABBREVIATIONS

AmiC: Amidase C bacterial protein

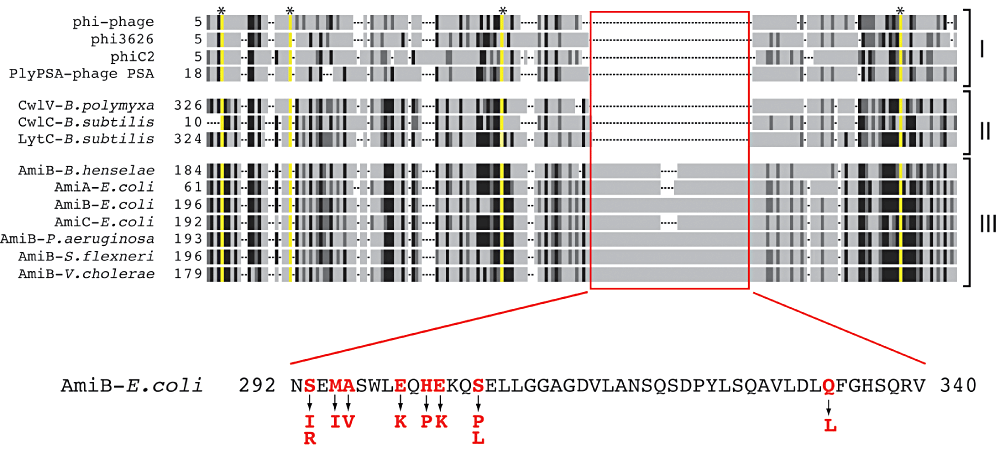
NAMLAA: N-acetylmuramoyl-L-alanine amidase

RMSD: Root-Mean-Square Deviation

PDB: Protein Data Bank

# INTRODUCTION

Intro: What NAMLAA amidases are, what they typically do in gram negative bacteria, the biochemical pathway involved, proteins involved, focus in on AmiC, focus in again on how it has been observed in gram positive bacteria and possibly other organisms, query why this is the case, and lead to bioinformatic work to identify the structures and sequences, and if there are functional differences in the AmiC protein family (as posited by the 3NE8(?) paper and CATH database)

* From annotation bibliography: Search for ‘AmiC E.coli’ in PubMed, 48 results
  + [**https://link.springer.com/article/10.1007/BF00272354**](https://link.springer.com/article/10.1007/BF00272354) **- action of the N-acetylmuramoyl-L-alanine amidase as cell wall lytic enzyme?**
  + Several studies using AmiC and the Tat exporter pathway mutants to explore antimicrobial effects or use in export of recombinant protein products from E. coli
  + Possible interaction with RNAse E?
  + AmiC might have a role as a exporter signal peptide? (AMIN type IV link from previous research, plus studies looking into its activity in Tat pathway)
  + Two possible candidates for interaction with AmiC? (DolP and ActS – ActS has a stronger link, 3 studies in particular, seems to prefer AmiC under standard growth conditions but AmiB in acidic pH?)
  + 4BIN and 3NE8 are AmiC/AmiB amidases, shown experimentally to have the mobile helix region.
  + From 3NE8 paper: 
  + 3NE8 paper posed 3 functions for LytC-type amidases (NAMLAA): lytic from bacteriophages, lytic from sporulating bacteria, and cell separation amidases. And that those with cell separation function have the self-regulating helix.
  + The theory at this stage was that the amidase\_3 in gram negative bacteria were intrinsically different to the amidase\_3 in gram positive bacteria, hypothesis being that function in gram negative bacteria was cell separation (in an autolytic cell cycle mechanism which requires an additional helix to ‘switch off’ the amidase activity when not in cell separation phase to avoid killing the parent cell), and in gram positive it had one of two roles (to assist in sporulation of child cells for sporulating bacteria in an autolytic cell cycle, or to kill parent cells upon infection for infectious bacteria acting in an endolytic fashion, both of which do not need a ‘switch’ as the protein needs to be on all the time to kill the parent cell). Additional group might be identified where amidase\_3 is in bacteriophages acting in the same pathway as the gram\_positive endolytic pathway to kill the target bacterial cell.
* Can we use the sequence data to identify the helix and functionally classify the NAMLA into sub-families? Because ‘NAMLAA’ or Amidase\_3 isn’t enough to distinguish these variety of functions on their own, clearly!
* CATH (<http://www.cathdb.info/version/v4_3_0/superfamily/2.60.40.3500/alignments>)
  + That link shows functional family separation across NAMLAA proteins, and seem to have different categories for gram negative and gram positive? Evidence that I’m going in the right direction here?
* Existing domain tools (reasons why functional difference in AmiC might not separate the different functions?)
  + UniProt/SwissProt: Provides assertion method (manual vs automatic), but draws in information from other tools or databases (does not appear to run the tools themselves in domain identification).
    - Includes both experimentally derived evidence and computational evidence. For amidase\_3 (amiC), had just automatic annotation from SMART ‘inferred from signature match’.
  + InterPro: Lists as ‘IPR002508’.
    - Incorporates the models from 13 member databases including SMART, adds the GO annotations on top of that (InterPro contains CATH-Gene3D (CATH for complete genomes), Pfam, CDD (multiple alignment databases), PROSITE, SMART, and others.)
    - I think matches the different signatures and profiles produced from the 13 databases and combines them into a consensus where they are different to assign them a family.
  + SMART: Lists as ‘Ami\_3’ (SM00646), reported in UniProt, Pfam, etc. Family includes phages as well as bacteria (therefore does not distinguish between specific function, although acknowledges phage endolytic and autolytic functional difference in-text – endolytic not having signal peptide and autopytlic having a C- or N-terminal cell wall binding domain).
    - Combines Uniprot, stable ensembl proteomes. Contains 137mill+ proteins from 537,000 species+.
    - 1998 paper: ‘Manually curated hidden Markov models’
    - Uses manual curation to annotate domains (using multiple sequence alignment and selecting a cut-off threshold for similarity to cluster each domain manually).
    - Gapped multiple sequence alignments (encompassing secondary structures of known tertiary structures), found candidate homologues using HMMer and two other iterative tools (MoST and WiseTools) which were statistically significant in their alignments but also includes some homologues experimentally shown to operate in the same biological context.
    - True positive/true negatives manually decided upon for each domain by assigning am SWise similarity score (scoring algorithm) threshold based on ‘published homology arguments’
    - Protein profiles generated from the alignments, then constructed hidden Markov models (assuming to identify patterns in the domains which can be applied to match with new sequences being entered?)
  + ECOD (domain ID e3ne8A1 for Amidase 3, using the protein family Amidase 3). F-class classification still contains both endolysins and autolysins.
    - Can be automatically assigned (was automatic for 3NE8).
    - F-group = domains with significant sequence similarity, mostly mapped Pfam families and some HHsearch clusters.
    - Also provides automatic groupings based on architecture of the structure, possible homologs based on high sequence/structure scores, functional similarity, features, topology.
    - Only uses structures derived in the PDB, emphasises distant evolutionary relationships.
    - Has a classification pipeline that gets more and more specific (starting with the SCOP superfamilies and pairing protein structures based on low lidentity, around 40%), but for F-class, uses hidden markov modelling like SMART does (HMMER and HHsearch-based clustering)
  + CDD/SPARCLE1: By the NCBI, architecture ID for NAMLA is 11436722
    - Searches based on protein sequence, uses RPS-BLAST to produce position-specific scoring matrices against a protein query (these are pre-calculated against the conserved domain database)
    - Manually curated at least for the NAMLA architecture
    - Does not distinguish between sporulating Clostridium difficile and the normal function cell cycle autolytic bacteria.
  + GO Annotations
    - A mix of manual annotations and computational predictions for gene products, indicated by evidence code.
    - Manual curation based on papers (functional evidence).
    - 99% are computational annotations: project annotations from one species to another based on orthology (like the EggNOG database), prediction based on curated rules, prediction based on sequence features, mapping of concepts to GO terms(?)
    - Annotations from: InterPro2GO, UniProt Keywords2GO, UniProt Subcellular Location2GO, EC2GO, UniRule2GO, Ensembl and Ensembl Genomes, UniPathway2GO, Gene Ontology Consortium, RNACentral
    - InterPro2GO: Manually curated file by looking between proteins and InterPro and selecting an appropriate GO term for all proteins in the family (using stats in keywords, comments, annotation conservation). 90-100% accurate.
    - Keyword2GO: Manually curated and based on literature/sequence analysis checks (from TrEMBL entries on keywords in nucleotide sequence, and using Spearmint decision tree program and RuleBase rules algorithm). 91-98% accurate.
  + PANTHER (N/A for NAMLA, no organism matches in their database, but uses protein placement in a phylogenetic tree to inform classification of sequences (TreeGrafter))
  + PROSITE (N/A didn’t have Amidase 3 as a domain, although describes it under the LytM entry)

# MATERIALS AND METHODS

The code used for all analysis can be found at the following GitHub repository: <https://github.com/sroseallen/Evolutionary-divergence-and-functional-investigation-of-NAMLAA-amidases>.

## Data Collection

### Identification of candidate AmiC structures

The FASTA sequence for AmiC protein structure 4BIN was downloaded from the Protein Data Bank (PDB)2–4. This sequence was used in three separate structure searches on 17/05/2023: a BLASTp search restricted to the PDB, a PDB search using the search term ‘amidase\_3’, and a SCOP search using identifiers 4001130 (N-acetylmuramoyl-L-alanine amidase-like family) and 8092183 (AmiC domain)5–7. Two further searches were performed on 19/05/2023 in Pfam (identifier PF01520) and Interpro (search term amidase\_3), however these did not identify any new experimental structures8,9. A final wider search for structures was performed in the PDB using the GO annotation ‘N-acetylmuramoyl-L-alanine activity’ and search terms ‘AmiA’, ‘AmiB’ and ‘AmiC’ in May 202310,11. The full longlist of experimental structures identified from these searches are listed in Appendix A.

Proteins were then shortlisted if they had a Pfam annotation of ‘amidase\_3’, a GO functional annotation of ‘N-acetylmuramoyl-L-alanine activity’, and an Interpro annotation of ‘Amidase\_3 domain’. BLASTp was then run using the PDB FASTA sequence for all structures in this shortlist (default parameters, against PDB database), and an additional structural alignment for all shortlisted structures using the Protein structure comparison service PDBeFold at European Bioinformatics Institute (<http://www.ebi.ac.uk/msd-srv/ssm>) 12,13 (parameters: against the entire PDB archive, lowest acceptable match of secondary structure alignment=50%, Q-score >0.3, match individual chains and connectivity, unique and best matches only, Normal precision). Structures were not added to the shortlist if alignment was to another chain in a model already on the list, or if the P-score was above 3. All shortlisted protein structures were downloaded from the PDB on 19/05/2023.

### Generating an AmiC sequence homologue list

BLASTp using FASTA protein sequences for each candidate structure was performed through the BLAST web browser on May 2023, and repeated in April 2024. Sequences were cropped to the Pfam annotation for the amidase\_3 domain as listed on the PDB record for each structure. The blastp algorithm was run using the default parameters with the following specifications: Database restricted to 59,610,710 sequences in RefSeq Select14, no organism or model exclusion, word size 0.05 (95% significance), word size 5, no restricting on maximum matches in the query range, BLOSUM62 matrix, gap costs existence 11, extension 1. The top 5000 aligned sequence results were downloaded for each search. A Python script was used to combine the sequences from BLASTp into a single list and remove duplicate entries. Notably, input sequences associated with the 1XOV, 3CZX, 3QAY, and 7TJ4 were not returned in their respective BLASTp searches as the sequences were not present within RefSeq Select, and were therefore manually added to the final BLASTp single list.

The HMMER 3 phmmer (REST) web service Python-based client 15 was used to search for further sequences using the below command, although outputs were not deemed informative for this purpose as only the top 50 alignments were provided without sequence ID/UniProt accession or the FASTA sequence itself.

python hmmer3\_phmmer.py --email 'sallen10@student.bbk.ac.uk' --database 'uniprotrefprot'

When searching InterPro for other relevant accessions (using the search term ‘N-acetylmuramoyl-L-alanine amidase’), several protein families of relevance were identified: 1 superfamily: [IPR036505](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR036505/) (PGRP domain superfamily), 6 families: [IPR010846](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR010846/) (AmiA-like), [IPR014234](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR014234/) (CwlD), [IPR017293](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR017293/) (SH3-domain containing protein), [IPR049745](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR049745/) (AmiC NAMLAA), [IPR050695](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR050695/) (amidase\_3), and [IPR051206](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR051206/) (amidase\_2). These families contained other domains which were not relevant for this stage of analysis, and were recorded for later domain-fusion analysis. 4 additional domains were also identified: [IPR002502](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002502/) (amidase\_2 domain), [IPR002508](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002508/) (amidase\_3 domain), [IPR021976](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR021976/) (CBD\_PlyG cell wall binding domain), and [IPR048586](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR048586/) (Cwp6 domain). Of these, only IPR002508 was used as a search term to download homologous sequences, since this was the only catalytic domain not associated with other protein families like amidase\_2. A Python script adapted from the InterPro website (<https://www.ebi.ac.uk/interpro/result/download/#/protein/UniProt/entry/InterPro/IPR002508/|fasta>) was used to download all IPR002508 sequences, UniProt accession IDs, taxonomic IDs and species names under the amidase\_3 entry IPR002508 and cropped the associated FASTA sequences to contain only the amidase\_3 domain as annotated by InterPro.

## Experimental Structural Analysis

### Initial candidate structural alignment

To visualise the alignment of the shortlisted structures and produce an all-vs-all pairwise structural alignment for later validation of the multiple sequence alignment, the MatchMaker tool within UCSF Chimera was run with the following parameters: secondary structure predictions not re-calculated, pairwise alignment, reference structure 4BIN in *E. coli*, Needleman-Waltsch alignment16,17. Multiple structure alignment was carried out using PDBeFOLD. Structure/sequence cropping was carried out in UCSF Chimera using a command script based on the Pfam boundaries of the ‘amidase 3’ domain as listed in the PDB. Confirmatory alignment of all structures to all other structures in the shortlist was carried out with a local download of US-align18 using the below command:

./USalign -dir chains\_amidase\_domain\_only/ chain\_list.txt -mm 4 -o sup -full T > output.txt

Comparison of RMSD scores between the amidase\_3 structure group and the AmiA-like structure group was calculated using a one-way ANOVA with the following R code:

df <- data.frame(structure\_type=c(rep("Amidase\_3",19), rep("AmiA-like",5)), RMSD=c(1.1019, 1.3471, 2.2253, 1.1724, 1.4097, 1.1283, 1.0965, 1.0891, 1.0517, 1.0529, 1.2284, 1.0970, 1.2498, 1.0451, 1.0681, 1.0429, 1.4485, 1.1393, 1.4056, 3.9191, 4.0545, 3.8910, 4.0163, 4.0122))

oneway.test(RMSD~structure\_type, df, var.equal=TRUE)

## Sequence Analysis

### Sequence annotation and filtering

Sequences from InterPro and BLASTp were combined using the BioPython19 module as part of a Python script, and duplicate sequences removed using a separate Python script. This script used taxonomic ID to identify unique sequences from the same species, and preferentially kept the RefSeq reference sequence from BLASTp over the InterPro sequence where applicable. This was to apply a level of caution to the InterPro sequences, which were mostly sourced from WGS preliminary or unreviewed sequences stored within UniProt, and therefore where a protein sequence has already been recorded for a species within RefSeq, the RefSeq sequence was used preferentially within the final sequence list as these have been annotated and curated by NCBI staff, and are non-redundant. Taxnomic ID was extracted within the InterPro API for InterPro sequences, and a separate Python script annotated BLASTp sequences with the taxonomic ID using a call to the NCBI REST API v220,21. If this API call failed, sequences were manually searched, as for some species (for example, species in the *Halomonas* genus) the species name did not match the name stored within the NCBI Taxonomy database22, but could be identified through manual searching of this database. Of note, some sequences from RefSeq had the same TaxID as other sequences from RefSeq; this is because the species name was used to call the REST API, and the species name was annotated at a higher order (for example, *Paenibacillus* genus instead of the species). All sequences from RefSeq were retained to avoid excluding novel species that were annotated at genus or family level.

The final list of sequences was annotated further with higher-order classification information for each species. This was done using a Python script to make an API call to Taxallnomy23 to annotate the family and phylum for each species. A second Python script then identified and removed sequences which were annotated with non-specific family names (such as ‘Fam\_of\_environmental samples’ or ‘uncultured’), sequences which failed the API Taxallnomy call, and any sequence which was partial (defined as a sequence length < 131, the threshold value determined as being more than 3 standard deviations (18.84) from the sample mean (187.63) of the amidase\_3 domain length in the candidate sequences).

### Multiple Sequence Alignment, validation, and analysis

The initial multiple sequence alignment was performed using a local download of kalign-324, chosen because kalign-3 was the most feasible tool to compute the alignment within reasonable computational time. The below command produced the initial alignment:

D:\msys2\usr\bin\kalign.exe -i keep\_seq\_taxid.txt -o D:\Documents\MSc\_Project\MSc\_Project\01\_DATA\Amidase\_3\04\_Multiple\_Alignment\alignment\_1.fa

The output from this was then filtered using OD-seq25, which is designed to detect outliers in large multiple sequence alignments by calculating an average distance matrix and predicting outliers using bootstrapping (1000 pseudo-replicates) to compute the mean and standard deviation of distance measures between an inspected sequence and the overall alignment. The following command was used to run a local instalment of OD-seq:

OD-seq -i alignment\_1.fa -r alignment\_1\_odseq.txt -o alignment\_1\_outliers.txt -c alignment\_1\_nonoutliers.txt -m affine -s 1

For OD-seq, ‘-r’ is the file name for all results, ‘-o’ is the file name for outlier sequences, ‘-c’ is the file name for non-outlier sequences, ‘-m’ is the gap penalty assignment method, and ‘-s’ is the number of standard deviations from the average distance permitted for sequence retainment. Affine was chosen as this gives a higher penalty (3 points) for opening a gap, and 1 point for each extension of an existing gap, meaning short gaps would be penalised in otherwise conserved regions without strongly penalising the long insertion region, as would be done for the Cumulative method. Alignment with kalign-3 was repeated as above after filtering. An R script was used to further filter out sequences which introduced a gap that contained information from <1% of all sequences in the alignment. Filtering was initially performed using 0.1% of sequences, however the alignment length remained around double the average amidase\_3 domain length from the candidate structures, and as amidase\_3 is a highly conserved domain, a shorter MSA obtained with a stricter threshold for information content was preferred to better ensure the sequence list contained ‘true’ amidase\_3 sequences. The MSA was visualised within Jalview26.

Validation was carried out according to the methods described by Modi and Dunbrack Jr.27 by comparing aligned residues in the structural alignment with the sequence alignment. Using the US-align all-vs-all pairwise alignment output, a Python script compared aligned residues in the structural alignment of each protein pair to the sequence alignment of the same proteins, and calculated the Positive Predictive Value (PPV): the ratio of sequence pairs aligned in both the sequences and structures to the total number of residues aligned in the sequences, and True Positive Rate (TPR): the ratio of sequence pairs aligned in both the sequences and structures to the total number of residues aligned in the structures.

### Conservation of regions in the final MSA

A conserved region was defined a continuous section of at least 3 columns where average occupancy of the columns was >90%, there was not >1 continuous column of <90% occupancy (gap-rich columns), and the region was not composed of >50% gap-rich columns. Regions were identified based on the JalView conservation quality scores, and boundaries determined using the above rules by viewing the alignment within JalView26,28,29. Logos were generated for conserved regions using WebLogo30, coloured by charge (positive=green, negative=red) and polarity (polar=blue, non-polar=black). Conservation scores were generated for each column of the MSA using Jalview web services28,29, and visualised by being mapped onto the residues for 4BIN and 3NE8 using custom UCSF Chimera attribute assignment files and rendering by attribute.

Sequences were annotated using an R script as being gram-positive or gram-negative based on the species phyla per the algorithm described for the R package AMR31, where bacteria within the phyla *Actinobacteria (*or *Actinomycetota), Chloroflexi, Firmicutes* (or *Bacillota*)or *Tenericutes* are gram-positive, and all else are gram-negative. To quantify the level of conservation of each species, species were annotated using a Python script depending on % of occupied columns (i.e. non-gap positions) across I-6, and species gram-status and phylum reviewed for patterns.

## Sequence clustering and phylogeny

### Feature identification and Sequence Clustering

For each insertion region identified, a Python script assigned each sequence a value of “0” if the sequence occupied <50% of the region with information, and a value of “1” if the sequence occupied ≥50% of the region, giving each sequence a binary fingerprint of 8 digits long based on which insertion regions the sequence appeared to have. These fingerprints were then used to cluster sequences based on occupancy in the insertion regions they contained – this was done using the Scikit-learn Python library to perform PCA with 8-components and a filtered 3-components32.

### Phylogeny

10 sequences from each unique binary fingerprint cluster were selected randomly using the ‘random’ Python package. Where a fingerprint had <10 sequences, all sequences were selected. A representative phylogenetic tree for these species was generated using the FastME webserver33. The FastME parameters used are as follows: Protein alignment, LG substitution model, Gamma distributed rates across sites, Gamma shape parameter 1, BioNJ algorithm34, Tree refinement with SPR\_BalME, and 100 bootstraps. The resulting phylogenetic tree was visualised using the Interactive Tree of Life (iTOL)35 and sequences coloured according to their binary fingerprint.

## Synthetic Structural Analysis

ColabFold36 was not used for synthetic structure predictions despite the accessibility and availability of options such as using a custom multiple sequence alignment, as a) the alignment from this research was based on an all-vs-all alignment set of sequences cropped to the amidase\_3 domain, rather than entire multi-domain protein sequences (and thus would not represent the entire protein prediction, possibly leading to poor confidence predictions for other domains in the AmiC protein), and b) the required custom alignment format requires sequences to be aligned in reference to the query sequence, and the alignment generated for this research is all-vs-all for the 20,403 (therefore the final alignment is longer than most query sequences, and is not compatible).

As such, an alternative approach was used. The AlphaFold Database37,38 contains predicted structures for each of the known proteins in UniProt, predicted using the monomer version of AlphaFold2. Therefore, 5 UniProt accessions from each binary fingerprint in the total dataset were selected and searched within the AlphaFold Database, and the structures retained for further inspection. In addition, 5 NCBI accessions from each fingerprint (or, if no structures of high enough quality in the AlphaFold database, further structures with Uniprot IDs), were selected and a structural prediction was generated using the OpenFold python notebook39 – an open source installation of AlphaFold2 which is powered for sequences for which very few template experimental structures are available in the PDB, which given the scarcity of structures found in the initial searches for data, is likely the most appropriate software for AmiC predictions. The following parameters were used for each sequence predicted: weight\_set=OpenFold, model\_mode=multimer, relax\_prediction=True(default).

To explore the possible associations of domains to different clusters, the full protein sequence was downloaded from NCBI RefSeq or UniProt respectively for structure predictions, using the sequence accession number in the file ‘BLAST\_Interpro\_all\_seq.txt’ as a search parameter. This allowed the full protein structure including other domains to be predicted (rather than just the amidase\_3 domain).

mmCIF Structures were examined and modelled using UCSF Chimera16 and colouring structure by pLDDT score using Render by Attribute (by residue, using “bfactor” which is equal to pLDDT in AlphaFold2 mmCIF files), and structures were also run through MolProbity40 and PAE Viewer41 to check for additional low confidence flags for structures not deposited in the AlphaFold Database. Low confidence flags are defined as a pLDDT score for a region of <70, or a high PSE score for proximity of two domains in the same structure.

# RESULTS

## EXPERIMENTAL STRUCTURAL ANALYSIS

### Generating a candidate AmiC structure shortlist

From an initial structural longlist constructed of possible NAMLAA structures across species, a final shortlist of 19 candidate AmiC structures was curated, which matched on Pfam annotation, GO annotation, and an Interpro domain annotation of amidase\_3 (Table 1). Searching of protein structure databases (see Methods) identified 24 potential NAMLAA structures, however five of these were annotated by Pfam8 as having an ‘AmiA-like’ domain rather than an ‘amidase\_3’ domain, and so were removed from the candidate list.

The lengths of the sequences for these structures varied widely in length from 175-638, however length of the annotated amidase\_3 domain varied less significantly, with a range from 159-215. 5J72 and 4RN7 are both from *Clostridium difficile*, however were deposited by different research groups and present different structures and sequences, where 5J72 is much longer. It is possible 4RN7 is a ‘cropped’ structure representing just the catalytic domain of this protein. Two of these structures (4M6H and 4M6I) are the same protein sequence from the same bacterium, however both were retained as one structure reflects the metal-bound form of the amidase, which may affect protein structure and may therefore be relevant in subsequent analyses. In creation of the shortlist, it is notable that a search within PDBeFOLD using the structure 4BIN had a low (<=17%) percentage of alignment of residues based on alignment in space for all identified structures. The species in this shortlist vary greatly, consisting of a mixture of gram-positive and gram-negative bacteria as well as bacteriophages, but tend to be reported with a zinc ion cofactor.

**Table 1: Shortlist of 19 candidate experimental structures of NAMLAA AmiC across different bacterial species.** Ligand, organism and sequence information taken from the PDB entry for each structure. Chain listed is the best-matched chain identified by the respective identification method for this shortlist. PDB entry reference is listed for each structure, as well as the publication where available.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PDB ID | Year in PDB | PDB  chain | Organism | Ligand | Full Sequence Length | Amidase domain length | Identification Method | Sequences identified from BLASTp |
| 4BIN2*,*3 | 2013 | A | *Escherichia coli* | Zn2+ ,Na+ | 403 | 215 | PDB search ‘AmiC’ | 5000 |
| 7B3N42,43 | 2020 | A | *Thermus parvatiensis* | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid | 177 | 159 | BLASTp search, 4BIN reference sequence | 5000 |
| 3CZX44 | 2008 | A | *Neisseria meningitidis* | Zn2+ | 182 | 169 | PDB search ‘Amidase\_3’ | 185 |
| 5J7245,46 | 2016 | A | *Clostridium difficile* | Citric Acid, Zn2+, Ca2+, Cl-, Na+ | 638 | 181 | BLASTp search, 4BIN reference sequence | 5000 |
| 7TJ447,48 | 2022 | B | *Staphylococcus aureus* | Zn2+ | 176 | 165 | BLASTp search, 4BIN reference sequence | 5000 |
| 4RN749 | 2014 | A | *Clostridium difficile* | Zn2+, Formic Acid, Glycerol, Ethanesulfonic Acid | 188 | 174 | BLASTp search, 4BIN reference sequence | 5000 |
| 5EMI50,51 | 2015 | A | *Nostoc punctiforme* | MES, MRD, Zn2+ | 180 | 169 | BLASTp search, 4BIN reference sequence | 5000 |
| 3NE852,53 | 2010 | A | *Bartonella henselae* | ACETATE ION, FORMIC ACID, GLYCEROL, Zn2+ | 234 | 215 | BLASTp search, 4BIN reference sequence | 5000 |
| 4M6G54,55 | 2013 | A | *Mycobacterium tuberculosis* | Zn2+ | 225 | 203 | PDB search ‘Amidase\_3’ | 2114 |
| 1JWQ56 | 2001 | A | *Paenibacillus polymyxa* | Zn2+ | 179 | 170 | BLASTp search, 4BIN reference sequence | 5000 |
| 4LQ657,58 | 2013 | A | *Mycobacterium tuberculosis* | SO4-, Zn2+, Cl-, Pt2+ | 218 | 203 | PDB search ‘Amidase\_3’ | 2114 |
| 4M6H59,55 | 2013 | A | *Mycobacterium tuberculosis* | None reported | 225 | 203 | PDB search ‘N-acetylmuramoyl-L-alanine activity’ | 2114 |
| 4M6I60,55 | 2013 | A | *Mycobacterium tuberculosis* | Zn2+ | 225 | 203 | PDB search ‘Amidase\_3’ | 2114 |
| 7AGO61,62 | 2020 | A | *Mycobacterium* | Zn2+, D-alpha-glutamine, Alanine | 233 | 203 | PDB search ‘Amidase\_3’ | 1942 |
| 7AGM63,62 | 2020 | A | *Mycolicibacterium smegmatis* | Zn2+ | 230 | 203 | PDB search ‘Amidase\_3’ | 2435 |
| 7AGL64,62 | 2020 | A | *Mycobacterium* | Zn2+ | 233 | 203 | PDB search ‘Amidase\_3’ | 1942 |
| 7RAG65,66 | 2021 | B | *Clostridium difficile* | Zn2+, 1,2-Ethanediol | 210 | 188 | PDB search ‘Amidase\_3’ | 5000 |
| 1XOV67,68 | 2004 | A | *Listeria monocytogenes* | Cl- ,GLUTAMIC ACID, LYSINE, 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL, SO4-, Zn2+ | 326 | 167 | PDB search ‘Amidase\_3’ | 1066 |
| 3QAY69,70 | 2011 | A | *Colneyvirus* | SO4-, Zn2+ | 180 | 172 | PDB search ‘Amidase\_3’ | 463 |

### Candidate structural alignment

Of the remaining 19 candidates, structural variation in the N-terminal domain of all proteins was shown, with these domains being unique to each protein. Specifically, 5J72 (*C. difficile*) had a large, unique CW-2 binding domain45,46. When cropping each protein to the C-terminal amidase\_3 domain only, the domain appeared to be highly conserved, with the exception of an additional α-helix present for some structures (Figure 1).

**Figure 1: Chimera visualisation of alignment of the 19 shortlisted structures.** A) Original structures as presented in the Protein Data Bank including longer N-terminal domains. B) Structures cropped to the C-terminal amidase\_3 domain boundaries; helixes are coloured in red, and beta sheets in purple. The additional α-helix is coloured in green, and is present for the following PDB structures: 4BIN, 3NE8.A close-up of a protein

Description automatically generated A close-up of a dna model

Description automatically generated

**B**

**A**

The 19 cropped candidate structures were then aligned for more detailed examination in PDBeFOLD. As the visualisation showed high variability outside of the amidase\_3 domain, these structures were cropped to this domain only before alignment. Overall, the RMSD demonstrated a close alignment of the carbon alpha backbone across the 19 structures (RMSD = 1.655), however a low Q-score (0.3561), which may reflect a longer alignment length from additional gaps. This is an improvement from the structural alignment performed with the uncropped structures (RMSD = 1.5525, Q-score 0.1255)(Appendix B). Notably, when examining the alignment of these structures, 4BIN and 3NE8 both had an additional helix structure which was absent in the alignment from the other structures (Figure 2), reflecting the visualisation of the alignment. These structures were both from gram-negative bacteria, however structures from other gram-negative bacteria 7B3N and 3CZX lack the helix in this alignment. Several bacteria also have a sequence consisting of a short helix and two beta strands near the start of the alignment.

**Figure 2: Multiple structural alignment representation of the 19 structures in PDBeFOLD.** Secondary structure predictions have been coloured; Helices = Red, Beta Sheets = Purple. The additional helix in 3NE8 and 4BIN has been highlighted in green.A screenshot of a computer

Description automatically generated

Confirmatory alignment with US-align produced an average TM-score of 0.83417 for the 19 cropped structures, with all structures having an individual TM-score above 0.8 with the exception of three structures from gram-negative bacteria; 4BIN, 3NE8, and 3CZX (Table 2). This is an improvement over the uncropped structures, with an average TM-score of 0.58520 (Appendix B). To validate removal of the ‘AmiA-like’ structures from the candidate list, a further structural alignment of the 19 candidates plus the five AmiA-like structures was performed, and showed that these five structures did not align well with the other 19; a one-way ANOVA for difference in means between the RMSD for the two groups produced a p-value of 2.26x10-16, indicating the AmiA-like group has significantly different RMSD scores to the other 19 structures (Appendix B).

|  |  |  |  |
| --- | --- | --- | --- |
| Structure | PDBeFOLD RMSD | PDBeFOLD Q-score | USalign TM-score |
| 3NE8 | 1.1502 | 0.5314 | 0.74304 |
| 4BIN | 0.9537 | 0.5699 | 0.77128 |
| 3CZX | 1.7016 | 0.5697 | 0.79843 |
| 7B3N | 1.7498 | 0.5651 | 0.85630 |
| 1JWQ | 0.9662 | 0.6479 | 0.88316 |
| 5J72 | 1.2414 | 0.6005 | 0.83297 |
| 4RN7 | 0.7880 | 0.6842 | 0.88708 |
| 7RAG | 1.0864 | 0.5987 | 0.81271 |
| 3QAY | 1.8619 | 0.5341 | 0.81289 |
| 7AGO | 0.7946 | 0.5863 | 0.81590 |
| 7AGL | 0.7904 | 0.5867 | 0.81593 |
| 7AGM | 0.7998 | 0.5858 | 0.81706 |
| 5EMI | 0.9011 | 0.6906 | 0.88620 |
| 7TJ4 | 1.1036 | 0.6442 | 0.88514 |
| 1XOV | 1.5857 | 0.5955 | 0.82427 |
| 4M6G | 0.7719 | 0.5885 | 0.81727 |
| 4LQ6 | 0.7409 | 0.5914 | 0.81818 |
| 4M6I | 0.7515 | 0.6882 | 0.88625 |
| 4M6H | 0.7498 | 0.6693 | 0.87099 |

**Table 2: PDBeFOLD and USalign scores from multiple alignment of the 19 cropped structures.**

## SEQUENCE ANALYSIS

### Generating an AmiC sequence homologue list

A BLASTp search for the top 5000 significantly aligned RefSeq sequences was performed for each of the candidate AmiC amidase\_3 domains, which identified in total 61,489 sequences. The number of identified sequences is given for each protein in Table 1. After removing duplicate sequences, BLASTp identified 18,247 candidate AmiC homologous sequences.

Interpro was then searched to identify further sequences. Of the search terms used to identify sequences potentially homologous to the AmiC amidase domain, the only domain that was not associated with amidase\_2 and was a catalytic domain instead of a binding domain was the ‘amidase\_3’ domain (IPR002508). Download of sequences annotated as IPR002508 identified 70,617 non-unique sequences, and after combining these with the sequences identified from BLASTp and removing duplicates, the final unique number of sequences was 36,233.

### Multiple Sequence Alignment and Validation

To identify conserved and insertion regions across species and cluster these accordingly, a multiple sequence alignment (MSA) for the sequence list was generated. To ensure a broad coverage of all possible AmiC homologues, sequence identification may have included non-AmiC proteins with a similar catalytic domain. To adjust for these less similar protein sequences pre-alignment, partial sequences (n=666) and sequences with a missing or unclassified taxonomic lineage (n=3,326) were also removed before alignment. After filtering, 32,241 sequences remained for the initial MSA.

To remove sequences which are unlikely to be AmiC homologues due to having an unusually divergent sequence from other homologues while still retaining sequences with insertions that are likely to be valuable in clustering of features, the initial multiple sequence alignment performed with k-align was further filtered in two stages:

* Outlier detection program OD-seq25 removed distant sequences (3,268) and removed none of the candidate structures.
* A separate R script removed sequences that introduced gaps which represented information from <1% of the data (see Methods). This removed 5 candidate sequences: 1JWQ (a *Paenibacillus* species), 3QAY (a bacteriophage), 3CZX (from *Neisseria meningitidis*), 7AGL and 7AGO (candidates from the *Mycobacteriaceae* family), 5J72 (*Clostridioides difficile*) and 7B3N (from the *Thermaceae* family).

The final post-filtered alignment comprised 20,304 sequences and was 399 residues long.

As described by Modi and Dunbrack Jr in 201927, the MSA was then validated by using a structural alignment of the candidate structures retained in the final alignment by examining the conserved (aligned) residues in the structural alignment versus the sequence alignment, and calculating the True Positive Rate (TPR) and Positive Predictive Value (PPV) for each. On average, TPR was 0.897 and PPV was 0.955 (TPR and PPV for each of the 9 structures are in Appendix C). There were 9 total potential insertions which are described in Table 3, the shortest being I-3 (5 residues) and the longest being I-6 (58 residues). To explore which regions these low-occupancy insertions refer to, conservation was mapped onto the 4BIN and 3NE8 structures (Figure 3), which showed high conservation in the core of the amidase\_3 structure and poor conservation of the outer helical structures. Region I-6, the longest insertion, maps onto both structures as being the proposed mobile helix.

**Table 3: Descriptive summary of final multiple sequence alignment of 20,304 sequences, length 399 residues.** Conserved region IDs = ‘C-X’, Insertion region IDs = ‘I-X’. Logos for conserved regions created using WebLogo30. Full web logo for all regions can be found in Appendix D.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Region | Region Columns | Region Length | Average Occupancy (%) | Consensus sequence | Logo |
| I-1 | 1-6 | 6 | 19.5% | -- | -- |
| C-1 | 7-14 | 8 | 97.3% | VVIDPGHG |  |
| I-2 | 15-57 | 43 | 30.1% | -- | -- |
| C-2 | 58-80 | 23 | 92.5% | TREKDVTLAIAKRLRALLEAQEP |  |
| I-3 | 81-85 | 5 | 26.1% | -- | -- |
| C-3 | 86-95 | 10 | 99.6% | ARVVLTRDGD |  |
| I-4 | 96-128 | 33 | 16.1% | -- | -- |
| C-4 | 129-153 | 25 | 95.5% | LRERVEIARKAGADLFVSIHADDAAF |  |
| I-5 | 154-175 | 22 | 27.9% | -- | -- |
| C-5 | 176-183 | 8 | 92.2% | GASVYALS |  |
| I-6 | 184-241 | 58 | 41.4% | -- | -- |
| C-6 | 242-251 | 10 | 85.2% | LADQVLLDEL |  |
| I-7 | 252-303 | 52 | 48.8% | -- | -- |
| C-7 | 304-354 | 51 | 92.9% | AGFAVLKEAPDKIPSVLVETGFISN  PEEEALLTSPAYQQKLAEAIADGIKR |  |
| I-8 | 355-399 | 45 | 13.5% | -- | -- |

A structure of a helix

Description automatically generatedA structure of a helix

Description automatically generated

**Figure 3: BLOSUM62 conservation scores mapped onto 4BIN and 3NE8.** Visualisation of structures was carried out using UCSF Chimera.Left: 4BIN, Right: 3NE8. Pink colouration = highly conserved residue, blue colouration = poorly conserved residue. The proposed mobile helix has been highlighted in each structure.

### Conservation of mobile region across species

Having defined the boundaries for conserved and less-conserved regions, and specifically the boundaries of the proposed mobile helix in region I-6, each species in the alignment was then examined to explore any species clustering patterns based on relative conservation of I-6. Based on the hypothesis that the mobile helix is linked to differing function in gram-negative (cell wall division) and gram-positive (endolytic behaviour) bacteria, species were also annotated with their predicted gram staining status (Figure 4).

A graph of different numbers and colors

Description automatically generated with medium confidenceA graph of a number of non-gap columns

Description automatically generated**Figure 4: Raw count of per-species occupancy of the I-6 region (left) and I-7 region (right), annotated by gram staining status.** Occupancy of other insertion regions is presented in Appendix E.

Looking at this conservation, it is evident that there is a clear binary difference in between ‘large insertion’ (>25 amino acids) and ‘low/no insertion’ characterised by the predicted gram-staining status of the bacteria. Some gram-negative bacteria lack this considerable insertion, aligning better with gram-positive bacteria with regard to this region. On further investigation, the majority of these bacteria share a common phylum (*Cyanobacteriota*). A considerable number of gram-positive bacteria occupied 20 columns in I-6, marking them as a unique group separate from both other gram-negative and gram-positive bacteria. With the exception of 12 species, these belonged to the *Streptomycetaceae* family. The same patterns were seen in the I-7 region, which for the 3NE8 and 4BIN structures maps onto a nearby long α-helix occurring just after the proposed mobile helix.

## Sequence clustering and phylogeny

### Feature identification and clustering

To quantify the degree of variation afforded by these insertion regions, 8-component PCA was conducted on the full dataset of sequences using the occupancy of each sequence in each insertion region. The scree plot (Figure 5B) shows that the first two PCs represent most of the total variance in the species, are most informative of the insertion regions, and therefore these two components were selected for the final PCA, producing the following plot:

A group of colorful dots

Description automatically generated

**A green and purple dots

Description automatically generated**A graph with blue bars

Description automatically generated with medium confidence A diagram of a number of red arrows

Description automatically generated

**Figure 5: PCA plots demonstrating variability for each insertion region.** A: Plot of PC1 and PC2, coloured by phylum. B: Scree plot for the 8-component PCA, the first two components represent most of the total variance seen in the insertion regions. C: Biplot of the 8 insertion regions projected onto PCA scatter plot of PC1 and PC2. D: PC1 and PC2 plot coloured by predicted gram staining status.

From the plot of PC1 and PC2 (Figure 5A), three major clusters emerge. Species with high variation in PC1 and low variation in PC2, which are largely *Actinomycetota*, species with mid-low variation in both PC1 and PC2, which are largely *Bacillota*, and species with high variation in PC2, which are largely *Pseudomonadota* and *Bacteroidota*. Using this clustering approach, separation between the gram negative phyla and the gram positive phyla is clear (Figure 5B), with an additional separation of two major gram positive phyla also demonstrated. The biplot reveals which features inform PC1 and PC2 (Figure 5D); larger occupancy in I-6, I-7 and I-8 is indicative of the species clustering in a homogenous gram-negative group, while higher occupancy in I-4 and I-5 in particular indicates belonging to the *Actinomycetota* phyla. This is demonstrated by the loading values for each feature, which are highest for I-6, I-7 and I-8 for PC2 and highest for I-2, I-3, I-4 and I-5 for PC1 (Table 4). Occupancy in I-1 is not informative towards any cluster, with a short eigenvector and little to no clustering around the centre of the plot, and seems to be the only informative feature for PC3 based on the loading scores.

|  |  |  |  |
| --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 |
| I-1 | 0.039796 | 0.091457 | 0.994439 |
| I-2 | 0.458354 | -0.17679 | 0.00627 |
| I-3 | 0.458183 | -0.17532 | 0.002198 |
| I-4 | 0.428548 | -0.2746 | 0.001025 |
| I-5 | 0.425983 | -0.27245 | -0.00017 |
| I-6 | 0.332375 | 0.448069 | -0.05178 |
| I-7 | 0.247136 | 0.529326 | -0.0395 |
| I-8 | 0.204286 | 0.546864 | -0.0825 |

**Table 4: Loading scores for PC1, PC2 and PC3 for the 8 insertion region features.**

### Phylogeny

Using the regions identified in Table 3, each sequence was then assigned an initial 8-digit binary fingerprint based on if that sequence had a higher-than-average occupancy in each of the 8 insertion regions, and then clustered based on these fingerprints. In total, there were 62 unique fingerprints, and a representative phylogenetic tree was generated based on these fingerprints using randomly selected sequences from each fingerprint cluster (Appendix G). A summary of the occurrences of these fingerprints is in Figure 6, where it becomes clear the most common combinations of insertion regions are I-5 alone; I-5, I-6, I-7 and I-8; and I-5, I-6, and I-7. Not every feasible combination was present in the sequence set, and perhaps notably there existed no species with every possible insertion region identified.

**Figure 6: Summary boxplot of 62 binary fingerprint clusters present in the overall sequence dataset of 20,304 sequences.**A graph with numbers and a blue line

Description automatically generated with medium confidence

Notably after this fingerprint clustering, there were no sequences with an I-6 region alone – species with an I-6 region exclusively occurred with other regions, such as the I-7 region. Looking at the representative phylogenetic tree, species with both these regions were predominately phyla which are predicted to be gram-negative, with some *Streptococcus* and *Streptomyces* species also containing both I-6 and I-7. Other associations were seen with specific phyla, such as the I-1 region being seen with *Pseudomonadota* and I-4 associated with gram-positive phyla. This random selection of species from each fingerprint identity does therefore seem to correlate with grouping of phyla, and also demonstrates a separation between species with the I-6 helix and species without.

### Sequence clustering validation

* If I can find an existing clustering algorithm for protein sequences, use this, do the results match up with the PCA model? Are the same patterns seen? (confirmatory results to prove the features I identified were relevant and useful in separating these different proteins out)
  + <https://huggingface.co/blog/AmelieSchreiber/plm-persistent-homology-msa-replacement> - clustering algorithm which doesn’t involve MSAs
    - <https://genome.cshlp.org/content/33/7/1145> - paper which I can cite for this methodology but blog puts things in a slightly more readable term for me
  + <https://www.nature.com/articles/s41586-023-06832-9> - AFCluster algorithm which clusters protein sequences on large-scale (paper uses alphafold database to show confirmation differences but also describes this clustering algorithm)
    - <https://github.com/HWaymentSteele/AF_Cluster> - github page for using the clustering algorithm

## SYNTHETIC STRUCTURE ANALYSIS

Due to the scarcity of experimental structures which are likely to be AmiC proteins in the PDB, many of the binary fingerprint clusters were not represented by experimental structures. As such, synthetic predictions were required to examine the probable structures of clusters, especially the highly represented I-5, I-6, I-7 and I-8 cluster.

4,776 of the sequences in the final sequence dataset had a corresponding structure predicted within the AlphaFold2 Protein Structure Database38. While sequences composed primarily of the amidase\_3 domain with minimal sequence outside of this, for larger proteins the database structures performed poorly. For these structures, such as those containing I-6, I-7, and I-8, the core amidase\_3 structure maintained generally high average pLDDT (>90), but outside of the core structure confidence was generally poor (pLDDT<50), especially with regard to position in physical space between multiple domains (Figure 7).

**Figure 7: Representative examples of structural predictions from the AlphaFold2 Protein Structural Database.** Left: UniProt ID A0A239XZL5, Sphingobacterium mizutaii, cluster group 00000000. Middle: UniProt ID A0A1H8MXL3, Duganella sp., cluster group 00001111. Right:UniProt IDA0A356EAJ1, Pasteurellaceae bacterium, cluster group 00001110. Larger structures appear to be multimeric, with loops of high uncertainty between possible domains and general disorder predicted. Structures coloured by pLDDT score; orange = <40, yellow = <50, light blue = <70, dark blue = <90

A blue and white ribbon

Description automatically generatedA close-up of a protein

Description automatically generated A structure of a protein

Description automatically generated with medium confidence

As such, a random selection of 3 sequences were taken from each cluster, and structures predicted for each using the multimeric version of AlphaFold2 using ColabFold. Even with the updated model, sequence coverage for regions not covered by dedicated domains remained poor.

A graph showing a number of positions

Description automatically generated with medium confidence**Figure X: Example sequence coverage for prediction of UniProt A0A0H3PCU7**

* Data sense check: Do these structures have the same patterns as expected from experimental data? What does the alphafold quality score say? Use a threshold of >70 for pLDDT for the NAMLA (less than this in some regions may = lack of data for training, disordered region, flexible region)
* Look at the structural and sequence differences between representatives from each cluster, what are these? What might they do? (eg that helix insertion in gram negative bacteria), link to discussion and try to find papers which describe possible function of these regions.
  + Specifically also look at the domains and domains in common across the clusters – are there any links here? Could feed in with the notes below about AMIN and other domains attached to the NAMLA amidase when talking about it in the discussion?
* Chimera representation and overlaps (colour insertion regions in, are they always in the same orientation even with other domains in the way? What do things look like in comparison to a 00000000 structure?)

Notes about domain fusion (for adding to discussion/research into structural similarities across species/fingerprint clusters)

* Start with AMIN domains
  + AMIN: PFAM PF11741: identified in 2765 species (all bacteria)
    - 5051 sequences with AMIN architecture, most of these are AMIN plus Amidase\_3 or AMIN plus LysM, but 633 also with secretin, secretin N and STN domains?
      * **Is this important?**
    - 3 instances in the PDB and 14 predicted alphafold structures (the 3 PDB have already been identified, see Day 1 notes)
  + SCOP search for AMIN domain-like family (4005368): 2 additional hits (4AQZ – another type Iva pilus machine, and 5H3K – hypothetical protein structure from *Synechocystis ‘Kazusa’*. No attached publication for this structure.)
  + AMIN: just 1 pathway – peptidoglycan recycling (<https://metacyc.org/META/NEW-IMAGE?type=NIL&object=PWY-7883&redirect=T> – takes place in cytosol?)
    - 7 PDB structures and 24,000+ alphafold predictions (4 new PDB structures: 4av2, 6ve2, 6ve3, 6ve4. All PilQ complexes from *Neisseria meningitidis* or *Pseudomonas aeruginosa* (the same typeIV secretion system as identified in Day 1) – is the E.coli amidase the only structure out there with AMIN that isn’t one of these secretion systems??
    - 242 domain architectures, 29,000+ protein structures (9 reviewed by UniProt/SwissProt)
      * Reviewed structures: Mix of PilQ and AmiC
  + AMIN: PFAM PF11741: identified in 2765 species (all bacteria)
    - 5051 sequences with AMIN architecture, most of these are AMIN plus Amidase\_3 or AMIN plus LysM, but 633 also with secretin, secretin N and STN domains?
      * **Is this important?**
    - 3 instances in the PDB and 14 predicted alphafold structures (the 3 PDB have already been identified, see Day 1 notes)
  + Identified some related domains which aren’t NAMLA but are listed under that GO annotation in PDB (N-acetylmuramoyl-L-alanine activity):
    - [PF01471](https://www.rcsb.org/search?q=rcsb_polymer_entity_annotation.annotation_id:PF01471%20AND%20rcsb_polymer_entity_annotation.type:Pfam&rt=polymer_entity) – PG\_binding\_1 domain (found in some of the Amidase\_2 family) – look into?
    - PF05036 – SPOR domain (binding to PG, involved with sporulation and division proteins like FtsN?)
    - PF01832 – Glycosaminidase (hydrolysing peptidoglycan)
    - PF00877 – NipC domain in TB (related to NIpD??)
    - PF12123 – CBC\_PlyG (cell wall binding domain in bacteria and viruses)
    - PF05257 – CHAP domain (related to amidase function, cell wall metabolism in bacteria, amidase domain of E. coli glutathionylspermidine synthetase?)
* Return to the shortlist of 19 and look at the N-terminal domains – what are these? What do they do?
* Use CATH searches and funfam matches to identify these
* Different domain fusions = different functions and different classes? Different clusters?
* From InterPro search for ‘Amidase\_3’ 09/04/2024:
  + IPR019606 – GerMN domain: ‘The GerMN domain is a region of approximately 100 residues that is found, duplicated, in the Bacillus GerM protein and is implicated in both sporulation and spore germination. It is also found in lipoprotein LpqB. The domain is present in a number of different bacterial species both alone and in association with other domains such as Gmad1 and Gmad2. It is predicted to have a novel α-β fold.’ **Could be a signifier of function if NAMLAA domain is found with this domain in a protein?**
  + [IPR021976](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR021976/) (CBD\_PlyG, cell wall binding domain), found with InterPro search for N-acetylmuramoyl-L-alanine amidase

# DISCUSSION

Discussion

* Candidate experimental structural alignment
  + High TM scores and average Q-scores indicate conservation for these structures – closely related
  + Including AmiA structure (x5) reduced average Q-score and TM-score significantly(Appendix C), therefore some confirmatory proof that these structures were all AmiC (as per 4BIN structure).
    - When cropping to what should be the most conserved portion of the enzyme given the patterns seen in the other 19 structures, the other 5 should in theory align very well if they are the same enzyme. However, RMSD was higher than 3 for all five structures, and Q-score was less than 0.1. These are poor scores, and alongside the consistently low RMSD scores of all the other 19 structures, provides some indication that these 5 structures - despite being identified through searches for NAMLAA – are indeed distinct from the highly conserved structures of the amidase\_3 enzymes. 3CZX had a slightly higher RMSD (2.22), however the Q-score was comparable to the other 19 structures and so was retained.
    - AmiA differences: Domain annotation, but also when including these in the PDBeFOLD initial alignment, the RMSD and Q-score become worse (RMSD: 2.858, Overall Q-score: 0.01216). There is also difference in the secondary structure, which perhaps causes the worse quality scores, and also notable that these did not appear in initial similarity searches and do not have a zinc ion ligand listed in their PDB entries list the others, instead having a sodium ion. What might this mean?
  + However, sequence/structure searching: Low % in alignment (low sequence identity), and low similarity matches in PDBeFOLD, possibly due to the N-terminal regions and uniqueness of these?
    - 3NE8 and 4BIN: When the initial 19 structures on the shortlist were aligned with PDBeFOLD and USAlign, these two proteins were the only ones with this apparent helix insertion region, and so this was used as a crude identifier for amidase\_3 function in the cell wall (based on inferences from literature and specifically the 3NE8 paper which had an early sequence alignment showing the helix was present in cell wall proteins and not in other amidase\_3 with functions like in Clostriudium difficile ie sporulating bacteria).
    - Clostrium difficle is sporulating, therefore the amidases here which had poor alignment structurally were probably in this functional class alone in the dataset (explains poorer TM scores for 7RAG, 5J72 when not cropped) (appendix C)
    - 5J72: Longer N-terminal domain than the others, no matches of significance when searched in PDBeFOLD, what is this? Why is it so different? Bacteriophage amidases coming with a long N-terminal domain (5J72), to bind to cell wall?
      * Notably, searching for NAMLAA in Clostridoides difficile, there are two ‘NAMLAA’ appearing – one type which is shorter, and one type with this long N terminus. Could this bacteria have two NAMLAA amidases conducting multiple functions?
    - 7AGL/M/O and 5LQ6 and 4M6H have a short insertion sequence near the start of the structure. These are all Mycobacterium or Mycolicibacterium. Is this something unique to this family of bacteria? Corresponds possibly to region I-1 in current alignment – second evolutionary marker, does it have functional consequence?
* Sequence analysis
  + Removal of structural candidate sequences at later stage filtering (which isn’t due to quality of the sequence, and more to do with that these sequences did not fit with the majority of other identified sequences well enough).
    - None of the candidate sequences with an associated structure were removed prior to multiple sequence alignment or with the OD-seq filtering – so the sequences weren’t removed due to poor quality or looking incomplete, therefore sequences removed because they were too distant from the others in the set.
    - Thermus family being separate anyway (from *Thermaceae*, which is known to be distinct from other bacterial families through conserved signature insertions/deletions71).
    - One Neisseria meningitidis (3CZX) removed as well explore why this might be?
    - 5J72 removed, which had a long n terminal, and the other clostridium not removed, 4RN7. Why? Are these different amidases with different roles? Does c. difficile have two different forms of the amidase in fact?
    - Colnyvirus (3QAY) being fairly distinct too – however the other phage in the candidate list (1XOV) was not removed and ended up in the final set – why?
    - 7AGL and 7AGO, both mycobacteria family being removed with the 1% information filter, although Mycolicibacterium being retained. Is there a difference functionally? Do some mycobacteria have amidase\_3 function and others do not? Cell wall structure being different to gram negative despite cell wall presence, etc etc.
      * Notably: of the 7 removed at this stage, 6 are retained by adjusting the filtering threshold from 1% information to 0.1% information, meaning they are not massively distinct and were removed because they introduced gaps which had information from between 29 and 290 sequences. Possibly, there are some very low-level low-information insertion regions from these bacteria and phages, but would add too much noise to the alignment to include on balance (the 0.1% filtered alignment was considerably longer (601) than the 1% alignment (400) and introduced several gaps in the middle of otherwise high-occupancy conserved regions).
  + Conservation from the alignment mapped onto structures: high conservation in the enzymatic core and poor conservation on the outside of this core – as expected, commonly seen for enzyme families. Specifically links in as the proposed mobile helix is also poorly conserved despite having a potential function, so likely that these enzymes have similar function (since both structures where available and sequences are highly conserved having similar sequence properties and overlapping structures in space), although not all have a helix which is present in the enzymatic core – therefore same enzyme, different function?
  + Look at the conserved region logos – any conserved residues in particular? Same amino acids, same polarity/charge? What does this mean for the enzymatic site function as these conserved regions map well onto the enzymatic core? Can we say again that despite differing species and differing gram status, the enzymatic function is retained across species, and the difference is in how the enzyme is used/activated (possibly by way of the mobile helix)?
  + Limitation: As sequence identity is so low for these proteins, possible that sequence included in the multiple sequence alignment are not true AmiC homologues, which may affect the final phylogeny if we assume all sequences are AmiC homologues (and possibly conflate non-helix containing AmiC sequence/structures with other proteins entirely, meaning the set of AmiC gram positives might not actually be AmiC – weakning the strength of the conclusion that AmiC with helix are different to AmiC without helix, might be that the only AmiC in there are ones with a helix). As not really much separation in the function of amidase\_3 domain in current domain annotations, this is to be expected, have tried to adjust for this known-unknown with quality filtering steps and validated with a structural alignment of candidate AmiC structures – this alignment also showed that there are large structural differences between known AmiC and AmiA proteins, and therefore sequences introducing large gaps in the alignment may not be true AmiC homologues, however certainty of this is not complete and could only be resolved by elucidating more structures of AmiC proteins (that have been confirmed to be AmiC experimentally through their biochemical role in the bacterial cell) to confirm there are AmiC with and without this helix that match with their functional role.
* Gram positive/negative splitting
  + General limitation: Acknowledge that this is a non-biological term, not very specific and boundaries can be blurred (see: Mycobacterium tuberculosis), less of a definitive term and more of a general term to relate the sequence/structure differences to a biological function. Also acknowledge that annotation of species as gram negative/positive is based on the classification of each species, and there is a chance that some of the enzymes have been mis-classified as the wrong species, especially for WGS shotgun sequences from InterPro. Therefore, have tried to define those ‘gram negative’ with a helix as ‘cell wall separation’ function, and those ‘gram positive’ as alternative sporulating or lytic functions, and state that this classification is a general term to group the families with/without the helix. Have also tried to apply quality filter to the InterPro sequences, and preferentially includes sequences from species that had been curated for the RefSeq database.
  + Gram negative having the I-6 insertion and I-7 insertion, gram positive tend not to. Supports the hypothesis that this insertion is relevant for a function in gram-negative that is not used in gram-positive bacteria – cell-wall separation would require such an insertion if the function of this insertion is to ‘switch off’ the enzymatic activity, whereas gram positive bacteria – which do not have a cell wall – would require the enzyme for a different reason that may not require such ‘switching off’ – for example, it would not be useful for an enzyme used in endolytic bacteria designed to kill neighbouring or parent bacterial cells to have an off switch, since uncontrolled amidase activity would result in cellular breakdown, which would be the goal of such endolytic activity. I-6 as a marker of function? What about I-7? What is this doing as the same patterns are seen? Is this also linked to gram-negative ‘switching off’?
  + However the separation is not perfect – there are gram negative bacteria lacking the insertion. In I-6, these low-occupancy gram negative species share *Cyanobacteriota* as their phylum for most (which all had an insertion 13-15), with some *Bacillota*, whereas those with the insertion are among *Pseudomonata*. This pattern is also seen in the I-7 region. Why?
    - Evolutionary terms – phylogenetic split? Look into when these phyla diverged, perhaps this marks a divergence.
    - Might these bacteria have alternative activity?
    - Do the bacteria in fact have multiple copies of the AmiC NAMLAA, where one copy is for cell wall activity and one is for endolytic activity? Do these bacteria have paralogs for multiple purposes?
  + Some gram positive in I-6 have more occupied columns than other gram positives appear unique. These are almost exclusively *Streptomycetaceae*. Why? What are these? Why do they have a considerable insertion region when the other gram positives don’t seem to?
    - May not be particularly significant after all; 506/1120 species with 20 occupied columns in I-6 are different Streptomyces strains that could not be filtered earlier (as their records did not have a species name, just a genus and an alphanumeric ‘species’ which appears to be a strain ID).
* Insertion region identification and what that means in relation to the clusters the insertions were found in
  + Cluserting with PCA: Insertion regions separate into two main clusters with the first two PCs, seem to correspond nicely to gram positive and negative, however is this truly showing anything or is it just representing how the I-6 and I-7 regions are big, and separate gram positive and negative well? Does this actually give any additional information?
  + Gram positive and negative cluster well – all gram positive cluster with other gram positive, but a few gram negative cluster with gram positive rather than the gram negative.
  + Universality of I-5 means anything?
  + I-6+I-7 are gram negative except for SOME streptococcus and streptomyces. Why might this be?
  + Not all fingerprint region combinations present (around 1/3 of possible combinations of 8 binary traits). Are they biologically viable? Or not? Why might this be?
* Bring in studies on the helix being used to block the enzyme activity when not needed, and why this might be absent for other bacteria
* If this aligns with the tree (is this a deletion? An insertion? Do proteins with/without the helix have a common ancestor, is the protein even related anymore if the helix is missing? Should this be indicative of a different protein family entirely?)
* Other structural findings go here (eg evidence 2d section, papers backing up what features which separate the structures do)
  + Example predicted structure for each fingerprint region – in FULL, the whole protein, not just the enzymatic region. Do they have similarities in N-terminal regions displayed?
  + Notably, of the structures within the AlphaFold Database, most had low average pLDDT. This may be due to other domains in the full AmiC proteins – most, if not all, of the structures predicted had at least one other domain, and the structures in the AlphaFold structure database currently were generated with the monomeric version of AlphaFold2.

# CONCLUSION

Conclusion

* Some statement on the markers for differentiating the protein by function, and argue that the absence of the helix could indicate different protein function (even though the enzyme acts by the same activity and technically has the same function, the bacterial action of the protein is so different between species clusters that perhaps there should be a separation in the way the domain is classified based on absence/presence of that helix near the structural core?)

# REFERENCES

1. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI’s conserved domain database. Nucleic Acids Res. *43*, D222-226. https://doi.org/10.1093/nar/gku1221.

2. Kerff, F., Rocaboy, M., Herman, R., Sauvage, E., and Charlier, P. (2013). Crystal structure of the E. coli N-acetylmuramoyl-L-alanine amidase AmiC. https://doi.org/10.2210/pdb4bin/pdb.

3. Rocaboy, M., Herman, R., Sauvage, E., Remaut, H., Moonens, K., Terrak, M., Charlier, P., and Kerff, F. (2013). The crystal structure of the cell division amidase AmiC reveals the fold of the AMIN domain, a new peptidoglycan binding domain. Mol. Microbiol. *90*, 267–277. https://doi.org/10.1111/mmi.12361.

4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.

5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

6. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. *42*, D310–D314. https://doi.org/10.1093/nar/gkt1242.

7. Andreeva, A., Kulesha, E., Gough, J., and Murzin, A.G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. *48*, D376–D382. https://doi.org/10.1093/nar/gkz1064.

8. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

9. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.

10. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

11. The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*, iyad031. https://doi.org/10.1093/genetics/iyad031.

12. Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268. https://doi.org/10.1107/S0907444904026460.

13. Krissinel, E., and Henrick, K. (2005). Multiple Alignment of Protein Structures in Three Dimensions. In Computational Life Sciences, M. R. Berthold, R. C. Glen, K. Diederichs, O. Kohlbacher, and I. Fischer, eds. (Springer), pp. 67–78. https://doi.org/10.1007/11560500\_7.

14. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-745. https://doi.org/10.1093/nar/gkv1189.

15. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

16. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. https://doi.org/10.1002/jcc.20084.

17. Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C., and Ferrin, T.E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics *7*, 339. https://doi.org/10.1186/1471-2105-7-339.

18. Zhang, C., Shine, M., Pyle, A.M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat. Methods *19*, 1109–1115. https://doi.org/10.1038/s41592-022-01585-1.

19. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

20. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.

21. Federhen, S. (2012). The NCBI Taxonomy database. Nucleic Acids Res. *40*, D136-143. https://doi.org/10.1093/nar/gkr1178.

22. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O’Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J. Biol. Databases Curation *2020*, baaa062. https://doi.org/10.1093/database/baaa062.

23. Sakamoto, T., and Ortega, J.M. (2021). Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. BMC Bioinformatics *22*, 388. https://doi.org/10.1186/s12859-021-04304-3.

24. Lassmann, T. (2020). Kalign 3: multiple sequence alignment of large datasets. Bioinformatics *36*, 1928–1929. https://doi.org/10.1093/bioinformatics/btz795.

25. Jehl, P., Sievers, F., and Higgins, D.G. (2015). OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics *16*, 269. https://doi.org/10.1186/s12859-015-0702-1.

26. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

27. Modi, V., and Dunbrack, R.L. (2019). A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. Sci. Rep. *9*, 19790. https://doi.org/10.1038/s41598-019-56499-4.

28. Troshin, P.V., Procter, J.B., Sherstnev, A., Barton, D.L., Madeira, F., and Barton, G.J. (2018). JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure. Bioinformatics *34*, 1939–1940. https://doi.org/10.1093/bioinformatics/bty045.

29. Troshin, P.V., Procter, J.B., and Barton, G.J. (2011). Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. Bioinformatics *27*, 2001–2002. https://doi.org/10.1093/bioinformatics/btr304.

30. Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190. https://doi.org/10.1101/gr.849004.

31. Berends, M.S., Luz, C.F., Friedrich, A.W., Sinha, B.N.M., Albers, C.J., and Glasner, C. (2022). AMR: An R Package for Working with Antimicrobial Resistance Data. J. Stat. Softw. *104*, 1–31. https://doi.org/10.18637/jss.v104.i03.

32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

33. Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol. *32*, 2798–2800. https://doi.org/10.1093/molbev/msv150.

34. Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. *14*, 685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808.

35. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res., gkae268. https://doi.org/10.1093/nar/gkae268.

36. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

37. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

38. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.

39. Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O’Donnell, T.J., Berenberg, D., Fisk, I., Zanichelli, N., et al. (2024). OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nat. Methods, 1–11. https://doi.org/10.1038/s41592-024-02272-z.

40. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. Publ. Protein Soc. *27*, 293–315. https://doi.org/10.1002/pro.3330.

41. Elfmann, C., and Stülke, J. (2023). PAE viewer: a webserver for the interactive visualization of the predicted aligned error for multimer structure predictions and crosslinks. Nucleic Acids Res. *51*, W404–W410. https://doi.org/10.1093/nar/gkad350.

42. Freitag-Pohl, S., and Pohl, E. (2022). AmiP amidase-3 from Thermus parvatiensis.

43. Jasilionis, A., Plotka, M., Wang, L., Dorawa, S., Lange, J., Watzlawick, H., van den Bergh, T., Vroling, B., Altenbuchner, J., Kaczorowska, A.-K., et al. (2023). AmiP from hyperthermophilic Thermus parvatiensis prophage is a thermoactive and ultrathermostable peptidoglycan lytic amidase. Protein Sci. Publ. Protein Soc. *32*, e4585. https://doi.org/10.1002/pro.4585.

44. Zhang, R., Zhou, M., Bargassa, M., Joachimiak, A., and Midwest Center for Structural Genomics (2011). The crystal structure of the putative N-acetylmuramoyl-L-alanine amidase from Neisseria meningitidis.

45. Renko, M., Usenik, A., and Turk, D. (2017). Cwp6 from Clostridium difficile.

46. Usenik, A., Renko, M., Mihelič, M., Lindič, N., Borišek, J., Perdih, A., Pretnar, G., Müller, U., and Turk, D. (2017). The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the Clostridium difficile Cell Wall Proteins Cwp8 and Cwp6. Structure *25*, 514–521. https://doi.org/10.1016/j.str.2016.12.018.

47. Page, J.E., Skiba, M.A., Kruse, A.C., and Walker, S. (2022). Structure of the S. aureus amidase LytH and activator ActH extracellular domains.

48. Page, J.E., Skiba, M.A., Do, T., Kruse, A.C., and Walker, S. (2022). Metal cofactor stabilization by a partner protein is a widespread strategy employed for amidase activation. Proc. Natl. Acad. Sci. *119*, e2201141119. https://doi.org/10.1073/pnas.2201141119.

49. Tan, K., Mulligan, R., Kwon, K., Anderson, W., Joachimiak, A., and Center for Structural Genomics of Infectious Diseases (2014). The crystal structure of N-acetylmuramoyl-L-alanine amidase from Clostridium difficile 630.

50. Büttner, F.M., and Stehle, T. (2016). N-acetylmuramoyl-L-alanine amidase AmiC2 of Nostoc punctiforme.

51. Büttner, F.M., Faulhaber, K., Forchhammer, K., Maldener, I., and Stehle, T. (2016). Enabling cell–cell communication via nanopore formation: structure, function and localization of the unique cell wall amidase AmiC2 of Nostoc punctiforme. FEBS J. *283*, 1336–1350. https://doi.org/10.1111/febs.13673.

52. Tan, K., Rakowski, E., Buck, K., Joachimiak, A., and Midwest Center for Structural Genomics (2010). The crystal structure of a domain from N-acetylmuramoyl-l-alanine amidase of Bartonella henselae str. Houston-1.

53. Yang, D.C., Tan, K., Joachimiak, A., and Bernhardt, T.G. (2012). A conformational switch controls cell wall-remodelling enzymes required for bacterial cell division. Mol. Microbiol. *85*, 768–781. https://doi.org/10.1111/j.1365-2958.2012.08138.x.

54. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the Mycobacterium tuberculosis peptidoglycan amidase Rv3717 in complex with L-Alanine-iso-D-Glutamine reaction product.

55. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., and Alber, T. (2013). Structural and Biochemical Analyses of Mycobacterium tuberculosis N-Acetylmuramyl-l-alanine Amidase Rv3717 Point to a Role in Peptidoglycan Fragment Recycling \*. J. Biol. Chem. *288*, 31549–31555. https://doi.org/10.1074/jbc.M113.510792.

56. Yamane, T., Koyama, Y., Nojiri, Y., Hikage, T., Akita, M., Suzuki, A., Shirai, T., Ise, F., Shida, T., and Sekiguchi, J. (2003). Structure of the catalytic domain of CwlV, N-acetylmuramoyl-L-alanine amidase from Bacillus(Paenibacillus) polymyxa var.colistinus.

57. Kumar, A., Kumar, S., Kumar, D., Mishra, A., Dewangan, R.P., Shrivastava, P., Ramachandran, S., and Taneja, B. (2013). Crystal structure of Rv3717 reveals a novel amidase from M. tuberculosis.

58. Kumar, A., Kumar, S., Kumar, D., Mishra, A., Dewangan, R.P., Shrivastava, P., Ramachandran, S., and Taneja, B. (2013). The structure of Rv3717 reveals a novel amidase from Mycobacterium tuberculosis. Acta Crystallogr. D Biol. Crystallogr. *69*, 2543–2554. https://doi.org/10.1107/S0907444913026371.

59. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the reduced, metal-free form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717.

60. Prigozhin, D.M., Mavrici, D., Huizar, J.P., Vansell, H.J., Alber, T., and TB Structural Genomics Consortium (2013). Structure of the reduced, Zn-bound form of Mycobacterium tuberculosis peptidoglycan amidase Rv3717.

61. Blaise, M. (2020). crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus bound to L-Alanine-D-isoglutamine.

62. Küssau, T., Van Wyk, N., Johansen, M.D., Alsarraf, H.M.A.B., Neyret, A., Hamela, C., Sørensen, K.K., Thygesen, M.B., Beauvineau, C., Kremer, L., et al. (2020). Functional Characterization of the N-Acetylmuramyl-l-Alanine Amidase, Ami1, from Mycobacterium abscessus. Cells *9*, 2410. https://doi.org/10.3390/cells9112410.

63. Blaise, M., and Alsarraf, H.M.A.B. (2020). Crystal structure of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium smegmatis.

64. Blaise, M. (2020). crystal structure of the apo form of the N-acetylmuramyl-L-alanine amidase, Ami1, from Mycobacterium abscessus.

65. Eckenroth, B.E., and Doublié, S. (2021). Structure of the CwlD amidase from Clostridioides difficile in complex with the GerS lipoprotein.

66. Feliciano, C.A., Eckenroth, B.E., Diaz, O.R., Doublié, S., and Shen, A. (2021). A lipoprotein allosterically activates the CwlD amidase during Clostridioides difficile spore formation. PLOS Genet. *17*, e1009791. https://doi.org/10.1371/journal.pgen.1009791.

67. Korndörfer, I.P., and Skerra, A. (2005). The crystal structure of the listeria monocytogenes bacteriophage PSA endolysin PlyPSA.

68. Korndörfer, I.P., Danzer, J., Schmelcher, M., Zimmer, M., Skerra, A., and Loessner, M.J. (2006). The Crystal Structure of the Bacteriophage PSA Endolysin Reveals a Unique Fold Responsible for Specific Recognition of Listeria Cell Walls. J. Mol. Biol. *364*, 678–689. https://doi.org/10.1016/j.jmb.2006.08.069.

69. Mayer, M.J., Garefalaki, V., Spoerl, R., Narbad, A., and Meijers, R. (2011). Catalytic domain of CD27L endolysin targeting Clostridia Difficile.

70. Mayer, M.J., Garefalaki, V., Spoerl, R., Narbad, A., and Meijers, R. (2011). Structure-Based Modification of a Clostridium difficile-Targeting Endolysin Affects Activity and Host Range. J. Bacteriol. *193*, 5477–5486. https://doi.org/10.1128/jb.00439-11.

71. Ho, J., Adeolu, M., Khadka, B., and Gupta, R.S. (2016). Identification of distinctive molecular traits that are characteristic of the phylum “Deinococcus-Thermus” and distinguish its main constituent groups. Syst. Appl. Microbiol. *39*, 453–463. https://doi.org/10.1016/j.syapm.2016.07.003.

# APPENDICES

## Appendix A: Longlist of candidate NAMLAA experimental structures identified from searches of the Protein Data Bank.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Organism | PDB ID | PDB chain | Year released in PDB | Ligand |
| *Bacteroides uniformis* | 4H4J | A | 2014 | None reported |
| *Bacteroides frafilis* | 2P1G | A | 2007 | None reported |
| *Bacteroides frafilis* | 2P1G | B | 2007 | None reported |
| *Legionella pneumophila* | 2IM9 | A | 2006 | None reported |
| *Bacteroides uniformis* | 4Q68 | A | 2014 | NAG, NA+ |
| *Bacteroides uniformis* | 4Q5K | A | 2014 | 2YP, NA+ |
| *Bartonella henselae* | 3NE8 | A | 2010 | ACETATE ION, FORMIC ACID, GLYCEROL, Zn2+ |
| *Escherichia coli* | 4BIN | A | 2013 | Zn2+ ,Na+ |
| *Neisseria meningitidis* | 3CZX | A | 2008 | Zn2+ |
| *Neisseria meningitidis* | 3CZX | B | 2008 | Zn2+ |
| *Neisseria meningitidis* | 3CZX | C | 2008 | Zn2+ |
| *Neisseria meningitidis* | 3CZX | D | 2008 | Zn2+ |
| *Thermus parvatiensis* | 7B3N | A | 2020 | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid |
| *Thermus parvatiensis* | 7B3N | B | 2020 | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid |
| *Thermus parvatiensis* | 7B3N | C | 2020 | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid |
| *Thermus parvatiensis* | 7B3N | D | 2020 | Zn2+, Na+, Cl-, Glycerol, SO42-, Ethanesulfonic Acid |
| *Bacillus( Paenibacillus polymyxa)* | 1JWQ | A | 2001 | Zn2+ |
| *Clostridium difficile* | 5J72 | A | 2016 | Citric Acid, Zn2+, Ca2+, Cl-, Na+ |
| *Clostridium difficile* | 4RN7 | A | 2014 | Zn2+, Formic Acid, Glycerol, EPE (Ethanesulfonic Acid) |
| *Clostridium difficile* | 7RAG | B | 2021 | Zn2+, 1,2-Ethanediol |
| *Clostridium difficile* | 3QAY | A | 2011 | SO4-, Zn2+ |
| *Clostridium difficile* | 3QAY | B | 2011 | SO4-, Zn2+ |
| *Clostridium difficile* | 3QAY | C | 2011 | SO4-, Zn2+ |
| *Clostridium difficile* | 3QAY | D | 2011 | SO4-, Zn2+ |
| *Mycobacteroides abscessus* | 7AGO | A | 2020 | Zn2+, D-alpha-glutamine, Alanine |
| *Mycobacteroides abscessus* | 7AGL | A | 2020 | Zn2+ |
| *Mycolicibacterium smegmatis* | 7AGM | A | 2020 | Zn2+ |
| *Nostoc punctiforme* | 5EMI | A | 2015/2016 | MES, MRD, Zn2+ |
| *Staphylococcus aureus* | 7TJ4 | B | 2022 | Zn2+ |
| *Listeria monocytogenes* | 1XOV | A | 2004 | Cl- ,GLUTAMIC ACID, LYSINE, 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL, SO4-, Zn2+ |
| *Mycobacterium tuberculosis* | 4M6G | A | 2013 | Zn2+ |
| *Mycobacterium tuberculosis* | 4LQ6 | A | 2013 | SO4-, Zn2+, Cl-, Pt2+ |
| *Mycobacterium tuberculosis* | 4M6I | A | 2013 | Zn2+ |
| *Mycobacterium tuberculosis* | 4M6I | B | 2013 | Zn2+ |
| *Mycobacterium tuberculosis H37Rv* | 4M6H | A | 2013 | None reported |
| *Mycobacterium tuberculosis H37Rv* | 4M6H | B | 2013 | None reported |
| *Citrobacter freundii* | 2Y2B | A | 2010 | 2-(2-ACETYLAMINO-4-HYDROXY-6,8-DIOXA-BICYCLO[3.2.1]OCT- 3-YLOXY)-PROPIONIC ACID, L-ALA-GAMMA-D-GLU-MESO-DIAMINOPIMELIC ACID, Zn2+ |
| *Escherichia coli* | 2WKX | A | 2010 | Zn2+, GLYCEROL, Cl- |
| *Staphylococcus epidermidis* | 3LAT | B | 2010 | 1,4-BUTANEDIOL, IMIDAZOLE, Zn2+, Cl- |
| *Escherichia phage T7* | 1ARO | B | 1997 | Hg2+ |
| *Citrobacter freundii* | 1J3G | A | 2003 | ZINC ION |
| *Citrobacter freundii* | 2Y28 | A | 2011 | ZINC ION |
| *Citrobacter freundii* | 2Y2C | A | 2011 | None reported |
| *Citrobacter freundii* | 2Y2D | A | 2011 | ZINC ION |
| *Citrobacter freundii* | 2Y2E | A | 2011 | ZINC ION |
| *Escherichia coli K-12* | 2BH7 | A | 2006 | SULFATE ION; ZINC ION |
| *Escherichia coli str. K-12 substr. MG1655* | 3D2Y | A | 2009 | GLYCEROL |
| *Escherichia coli str. K-12 substr. MG1655* | 3D2Z | A | 2009 | CHLORIDE ION; ZINC ION |
| *Pseudomonas aeruginosa PAO1* | 4BJ4 | A | 2013 | CITRATE ANION |
| *Pseudomonas aeruginosa PAO1* | 4BOL | A | 2013 | ZINC ION; D-alanyl-N-[(2S,6R)-6-amino-6-carboxy-1-{[(1R)-1-carboxyethyl]amino}-1-oxohexan-2-yl]-D-glutamine |
| *Pseudomonas aeruginosa PAO1* | 4BPA | A | 2013 | ZINC ION |
| *Pseudomonas aeruginosa PAO1* | 4BXD | A | 2013 | ZINC ION |
| *Pseudomonas aeruginosa PAO1* | 4BXE | A | 2013 | DI(HYDROXYETHYL)ETHER; ACETATE ION |
| *Pseudomonas aeruginosa PAO1* | 4BXJ | A | 2013 | GLYCEROL |
| *Bacillus anthracis* | 1YB0 | A | 2005 | PHOSPHATE ION; ZINC ION |
| *Bacillus anthracis* | 2AR3 | A | 2006 | PHOSPHATE ION; ZINC ION |
| *Bacillus subtilis* | 3HMB | A | 2010 | ZINC ION |
| *Bacillus subtilis* | 3RDR | A | 2011 | CHLORIDE ION; ZINC ION |
| *Clostridium intestinale* | 6SSC | A | 2020 | ZINC ION; GLYCEROL; PHOSPHATE ION |
| *Clostridium intestinale URNW* | 6SRT | A | 2020 | PHOSPHATE ION; GLYCEROL; ZINC ION |
| *Clostridium perfringens str. 13* | 7F5I | A | 2022 | SODIUM ION; ZINC ION; GLUTAMIC ACID |
| *Staphylococcus aureus subsp. aureus NCTC 8325* | 4KNK | A | 2014 | ZINC ION; 1,2-ETHANEDIOL; DI(HYDROXYETHYL)ETHER; FORMIC ACID; SODIUM ION; IMIDAZOLE |
| *Staphylococcus aureus subsp. aureus NCTC 8325* | 4KNL | A | 2014 | N-acetyl-alpha-muramic acid; D(-)-TARTARIC ACID; (4S)-2-METHYL-2,4-PENTANEDIOL; OXAMIC ACID; FORMIC ACID; SODIUM ION; IMIDAZOLE |
| *Streptococcus pneumoniae TIGR4* | 4IVV | A | 2014 | 1,2-ETHANEDIOL; ZINC ION |
| *Streptococcus pneumoniae TIGR4* | 4X36 | A | 2015 | CHOLINE ION; ZINC ION; GLYCEROL |
| *Streptococcus pneumoniae TIGR4* | 5CTV | A | 2016 | None reported |
| *Bacillus phage Gamma* | 2L47 | A | 2011 | ZINC ION |
| *Escherichia phage T7* | 1LBA | A | 1994 | ZINC ION |
| *Staphylococcus phage G15* | 4OLS | A | 2014 | ZINC ION; FE (III) ION; MAGNESIUM ION |
| *Thermus phage 2119* | 6SU5 | A | 2020 | PHOSPHATE ION; ZINC ION; GLYCEROL |
| *Thermus phage 2631* | 6FHG | A | 2019 | ZINC ION |
| *Enterococcus phage phiM1EF22* | 7D55 | A | 2021 | None reported |
| *[Bacillus thuringiensis] serovar konkukian* | 2KQ8 | A | 2010 | None reported |
| *Clostridium perfringens* | 2KRS | A | 2010 | None reported |
| *Staphylococcus epidermidis* | 7KWI | A | 2021 | None reported |
| *Kayvirus* | 5O1Q | A | 2017 | None reported |
| *Staphylococcus phage 2638A* | 7AQH | A | 2020 | 4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID |
| *Staphylococcus phage G15* | 2MK5 | A | 2014 | None reported |
| *Bacillus subtilis* | 1X60 | A | 2005 | None reported |
| *Bacillus phage Gamma* | 2L48 | A | 2011 | None reported |
| *Kayvirus kay* | 4CSH | A | 2014 | SODIUM ION; ZINC ION; 2-(N-MORPHOLINO)-ETHANESULFONIC ACID; CALCIUM ION; GLYCEROL |
| *Kayvirus kay* | 4CT3 | A | 2014 | MERCURY (II) ION; METHYL MERCURY ION; CALCIUM ION; 4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID; GLYCEROL; CHLORIDE ION; S-(METHYLMERCURY)-L-CYSTEINE |
| *Staphylococcus phage G15* | 4OLK | A | 2014 | 2-[3-(2-HYDROXY-1,1-DIHYDROXYMETHYL-ETHYLAMINO)-PROPYLAMINO]-2-HYDROXYMETHYL-PROPANE-1,3-DIOL; CALCIUM ION |
| *Streptococcus pneumoniae* | 1GVM | A | 2002 | 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL; DECYLAMINE-N,N-DIMETHYL-N-OXIDE; CHOLINE ION |
| *Streptococcus pneumoniae* | 1H8G | A | 2002 | CHOLINE ION |
| *Streptococcus pneumoniae* | 1HCX | A | 2001 | DECYLAMINE-N,N-DIMETHYL-N-OXIDE; CHOLINE ION; 2,2':6',2''-TERPYRIDINE PLATINUM(II) Chloride |
| *Streptococcus pneumoniae* | 2BML | A | 2005 | HEXAETHYLENE GLYCOL; SULFATE ION; TETRAETHYLENE GLYCOL; DEXTROFLOXACINE; 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL |
| *Clostridium virus phiCD27* | 4CU5 | A | 2014 | None reported |
| *Clostridioides difficile* | 5OQ2 | A | 2017 | 1,2-ETHANEDIOL; PHOSPHATE ION |
| *Clostridioides difficile* | 5OQ3 | A | 2017 | DI(HYDROXYETHYL)ETHER; 1,2-ETHANEDIOL; CHLORIDE ION |
| *Staphylococcus aureus subsp. aureus Mu50* | 6FXO | A | 2019 | CHLORIDE ION |
| *Staphylococcus epidermidis* | 4EPC | A | 2012 | None reported |
| *Mycobacterium tuberculosis H37Rv* | 6EWY | A | 2018 | None reported |
| *Mycobacterium tuberculosis* | 3PBC | A | 2011 | None reported |
| *Mycobacterium tuberculosis* | 3S0Q | A | 2011 | None reported |
| *Mycobacterium tuberculosis* | 4Q4G | A | 2014 | None reported |
| *Mycobacterium tuberculosis* | 4Q4N | A | 2014 | None reported |
| *Mycobacterium tuberculosis H37Rv* | 2XIV | A | 2010 | GLYCEROL |
| *Mycobacterium tuberculosis H37Rv* | 3NE0 | A | 2010 | None reported |
| *Mycobacterium tuberculosis H37Rv* | 4Q4T | A | 2014 | GLYCEROL; FORMIC ACID |
| *Enterobacter lignolyticus* | 5VJ0 | A | 2017 | Protoporphyrin IX |
| *Staphylococcus phage 2638A* | 6YJ1 | A | 2020 | ZINC ION |
| *Bombus* | 5XZ4 | A | 2018 | 2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL; SULFATE ION |
| *Branchiostoma belcheri tsingtauense* | 4Z8I | A | 2015 | ZINC ION |
| *Branchiostoma belcheri tsingtauense* | 4ZXM | A | 2015 | None reported |
| *Camelus dromedarius* | 2R2K | A | 2007 | L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 2R90 | A | 2007 | None reported |
| *Camelus dromedarius* | 2Z9N | A | 2007 | L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3C2X | A | 2008 | SULFATE ION; L(+)-TARTARIC ACID; GLYCEROL |
| *Camelus dromedarius* | 3CG9 | A | 2008 | L(+)-TARTARIC ACID; alpha-L-rhamnopyranose |
| *Camelus dromedarius* | 3COR | A | 2008 | L(+)-TARTARIC ACID; 2-acetamido-2-deoxy-beta-D-galactopyranose |
| *Camelus dromedarius* | 3CXA | A | 2008 | L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3NG4 | A | 2010 | GLYCEROL; 2-acetamido-2-deoxy-beta-D-glucopyranose; S,R MESO-TARTARIC ACID |
| *Camelus dromedarius* | 3NNO | A | 2010 | L(+)-TARTARIC ACID; alpha-L-rhamnopyranose |
| *Camelus dromedarius* | 3NW3 | A | 2010 | LACTIC ACID; 2-acetamido-2-deoxy-alpha-D-glucopyranose; ALANINE; D-GLUTAMINE; GLYCEROL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3O4K | A | 2010 | L(+)-TARTARIC ACID; GLYCEROL; (2S)-1-({3-O-[2-(acetylamino)-4-amino-2,4,6-trideoxy-beta-D-galactopyranosyl]-alpha-D-glucopyranosyl}oxy)-3-(heptanoyloxy)propan-2-yl (7Z)-pentadec-7-enoate |
| *Camelus dromedarius* | 3OGX | A | 2010 | L(+)-TARTARIC ACID; GLYCEROL |
| *Camelus dromedarius* | 3QJ1 | A | 2011 | DI(HYDROXYETHYL)ETHER; GLYCEROL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3QS0 | A | 2011 | L(+)-TARTARIC ACID; GLYCEROL; 2-acetamido-2-deoxy-beta-D-glucopyranose |
| *Camelus dromedarius* | 3QV4 | A | 2011 | 1,2-ETHANEDIOL; S,R MESO-TARTARIC ACID; D-GLUTAMIC ACID; ALANINE; DI(HYDROXYETHYL)ETHER; GLYCEROL |
| *Camelus dromedarius* | 3RT4 | A | 2011 | L(+)-TARTARIC ACID; (R)-((2R,3S,4R,5R,6R)-3-HYDROXY-2-(HYDROXYMETHYL)-5-((R)-3-HYDROXYTETRADECANAMIDO)-6-(PHOSPHONOOXY)TETRAHYDRO-2H-PYRAN-4-YL) 3-HYDROXYTETRADECANOATE |
| *Camelus dromedarius* | 3T2V | A | 2011 | (2S,3R)-2-hexyl-3-hydroxynonanoic acid; GLYCEROL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3T39 | A | 2011 | (3R,4S,5R)-3,4,5-TRIHYDROXYCYCLOHEX-1-ENE-1-CARBOXYLIC ACID; S,R MESO-TARTARIC ACID; GLYCEROL |
| *Camelus dromedarius* | 3TRU | A | 2011 | (3R,4R)-3-[(1-carboxyethenyl)oxy]-4-hydroxycyclohexa-1,5-diene-1-carboxylic acid; GLYCEROL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 3UIL | A | 2012 | LAURIC ACID; GLYCEROL |
| *Camelus dromedarius* | 3UMQ | A | 2012 | GLYCEROL; butanoic acid |
| *Camelus dromedarius* | 3USX | A | 2012 | GLYCEROL; MYRISTIC ACID |
| *Camelus dromedarius* | 4FNN | A | 2012 | STEARIC ACID |
| *Camelus dromedarius* | 4GF9 | A | 2012 | GLYCEROL; (R)-((2R,3S,4R,5R,6R)-3-HYDROXY-2-(HYDROXYMETHYL)-5-((R)-3-HYDROXYTETRADECANAMIDO)-6-(PHOSPHONOOXY)TETRAHYDRO-2H-PYRAN-4-YL) 3-HYDROXYTETRADECANOATE; STEARIC ACID |
| *Camelus dromedarius* | 4OPP | A | 2014 | 11-cyclohexylundecanoic acid; GLYCEROL; L(+)-TARTARIC ACID; 2-acetamido-2-deoxy-beta-D-glucopyranose |
| *Camelus dromedarius* | 4ORV | A | 2014 | GLYCEROL; 2-acetamido-2-deoxy-beta-D-glucopyranose; L(+)-TARTARIC ACID; 7-phenylheptanoic acid |
| *Camelus dromedarius* | 4OUG | A | 2014 | (R)-((2R,3S,4R,5R,6R)-3-HYDROXY-2-(HYDROXYMETHYL)-5-((R)-3-HYDROXYTETRADECANAMIDO)-6-(PHOSPHONOOXY)TETRAHYDRO-2H-PYRAN-4-YL) 3-HYDROXYTETRADECANOATE; GLYCEROL; L(+)-TARTARIC ACID; PALMITIC ACID |
| *Camelus dromedarius* | 4Q8S | A | 2014 | GLYCEROL; L(+)-TARTARIC ACID; 2-acetamido-2-deoxy-beta-D-glucopyranose; 4-nitrophenyl hexadecanoate |
| *Camelus dromedarius* | 4Q9E | A | 2014 | L(+)-TARTARIC ACID; GLYCEROL; 4-nitrobenzaldehyde; 2-acetamido-2-deoxy-beta-D-glucopyranose |
| *Camelus dromedarius* | 5DWF | A | 2015 | GLYCEROL; 1,2-ETHANEDIOL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 5E0A | A | 2015 | L(+)-TARTARIC ACID; 2-acetamido-2-deoxy-beta-D-glucopyranose |
| *Camelus dromedarius* | 5E0B | A | 2015 | GLYCEROL; N-acetyl-beta-muramic acid; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 5XGY | A | 2017 | GLYCEROL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 6A89 | A | 2018 | alpha-D-ribofuranose; L(+)-TARTARIC ACID; GLYCEROL; 2,6-DIAMINOPIMELIC ACID; 1,2-ETHANEDIOL |
| *Camelus dromedarius* | 6IDM | A | 2018 | L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 7DY5 | A | 2021 | HEXANOIC ACID; GLYCEROL; CHLORIDE ION; 1,2-ETHANEDIOL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 7XFW | A | 2022 | GLYCEROL; HEXANOIC ACID; CHLORIDE ION; ACETATE ION; 1,2-ETHANEDIOL; L(+)-TARTARIC ACID |
| *Camelus dromedarius* | 7XFX | A | 2022 | HEXANOIC ACID; L(+)-TARTARIC ACID; GLYCEROL; ACETATE ION; CHLORIDE ION; 1,2-ETHANEDIOL |
| *Camelus dromedarius* | 7XFY | A | 2022 | HEXANOIC ACID; GLYCEROL; ACETATE ION; CHLORIDE ION; L(+)-TARTARIC ACID; 1,2-ETHANEDIOL |
| *Camelus dromedarius* | 7XU8 | A | 2022 | HEPTANOIC ACID; (4S)-2-METHYL-2,4-PENTANEDIOL; SODIUM ION; CARBONATE ION; HYDROGEN PEROXIDE; 1,2-ETHANEDIOL |
| *Drosophila melanogaster* | 1OHT | A | 2003 | L(+)-TARTARIC ACID; ZINC ION; 1,2-ETHANEDIOL |
| *Drosophila melanogaster* | 7NSX | A | 2021 | ZINC ION |
| *Drosophila melanogaster* | 7NSY | A | 2021 | None reported |
| *Drosophila melanogaster* | 7NSZ | A | 2021 | SODIUM ION; 4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID; ZINC ION |
| *Drosophila melanogaster* | 7NT0 | A | 2021 | GLCNAC(BETA1-4)-MURNAC(1,6-ANHYDRO)-L-ALA-GAMMA-D-GLU-MESO-A2PM-D-ALA; ZINC ION |
| *Homo sapiens* | 1SK3 | A | 2004 | SULFATE ION; NICKEL (II) ION |
| *Homo sapiens* | 1SK4 | A | 2004 | SODIUM ION |
| *Homo sapiens* | 1TWQ | A | 2004 | NICKEL (II) ION |
| *Homo sapiens* | 1YCK | A | 2005 | None reported |
| *Homo sapiens* | 2APH | A | 2006 | SULFATE ION |
| *Homo sapiens* | 2EAV | A | 2007 | NICKEL (II) ION |
| *Homo sapiens* | 2EAX | A | 2007 | None reported |
| *Alvinella pompejana* | 3EP1 | A | 2009 | None reported |
| *Arabidopsis thaliana* | 6ACV | A | 2018/2019 | DNA? |
| *Manduca sexta* | 6CKH | A | 2019 | None reported |
| *Clostridioides difficile* | 7ACV | A | 2022 | None reported |
| *Clostridioides difficile* | 7ACW | A | 2022 | None reported |
| *Clostridioides difficile* | 7ACX | A | 2022 | SULFATE ION; 2-acetamido-2-deoxy-beta-D-glucopyranose |
| *Clostridioides difficile* | 7ACX | B | 2022 | None reported |
| *Clostridioides difficile* | 7ACY | B | 2022 | SO42- |
| *Clostridioides difficile* | 7ACZ | B | 2022 | None reported |
| *Clostridioides difficile* | 5J6Q | A | 2017 | SO42-, Cl- |
| *Listeria phage* | 6HX0 | A | 2019 | None reported |
| *Geobacter sulfurreducens* | 3DP5 | A | 2008 | Heme C and SULFATE ION |
| *Mycolicibacterium smegmatis* | 4TM7 | A | 2015 | SULFATE ION, COPPE (II) ION, COPPER (I) ION, 1,2-ETHANDIOL, CHLORIDE ION |
| *Homo sapiens* | 4R7P | A | 2015 | UPG, SULFATE ION, 1,2-ETHANEDIOL, ACETATE ION |
| *Staphylococcus aureus* | 6G56 | A | 2018 | SULFATE ION, GLYCEROL |
| *Homo sapiens* | 4CC9 | C | 2013 | None reported |
| *Thermochaeotoides thermophila* | 6CfZ | F | 2018 | None reported |
| *Pseudomonas aeruginosa* | 1PEA | A | 1996 | ACETAMIDE |
| *Pseudomonas aeruginosa* | 1QNL | A | 1999 | ACETAMIDE |
| *Pseudomonas aeruginosa* | 1QO0 | A,B | 1999 | BUTYRAMIDE |

## Appendix B: Grouped multiple structural alignments of experimental structures with PDBeFOLD

i) PDBeFOLD and USAlign results for the uncropped 19 structures. Overall RMSD: 1.5525, overall Q-score: 0.1255.

|  |  |  |
| --- | --- | --- |
| PDB Structure | RMSD | Q-score |
| 1JWQ | 0.9407 | 0.6663 |
| 1XOV | 1.5939 | 0.3223 |
| 3CZX | 1.6146 | 0.5581 |
| 3NE8 | 1.029 | 0.5186 |
| 3QAY | 1.7692 | 0.54 |
| 4BIN | 0.813 | 0.3507 |
| 4LQ6 | 0.6333 | 0.5888 |
| 4M6G | 0.6425 | 0.5799 |
| 4M6H | 0.6263 | 0.6607 |
| 4M6I | 0.6243 | 0.6787 |
| 4RN7 | 0.713 | 0.6666 |
| 5EMI | 0.8254 | 0.6766 |
| 5J72 | 1.1552 | 0.1788 |
| 7AGL | 0.687 | 0.5871 |
| 7AGM | 0.6632 | 0.5729 |
| 7AGO | 0.6897 | 0.576 |
| 7B3N | 1.7739 | 0.5743 |
| 7RAG | 0.9918 | 0.5995 |
| 7TJ4 | 1.082 | 0.6586 |

ii) PDBeFOLD results for the cropped 19 candidate structures plus five cropped AmiA-like structures identified from initial protein structure database keyword search. Overall RMSD: 3.0837, overall Q-score: 0.0212.

|  |  |  |  |
| --- | --- | --- | --- |
| PDB Structure | Type | RMSD | Q-score |
| 1JWQ | Amidase\_3 | 1.1019 | 0.2855 |
| 1XOV | Amidase\_3 | 1.3471 | 0.2873 |
| 3CZX | Amidase\_3 | 2.2253 | 0.2201 |
| 3NE8 | Amidase\_3 | 1.1724 | 0.2396 |
| 3QAY | Amidase\_3 | 1.4097 | 0.2746 |
| 4BIN | Amidase\_3 | 1.1283 | 0.2491 |
| 4LQ6 | Amidase\_3 | 1.0965 | 0.2508 |
| 4M6G | Amidase\_3 | 1.0891 | 0.2512 |
| 4M6H | Amidase\_3 | 1.0529 | 0.2869 |
| 4M6I | Amidase\_3 | 1.0517 | 0.2952 |
| 4RN7 | Amidase\_3 | 1.2284 | 0.2838 |
| 5EMI | Amidase\_3 | 1.0970 | 0.3009 |
| 5J72 | Amidase\_3 | 1.2498 | 0.2716 |
| 7AGL | Amidase\_3 | 1.0451 | 0.2535 |
| 7AGM | Amidase\_3 | 1.0681 | 0.2523 |
| 7AGO | Amidase\_3 | 1.0429 | 0.2537 |
| 7B3N | Amidase\_3 | 1.4485 | 0.2783 |
| 7RAG | Amidase\_3 | 1.1393 | 0.2682 |
| 7TJ4 | Amidase\_3 | 1.4056 | 0.2718 |
| 2IM9 | AmiA-like | 3.9191 | 0.0717 |
| 2P1G | AmiA-like | 4.0545 | 0.0447 |
| 4H4J | AmiA-like | 3.8910 | 0.0916 |
| 4Q5K | AmiA-like | 4.0163 | 0.0876 |
| 4Q68 | AmiA-like | 4.0122 | 0.0885 |

## Appendix C: All validation scores for the multiple sequence alignment

**Heatmaps demonstrating A: the True Positive Rate (TPR) and B: the Positive Predictive value (PPV) for each pairwise alignment for 9 candidate AmiC structures**, where TPR is the ratio of sequence pairs aligned in both the sequences and structures compared to the total number of residues aligned in the structures, and PPV is the ratio of aligned sequence and structure residue pairs to the total number of residues aligned in the sequences.

**A**

A diagram of different numbers

Description automatically generated with medium confidence

**B**

A diagram of a number of numbers

Description automatically generated with medium confidence

## Appendix D: Full Logo for the final Multiple Sequence Alignment

**WebLogo displaying conservation across the entire multiple sequence alignment, including conserved and less-conserved regions.** X-axis represents column of alignment. Y-axis represents probability of observing residue in that column (always sums to 1.0).

A chart of letters and numbers

Description automatically generated

## Appendix E: Occupancy plots for each insertion region

Raw counts of per-species occupancy of regions I-1 to I-8, annotated by gram staining status. A=I-1, B=I-2, C=I-3, D=I-4, E=I-5, F=I-6, G=I-7, H=I-8.

**Bhhhhhhhhhhhhh**

**A**

A graph of a number of non-gap columns

Description automatically generated A graph of a number of non gap columns

Description automatically generated

**C**

A graph of a number of non-gap columns

Description automatically generated A graph of numbers and a number of non-gap columns

Description automatically generated

**D**

**F**

**E**

A graph of a number of non-gap columns

Description automatically generated A graph of different numbers and colors

Description automatically generated with medium confidence

A graph of a number of non-gap columns

Description automatically generated A graph of numbers and columns

Description automatically generated

**H**

**G**

## Appendix F: PCA scatter plots for components PC1/PC3 and PC2/PC3

A close-up of a computer screen

Description automatically generated

A white screen with a black border

Description automatically generated with medium confidence

## Appendix G: Representative phylogenetic tree

Representative phylogeny for each unique combination of fingerprints comprised of the 8 predicted insertion regions seen across the sequence dataset. From innermost circle (dark blue) to outermost circle: I1 presence, I2 presence, I3 presence, I4 presence, I5 presence, I6 presence, I7 presence, I8 presence, Phylum (colour legend for phylum is below the phylogenetic tree). Actual branch lengths are not represented to more clearly depict groupings of species.

A circular object with different colored lines

Description automatically generated



NOTES (to remove prior to submission, for input into discussion etc)

From Martin Steinegger talk 29/02/2024:

* Classifier trained on mills of protein sequences: random forest, used a large transformer to train on small number of data points (169)
* Combined with homology search to rank candidate proteins according to protein performance/function being similar to an input protein, even with low sequence identity
* (**ColabFold and ESM1b databases used, notebook available for homologous protein search – seekrank.steineggerlab.com) – pre print out soon 2024 ‘discovery of highly active…’)**
* Predicted structure databases released in 2022: Uniprot TrEMBL, ESMatlas-MGnify
* Isoform.io: Structure-guided isoform identification for the human transcriptome – predictable structures more likely to be real proteins than unpredictable structures (pLDDT: higher pLDDT = more likely the protein is real. Over 70 is pretty good.). Some isoforms have better structural prediction than MANE select, this website lists the 940 proteins this is the case for.
* Foldseek: 1D. Reduce 3D structure into 1D string. Each residue represented by a structural state letter. Describes tertiary interactions – the nearest 3D neighbour in space (simplified to the virtual centre of a residue, distance between these virtual centres measured), features extracted from the nearest neighbour (torsion angles and distances), discretise these different feature combinations by feeding into an encoder/decoder. Creates 20 different states. Called 3Di sequences.
  + Foldseek-TM: Improvement to Foldseek with higher precision/recall to TM-align?
  + ‘giant genes are rare but implicated in cell wall degradation by predatory bacteria’, west-roberts et al 2023, biorxiv pre-print (might be worth a read?)
* ‘Fast and accurate protein structure search with Foldseek’, van Kempen and Kim et al 2023 Nature biotech <https://www.nature.com/articles/s41587-023-01773-0>
* Clustering of AlphaFold TrEMBL with structures:
* A diagram of a protein universe using mms

  Description automatically generated
  + **Clustering predicted structures at the scale of the known protein universe, Barrio and Yeo et al** [**https://www.nature.com/articles/s41586-023-06510-w**](https://www.nature.com/articles/s41586-023-06510-w)
  + Evolutionary analysis, ‘dark clusters’ (no annotation in PDB, Pfam, TIGRfams) and new domain annotation from structural similarity!
    - **Lowest common ancestor analysis**: take tree of life and your cluster, and compute where in the tree these proteins in the cluster would meet. Quantifies how specific a cluster is (is it unique to one organism? One family? Etc.)
    - **Cluster.foldseek.com: Any uniport identifier, identify clusters (AFDB clusters)**
* Phylogeny: 3di can help with multiple sequence alignments! Pre-print here: (<https://www.biorxiv.org/content/10.1101/2023.12.12.571181v2.full.pdf>)
* Structure alignment: <https://github.com/steineggerlab/foldmason>
* About AlphaFold2 (for discussion/notes for viva):
  + Struggles with ‘orphan’ proteins with few close relatives, especially if there are known structures in the PDB, but can produce valid predictions if there are many relatives, few known structures. Also does not capture confirmational changes, so will not predict the active states of these proteins. It can’t model transmembrane orientations, ligand binding, post-translational modifications (AlphaFold3!). Is also not good for regions of disorder and low-confidence regions.
  + It is however good for single protein chains – if the protein is made from multiple domains connected by flexible linkers, the two domains in space may be arranged with low-confidence prediction (so will likely only be able to say with confidence which domains are attached to NAMLAA, but not with high certainty where they interact)
  + RMSD bigger than 2-3A suggests substantial difference; median for same protein = 0.6A, 1A average difference for alphafold vs experimental. (EBI AlphaFold training course). Low-confidence regions go to 2A or higher, so 2A+ indicates difference to the true experimental structures. Side chains may be lower confidence.
  + AlphaFold2 uses an MSA to generate a set of pair representations modelling relationship between every pair of amino acid residues in the input sequence. If two AA are in close contact, mutations in one will probably be followed by mutations in the other (co-evolution/covariation). And, if two regions change and evolve independently, they are likely not in direct contact. High quality MSA is vital for prediction.
  + Inputs: sequence, possibly also an MSA (likely would be good here because of low no. experimental structures in AlphaFold, a curated MSA would be good!) Can also supply template structures if needing a specific state
  + Measures of confidence: pLDDT and PAE.
    - pLDDT is local confidence, 0-100 scale, higher score=higher confidence; is how well the prediction would agree with an experimental structures, based on local carbon alpha distance difference. >90 = highest accuracy category, >70 indicates correct backbone and slight error with side chains. Low confidence comes from either the region being flexible/disordered, or lack of information to correctly predict the region. Might also coincide with missing residues in experimental structures. Supplied with mmcIF file on a per-residue basis in the B-factor field, structures can be coloured according to residue
      * MolProbity <http://molprobity.biochem.duke.edu/> (can flag regions of low confidence for further examination)
    - PAE is predicted aligned error – confidence in relative position of two residues in the predicted structure. A measure of confidence that the domain placement in the structure is correct. Low score=low error=higher confidence. Looking for ‘dark green’ on the PAE plot. Disordered regions may have low pLDDT and high PAE (because the local structure is uncertain, the position of the local structure to other structures is also uncertain). Supplied in JSON format, plot as PAE plot.
      * PAE viewer <https://subtiwiki.uni-goettingen.de/v4/paeViewerDemo> (can look specifically at multimers and the confidence of linker regions)
  + ColabFold predictions ranked by pLDDT and give a pLDDT plot scoring each amino acid position – plddt.png.
  + Reasons this is good: sequence searches can miss out related proteins because the sequences differ while structures tend to be more conserved and may have similar function.

Note to self: after draft submitted, go back to scripts and clean up code (don't forget to run black on each python file and do type hints/descriptions for each function! And also put code INTO functions! And combine the scripts better so you don't have a confusing list of scripts in the scripts file!)

**Need to also check I’ve cited all R and Python packages used**