NAR Genomics and Bioinformatics, 2024, **6**, lqae025
https://doi.org/10.1093/nargab/lqae025
Advance access publication date: 7 March 2024
Standard Article

Mining metagenomics data for novel bacterial nanocompartments

Naail Kashif-Khan [1,2], Renos Savva [2] and Stefanie Frank [1,*]

[1]Department of Biochemical Engineering, University College London, Bernard Katz Building, Gower Street, London WC1E 6BT, UK
[2]Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

[*]To whom correspondence should be addressed. Tel: +44 2076799567; Email: stefanie.frank@ucl.ac.uk

Abstract

Encapsulin nanocompartments are prokaryotic protein-based organelles. They display diverse natural functions, including mineral storage and stress response. Encapsulins also have applications in synthetic biology, drug delivery, vaccines, and metabolic engineering. Discovering novel encapsulins is challenging due to inconsistent annotations, and data contamination due to similarity with phage proteins. Previous studies have discovered thousands of encapsulin sequences from bacteria and archaea, but metagenomics databases were not specifically interrogated. Metagenomics can provide information on a much larger diversity of unculturable organisms and environmental samples than conventional sequencing experiments, and metagenomic protein databases have shed light on previously unexplored regions of the protein universe. This study leverages developments in deep learning for structure and function prediction, to produce a dataset of over 1300 novel putative encapsulin sequences from the MGnify Protein Database. Some well-known encapsulins and their cargo proteins were identified, predominantly peroxidases and ferritin-like proteins. A potentially novel encapsulin-associated biosynthetic gene cluster involved in producing cytotoxic or antimicrobial saccharides was discovered using biosynthetic gene cluster prediction. Finally, a cluster of predicted structures with novel features not seen in experimentally solved encapsulin structures was discovered using large-scale, deep learning-based structure prediction of putative metagenomic encapsulins.

Introduction

Encapsulin nanocompartments

Encapsulin nanocompartments are icosahedral protein-based organelles found in bacteria and archaea (1). Encapsulin organelles serve a wide range of physiological functions, including mineral storage (2), oxidative stress response (3), enzyme catalysis (4) and secondary metabolism (5,6) These protein nanostructures have many potential applications in synthetic biology and biomedicine, for example metal ion loading for use as imaging agents in biomedicine (7,8), antigen display for protein-based vaccines (9), as recently demonstrated with the surface display of SARS-CoV-2 antigens in animal models (10), and packaging of proteins and RNA towards drug delivery applications (11,12). Encapsulins may also be a promising platform for metabolic engineering via loading of heterologous enzymes; this approach may protect unstable proteins from degradation, increase reaction rates, and enable the use of reaction pathways with toxic intermediates (13).

Encapsulin monomers spontaneously self-assemble into full-sized capsids and are capable of encapsulating cargo proteins in a specific manner (as shown in Figure 1A). Encapsulin cargo proteins contain C-terminal cargo loading peptides (CLPs) (14) or longer N-terminal domains (NTDs) (4) responsible for targeting them to the capsid interior. Encapsulins display similar icosahedral symmetry to virus capsids and, as such, are assigned triangulation numbers (T-numbers) based on the number of subunits and size of the capsid assembly (Figure 1B). Encapsulin proteins share a common ancestor with HK97-fold phage major capsid proteins, and as such show sequence and structural similarity with this family of viral proteins (15). This shared evolutionary history makes discovery of encapsulins from protein sequence databases difficult, since encapsulin sequences are often misannotated as phage capsid proteins, bacteriocins or linocins (16), and search hits can be 'contaminated' with phage capsid proteins (5).

Encapsulins are currently grouped into four families based on their cargo type and Pfam annotation (Figure 1C) (5). Family 1 currently includes encapsulins from Pfam family PF04454 (Encapsulating Protein for Peroxidase). As the Pfam name suggests, these encapsulins are associated with cargo proteins from the dye-decolourizing peroxidase (DyP) family, or iron-binding cargo proteins like ferritins, rubrerythrin, hemerythrin, or manganese catalase-like proteins (5). Almost all experimentally solved encapsulin protein structures are derived from family 1. Family 2 is the largest encapsulin family, whose members are most often associated with four different types of cargo enzymes; these are cysteine desulfurase, polyprenyl transferase, xylulose kinase, and terpene cyclase. Family 2 encapsulins are not typically associated with any single Pfam family. Family 2 encapsulin capsid proteins can also be found fused to cyclic NMP-binding domains (5). Family 3 includes encapsulins from the Pfam family PF05065 (Phage capsid family), which are found within biosynthetic gene clusters (BGCs)—sets of genes encoding enzymes responsible for the synthesis of a variety of natural products. Finally, family 4 encapsulins are part of Pfam family

Received: October 16, 2023. **Revised:** January 5, 2024. **Editorial Decision:** February 19, 2024. **Accepted:** February 21, 2024

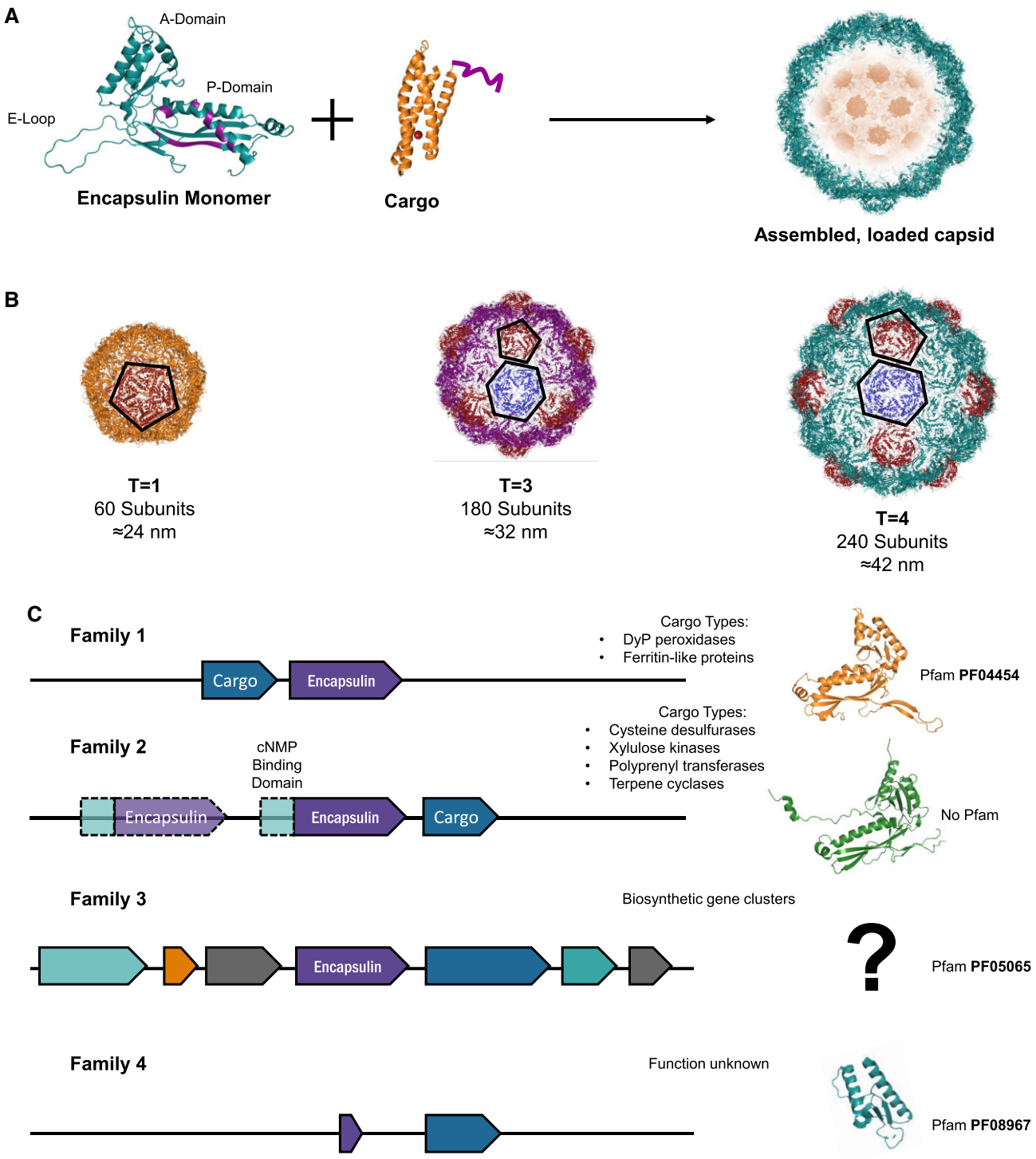

Figure 1. Encapsulin Structure and Function (**A**) Encapsulins spontaneously self-assemble from multiple copies of a single monomer. Cargo proteins contain peptide sequences (shown in purple) which target them to the interior of the capsid in a specific manner. PDB codes **6NJ8** and **6N63** for encapsulin and cargo respectively. (**B**) Encapsulins are assigned a triangulation number (T-number) which describes the icosahedral geometry of the capsid. The number of subunits is equal to the T-number multiplied by 60. T = 1 capsids are formed only of pentameric units whereas higher T-numbers make use of hexameric and pentameric units (outlined). (**C**) Encapsulins are classified into four families based on function. Family 1 encapsulins typically encapsulate peroxidases or ferritin-like proteins. Family 2 encapsulins can contain a cNMP-binding domain and are sometimes found as fusions of two encapsulin monomers. These encapsulins can contain four different enzyme types. Family 3 encapsulins are typically found within biosynthetic gene clusters, whilst family 4 encapsulins are truncated forms of the full HK97 fold.

PF08967 (DUF1884 domain-containing protein). These proteins display a truncated form of the HK97-fold containing only the A-domain. It is currently unknown whether these proteins are capable of self-assembly into an icosahedral particle like known encapsulins, or whether they encapsulate any cargo proteins. Despite this, previous work (5) has considered this small family of proteins as encapsulins and classified them based on the presence of putative 'cargo' proteins – hydrogenase, osmotic shock-associated proteins, deoxyribose phosphate aldolase, or glyceraldehyde-3-phosphate dehydrogenase.

Previous work and aims of this work

The most recent bioinformatics survey yielded a dataset of approximately 6000 encapsulin sequences (5). However, this work did not include metagenomics databases. These resources, compared to conventional genomic databases, contain large numbers of novel protein sequences and novel protein folds (17,18). Recent advances in deep learning have produced state-of-the-art bioinformatics tools for protein structure prediction (18,19), functional annotation (20), and biosynthetic gene cluster prediction (21). These tools are critical for understanding metagenomic proteins, which are otherwise difficult to analyse due to their diversity relative to annotated proteins in conventional databases.

Since encapsulins share significant sequence and structural similarity with HK97-fold phage capsid proteins, any bioinformatics search is likely to return many viral protein sequences. Previous work filtered out phage capsids by screening the genomic context of each search hit and removing any hits whose genome neighbourhood contained phage-associated proteins (5). However, when metagenomics databases are searched, the full genomic sequence for each returned candidate may be unavailable. Instead, only a single contig of variable length is available for each returned search sequence, and retrieving these contigs may require integrating data from different databases. In this case, an in-depth curation and filtering pipeline is required.

In light of this, the goal of this work was to utilize deep learning tools and bioinformatics to uncover novel encapsulin candidates with new structural and functional properties, specifically from metagenomics databases. Here, we describe how we have extended previous encapsulin discovery approaches (5) to public metagenomic data sources, by incorporating protein structure prediction and the use of predicted structure databases (18), combined with biosynthetic gene cluster prediction, and Pfam annotations generated using deep learning tools. Search hits were filtered to remove contaminating phage protein sequences using an in-depth curation pipeline which resulted in a dataset of 1326 novel putative encapsulin sequences. Importantly, many of the sequences we found share little or no sequence identity with currently known encapsulins and their predicted structures show novel conformational features. The new candidates we present here are expected to considerably expand the known array of biological functions observed in this class of protein nanocompartment, and will need to be experimentally validated.

Materials and methods

Search strategy

To discover novel encapsulin sequences, a combined sequence annotation and structure-based search approach was used. The 2022/05 release of the MGnify Protein Database (22) was filtered to recover all accessions with Pfam annotations from clan CL0373 (phage coat), which contains all HK97 fold-associated Pfam families. These annotations are generated using a convolutional deep neural network tool instead of the traditional Hidden Markov Model (HMM) method used by Pfam (23), which has been demonstrated to assign function more accurately in cases where sequence homology is remote or non-existent (20). In tandem, structure searches were performed against the 2023/02 release of ESM Atlas (18) using experimentally solved structures of the T = 1 encapsulin from *T. maritima*, the T = 3 encapsulin from *M. xanthus*, the T = 4 encapsulin from *Q. thermotolerans*, and the T = 1 encapsulin from *S. elongatus* (PDB codes 7MU1, 7S2T, 6NJ8 and 6X8M respectively). ESM Atlas structures with pTM scores of 0.7 and higher were downloaded using aria2c (24), and structure searches were performed using Foldseek (25) with the 'easy search' workflow and a minimum coverage of 0.5. These two searches gave an initial dataset of ∼800 000 sequences. Structure database search was carried out on VM.Standard.E4.Flex cloud instance with 64 cores and 1024GB of RAM (Oracle Cloud).

Removing phage-associated sequence contamination

Genomic contigs for each returned search sequence were retrieved using a combination of text-based filtering using bash, and API calls using Python (Supplementary Figure S1). To retrieve contigs for each returned search sequence, an MGYC contig accession was obtained from the MGnify Protein Database for each candidate. For each MGYC accession, a corresponding European Nucleotide Archive analysis accession (ERZ) and MGnify contig name was also retrieved from the MGnify Protein Database. Lastly, a MGnify analysis accession (MGYA) was obtained for each ERZ accession using the MGnify API. This API was also used to obtain protein coding sequences (CDS) for each search hit, using the hit's respective ERZ and MGYA accession and contig name. Any returned search sequence with missing MGYC accession contig CDS was removed from the dataset. These retrieval steps yielded a filtered dataset of ≈ 372000 putative encapsulin sequences with accompanying contig nucleotide sequences and CDS. Detailed metrics for the number of sequences removed at every filtering step are shown in Supplementary Figure S3.

As in previously published analyses (5), a custom mmseqs2 database (26) was prepared from two phage proteome datasets: one containing all proteins from Bacteriophage HK97, and one containing proteins from a broader set of prokaryotic tailed dsDNA viruses (UniProt proteome accessions UP000002576 and UP000391682, respectively). This database was searched using candidate contig protein sequences as query; searches were performed using mmseqs2 (26) with the iterative search function, a starting sensitivity of 4, a final sensitivity of 7, and 5 sensitivity steps (all further searches in this study used these parameters unless otherwise stated). Any candidates whose contigs contained mmseqs2 hits against these two phage proteomes were removed from

the dataset, leaving ≈340 000 putative encapsulin sequences whose contigs produced no matches against these phage proteomes.

Any returned search sequence with contigs under 25 kb in length or containing fewer than 10 protein sequences was then removed from the dataset—this was done to ensure that every candidate has enough genomic context available to confidently filter out phage-associated proteins and provide functional information for putative encapsulins. This drastically reduced the size of the dataset to 13 031 putative encapsulin sequences and associated contigs.

Next, every contig protein's Pfam annotations were retrieved from the MGnify Protein Database. These annotations were screened against a manually curated set of 279 phage-associated Pfam families, and any candidates whose contigs contained these proteins were removed from the dataset. This removed a substantial number of putative encapsulin hits from the dataset, yielding a filtered set of 1550 sequences.

Finally, to ascertain maximum sequence identity with proteins in conventional databases, putative encapsuling candidates were searched against the UniRef90 database (downloaded 2023-03-30) (27) using mmseqs2 with previously mentioned parameters, and '–max-accept' set to 1. UniRef90 database search was carried out on a VM.Standard.E4.Flex cloud instance with 64 cores and 1024GB of RAM (Oracle Cloud). The taxonomy ID of each encapsulin candidate was used to retrieve its taxonomic lineage using the UniProt API, and 2 encapsulin sequences showing >95% identity to UniRef90 sequences from the superkingdom 'viruses' were removed from the dataset.

Structure prediction and analysis

Where available, putative encapsulin structure predictions were retrieved from ESM Atlas using the public API. However, most candidate sequences had no available structure prediction data. For these candidates, structure prediction was carried out using ESMFold (18) in Google Colaboratory (28) with a chunk size of 64 for sequences larger than 700 amino acids, and 128 for sequences smaller than 700 amino acids. Structures for putative encapsulins longer than 900 amino acids were not predicted due to computational constraints (however, there were only 38 encapsulin sequences longer than 900 residues, see Supplementary Figure S2). Any structure predictions with a mean pLDDT value under 70 were removed from the dataset.

Confident predicted structures were analysed using DALI (29) to compute all-against-all pairwise Z-scores. Experimentally solved structures for four well-characterized encapsulin proteins were also included, from *T. maritima*, *M. xanthus*, *Q. thermotolerans* and *S. elongatus* (PDB codes 7MU1, 7S2T, 6NJ8 and 6X8M respectively). The similarity matrix was manually inspected, and a set of 130 structures showing extremely low similarity to all others were removed and manually assigned to their own dissimilar cluster. The remaining matrix was used as input for hierarchical clustering with complete linkage using the scipy.cluster.hierarchy package (30). The protein sequences within each cluster were then clustered at 80% sequence identity cutoff using mmseqs2 to reduce redundancy and facilitate easier manual inspection (26). Predicted structures for each cluster were visually inspected using PyMOL (31). All plots were created and inspected using the Plotly package in Python (32).

Representative sequences from each cluster of ESMFold predicted structures were also predicted using AlphaFold2 (19). This was done to demonstrate that structure prediction with ESMFold is comparable and does not lead to exclusion of encapsulin structures due to inferior performance. Structures were predicted with AlphaFold2 v2.3.0 using default MSA settings, and a maximum template cutoff date of 1 December 2023. For structure clusters with fewer than 15 sequences (after mmseqs2 clustering at 80% identity), all sequences were predicted. In the case of larger structure clusters, a single representative sequence was chosen based on the lowest mean DALI Z-Score to every other member of the cluster.

Encapsulin cargo type annotation

Initially, all contig protein Pfam annotations were manually inspected to assign encapsulin candidate cargo type and biological function. Two sets of Pfam annotations were considered in this study: Pfam family annotations from the MGnify Protein Database which are generated using ProtENN, a deep learning tool based on convolutional neural networks (33), and more conventional HMM-based Pfam assignments generated using HMMScan as part of DeepBGC (21). A comprehensive set of cargo types has previously been published (5), however that work did not assign every cargo type a Pfam family or set of Pfam families. For this current study, the known cargo Pfam families were enriched with further manually curated Pfam families (Supplementary Table S1), which were used to annotate some family 1, 2 and 4 encapsulin cargo proteins. However, since most putative encapsulins still had no family or cargo protein assigned, a more involved strategy was required.

Additional Family 1 cargo proteins were identified by searching an mmseqs2 database containing all contig proteins, using as query the family 1 cargo loading peptide (CLP) consensus sequences and secondary cargo CLP sequences (34) (Supplementary Data File S1). Search parameters in mmseqs2 were optimized for short query sequences by using the PAM30 Matrix, an E-value cutoff of 200000, and setting 'spaced-kmer-mode' to 0.

Additional Family 2 cysteine desulfurase (CyD) cargo protein candidates were identified using the same search parameters with a conserved motif (LARLANEFFS) found in the disordered N-terminal domain (NTD) of CyD from the *S. elongatus* family 2 encapsulin system (4). Further Family 2 cargo protein candidates were discovered using Hidden Markov Model (HMM)-based searches. For the four known cargo types (cysteine desulfurase, xylulose kinase, polyprenyl transferase, and terpene cyclase) sequence accessions were collected (5) and sequences retrieved from UniProt (27). Multiple sequence alignments (MSAs) for each cargo type were built using Clustal Omega with default parameters (35), and HMMs produced from these MSAs using the hmmbuild utility from HMMer with default settings (36). The hmmsearch utility from HMMer was used to search these profile HMMs against all putative cargo proteins and any hits with *E*-value <1 were reported.

Further cargo annotations were carried out using sequence similarity by searching all putative cargo proteins against the NCBI non-redundant protein database (37) using mmseqs2 with the previously mentioned parameters and –max-accept set to 30. NCBI non-redundant database search was carried

out on a VM.Standard.E4.Flex cloud instance with 64 cores and 1024GB of RAM (Oracle Cloud).

Biosynthetic gene cluster prediction

BGCs were predicted from putative encapsulin-containing metagenomic contigs using two different approaches. The antiSMASH 6.1.1 package (38) was used to predict BGCs from contigs with the following settings: Prodigal was used as the genefinding tool, ClusterBLAST was used with the general, subclusters and knownclusters settings, active site finder was enabled, and the pfam2go and clusterhmmer options were enabled. antiSMASH outputs in HTML format were parsed using the Beautifulsoup4 package in Python (39). In parallel, DeepBGC (21) was also used to predict BGCs, using Prodigal in metagenomic mode for gene finding, the 'deepbgc' detector, and classifiers 'product_class' and 'product_activity'. Predicted clusters were filtered to remove any clusters ending more than 10 kb upstream of the putative encapsulin gene or beginning 10 kb downstream of that gene.

Results

Putative encapsulin sequences from MGnify Protein database

Interrogation of the MGnify protein database returned 1548 putative encapsulin sequences filtered from ≈ 800000 initial hits, of which 1326 showed sub-95% identity with any sequence in UniRef90 (Figure 2A). Supplementary Figure S3 shows a detailed breakdown of the filtering pipeline and the number of sequences filtered at each step. Tracing the species of origin is non-trivial because these sequences come from unassembled contigs. However, biome annotations for every protein sequence in the MGnify database are provided. These annotations (Figure 2B and C) provide categorical information about metagenomics samples according to the Environment Ontology (40). Putative encapsulins were found in 2000 different metagenomics samples across diverse environments. Whilst most samples are sourced from aquatic environments, a sizeable proportion (29%) are associated with host-associated biomes, most of which are microbiota of the digestive systems of humans, other mammals, or birds. The presence of encapsulins in host-associated pathogens aligns with previous results (5) and may support the hypothesis that these proteins serve roles in bacterial pathogenicity (41).

Putative encapsulins are associated with biosynthetic gene clusters (BGCs)

Multiple techniques were used to identify associated cargo proteins corresponding to putative metagenomic encapsulins. These included sequence similarity searches, and functional annotation searches using two different methods: Profile Hidden Markov Model searches, and sequence searches using known encapsulin cargo loading peptide (CLP) sequence motifs. Despite this broad strategy, only 177 cargo protein candidates were identified for 1550 putative metagenomic encapsulins (Figure 3A). Most encapsulin cargo candidates belonged to encapsulin families 1 and 2 and included DyP-type peroxidases, ferritin-like domains, cysteine desulfurases, and polyprenyl transferases. Figure 3B shows the predicted structure of a representative putative encapsulin and its candidate corresponding ferritin-like cargo protein. Both show low sequence identity to any protein in conventional databases (46.2% and 32.7% for encapsulin and cargo) but are annotated as encapsulin and cargo by Pfam family. Predicted structures of these two proteins also corroborate their sequence-based functional annotations – the encapsulin clearly displays the HK97 fold while the cargo protein gives significant hits against crystal structures of ferritins from *E. coli* when searched with Foldseek (25). Interestingly, this encapsulin was annotated with Pfam PF05065 ('phage capsid family'), a label which was previously assigned to family 3 encapsulins (5) but which is seen here in a putative family 1 encapsulin system.

Family 3 encapsulins are found within biosynthetic gene clusters (BGCs), defined as groups of genes in close genomic proximity which encode pathways producing specialized products known as secondary metabolites (42). BGC prediction tools were used to identify putative encapsulin-associated BGCs from MGnify contig data (see Materials and methods).

These predictions uncovered a potentially novel encapsulin-associated BGC, the Saccharide BGC (Figure 4A). These putative BGCs are predicted by DeepBGC (21) to produce antimicrobial or cytotoxic saccharides, and all encode at least one glycosyl transferase enzyme, although most contain multiple such enzymes. Other enzymes that can be found in such BGCs include carbohydrate epimerases and dehydratases, methyltransferases, and oxygenases. Proteins from several Saccharide BGC systems were analysed using BLAST searches (37), ESMFold structure prediction (18), and Foldseek searches against the PDB. However, none of these searches gave any significant matches (*E*-value $< 10^{-3}$ for BLAST, or TM-score and probability >0.5 for Foldseek) to proteins of known structure or function, indicating a low degree of homology to proteins in these databases.

Several putative encapsulins were found within predicted BGCs with known cargo proteins that contain capsid targeting peptides or domains (Figure 4B). Several putative Saccharide BGCs also contains cysteine desulfurases, a known family 2 cargo. A putative encapsulin was found downstream of a putative hemerythrin cargo, but upstream of a polyketide synthase-like BGC, which encodes enzymes involved in the synthesis of chalcones. Putative encapsulins with ferritin cargos were also found in N-acetylglutaminyl glutamine (NAGGN) and non-ribosomal peptide synthetase (NRPS)-like clusters, both of whose general function is to synthesise short modified peptides (43–45).

The encapsulin-associated NAGGN BGC described here encodes the asparagine synthase and acetyltransferase enzymes needed to produce the osmoprotective peptide NAGGN (45). NRPS-like encapsulin BGCs encode the phosphate/AMP binding proteins usually associated with NRPS BGCs, but are missing the key peptidyl carrier protein (PCP), which is required for non-ribosomal peptide synthesis (46). However, such systems encode several potentially encapsulated enzymes, including gluconolactonases, thioesterases, aldo-keto reductases, and nitroreductases. Encapsulins have been previously reported as part of NRPS operons, but these partial 'NRPS-like' systems lacking a full complement of enzymes have not been seen previously.

Structural prediction of clusters with novel features

Following cargo annotation, protein structures of putative encapsulin hits were predicted using ESMFold, and confident structures were compared to each other using DALI. The

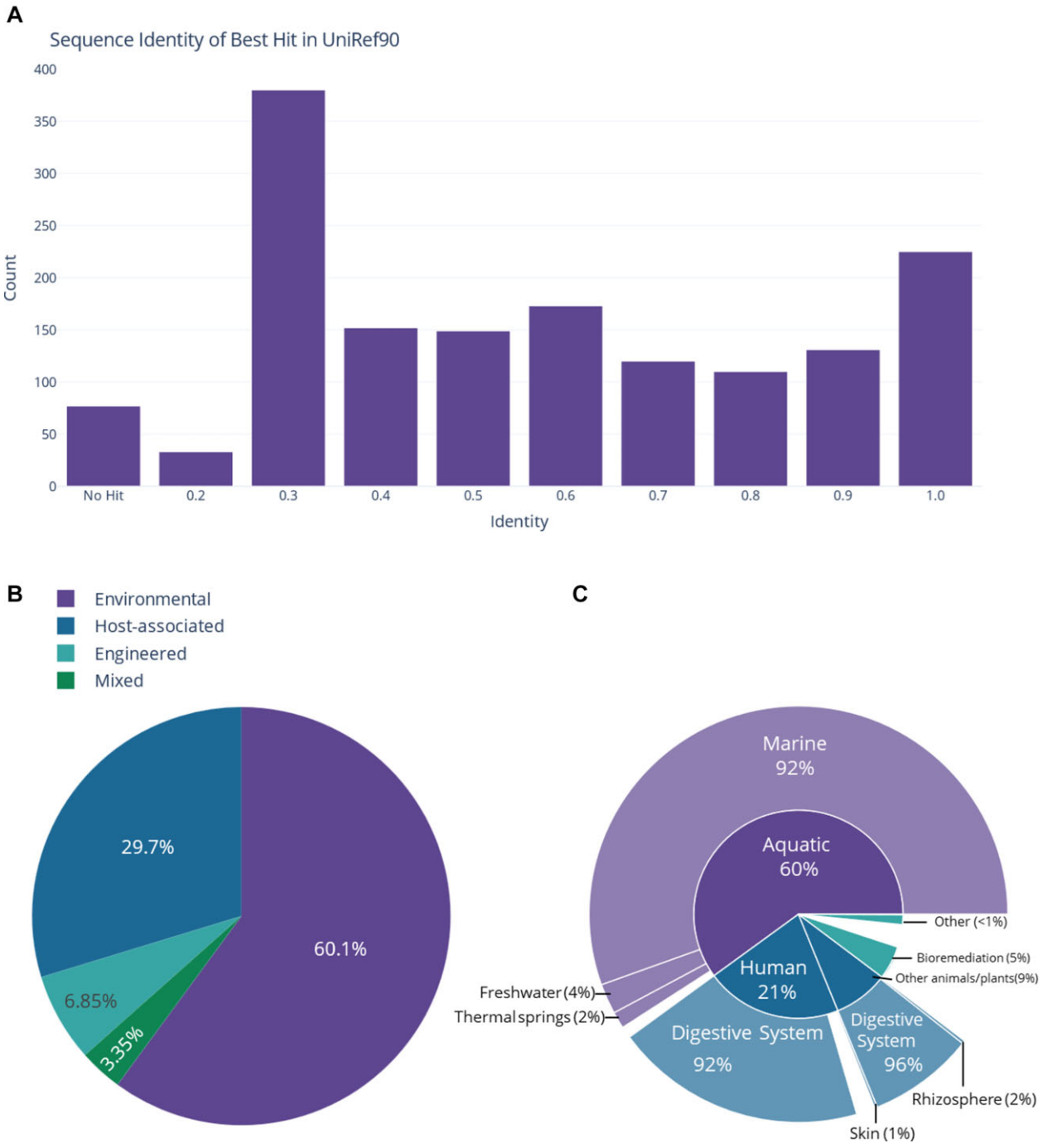

Figure 2. Putative metagenomic encapsulin sequences. (**A**) Histogram showing sequence identity of the closest match for 1550 putative metagenomic encapsulins when searched against UniRef90. X-axis values indicate the centre of each bin. 77 sequences showed no significant hits in Uniref90, and 225 sequences showed over 95% sequence identity to the best hit in UniRef90. (**B**) Breakdown of biome data for the 2000 metagenomic samples where putative encapsulin sequences were found. The four categories in the pie chart represent the four top-level categories of the Environment Ontology used in the MGnify database—environmental, host-associated, engineered and mixed. (**C**) Sunburst plot showing a breakdown of the biomes within each category from (**B**). 'Other animals' includes mammals and birds. 'Other' engineered types include laboratory samples, food production, fermented beverage production, and bioreactors/biogas sites. These categories contain many sparsely populated subcategories which are omitted for clarity. The 'Mixed' biome has no subcategories and is thus omitted.

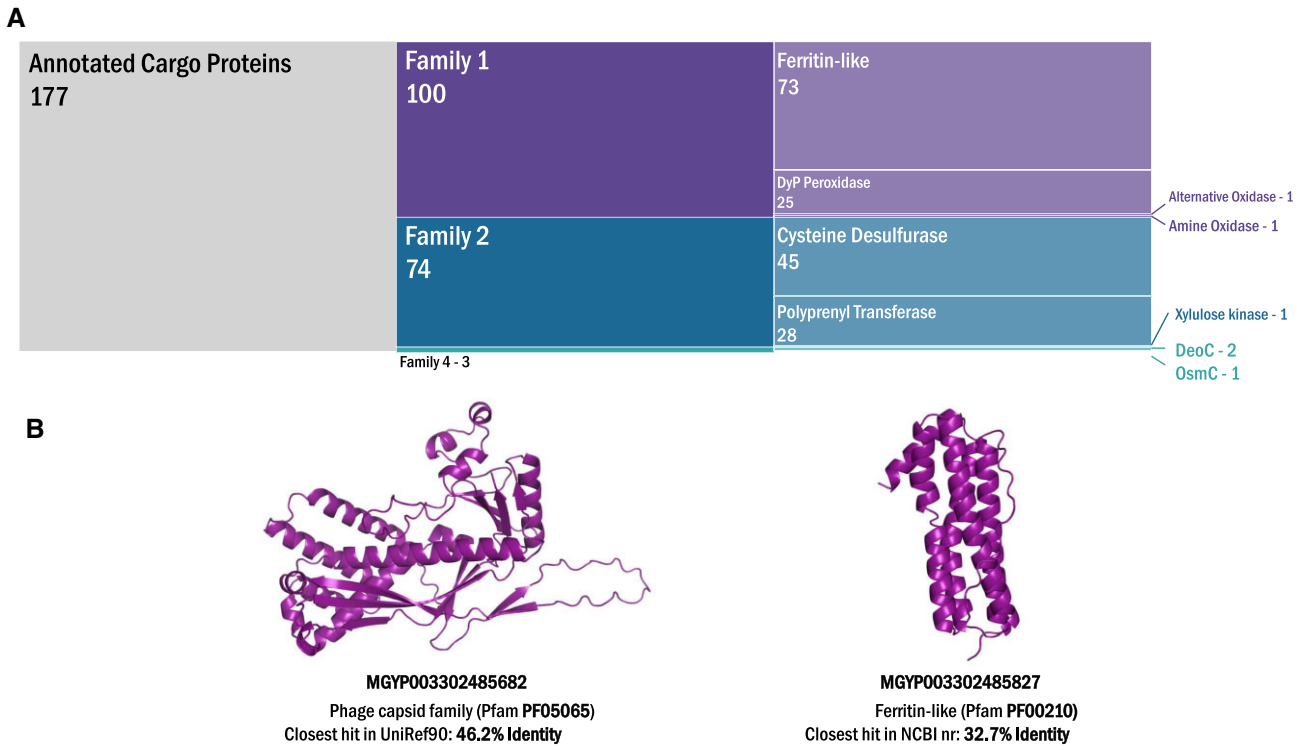

Figure 3. Metagenomic encapsulin cargo protein (**A**) Icicle plot showing a breakdown of annotated cargo proteins by family. Out of 1550 putative encapsulin hits, 177 were annotated with putative cargo proteins. Most cargo proteins were from families 1 and 2, with 3 proteins from family 4. (**B**) ESMFold predicted structures of a putative encapsulin hit and its respective ferritin-like cargo protein. Despite low sequence identity to any sequence in conventional databases, both are identified as encapsulin and ferritin-like proteins by functional annotation and predicted structures respectively.

resulting all-against-all similarity matrix (Figure 5A) showed clear patterns of clustered structures that share similarity with each other. These clusters also correspond to distinct regions of protein feature space (Figure 5B). Encapsulin hit sequences explore a wide range of lengths and charge properties, and in several cases, predicted encapsulin structures in the same cluster show similar length and isoelectric points. Notably, three out of the four experimental structures analysed here fall within the largest cluster of structures, suggesting that other clusters may contain predicted structures with novel features. Representatives from each cluster of ESMFold predicted structures were also predicted using AlphaFold2 (see Supplementary Figure S6), and across this set of representatives the two methods showed good agreement as measured by TM-Score and comparison of pLDDT values. This appears to rule out the possibility that results are influenced by artefacts from ESMFold structure prediction.

Predicted encapsulin structures from several clusters were manually inspected to reveal potentially novel features. As could be expected, the predicted structures from Cluster 8 (which contained 3/4 of the experimental structures analysed) all resembled known encapsulin structures from the literature (Figure 6A), with E-loops either in the 'T = 1-like' conformation or in a position resembling the T = 3 or T = 4 encapsulins. However, encapsulins from other clusters displayed some conformational diversity compared to experimentally solved structures—for example putative encapsulins from Cluster 1 showed insertions in the A-domain and E-loop. Interestingly, Cluster 6 encapsulin candidates all displayed several interesting features not seen in known encapsulins (Figure 6B). This included large insertion domains in the E-loop which could not be identified by sequence or structure searches against existing databases. These predicted structures also showed insertion of a small β-strand in the G-loop region, which is not seen in the known encapsulins. The positioning of these insertions in E- and G-loops indicates that these domains could decorate the outside of the assembled capsid shell. All predicted structures presented in Figure 6 showed acceptable confidence metrics (Supplementary Figure S5).

Discussion

In this work, a dataset of novel putative encapsulin sequences is presented, leveraging the rapid growth in metagenomic databases, and the wealth of new sequence diversity contained within them. This diversity presents many opportunities for discovery of novel proteins; however, it also brings to light several challenges which were encountered in this work. Analysing metagenomic encapsulin hits by functional prediction is a non-trivial task, as seen in the relatively slim proportion of candidate encapsulin sequences that could be annotated with a feasible cargo type. This could be because of low sequence identity (often sub-30%) of putative cargo proteins with any protein of known function, or it could also indicate that putative encapsulins in this dataset are associated with novel cargo proteins whose function has not previously been observed in known encapsulin systems. The scarcity of genomic context surrounding metagenomic encapsulin hits also makes removing phage proteins difficult, requiring a much more involved search and manually intensive filtering strategy compared to previous work (5). Many initial candidate sequences had to be removed due to small contig sizes (see Materials and methods), and contigs could not be retrieved

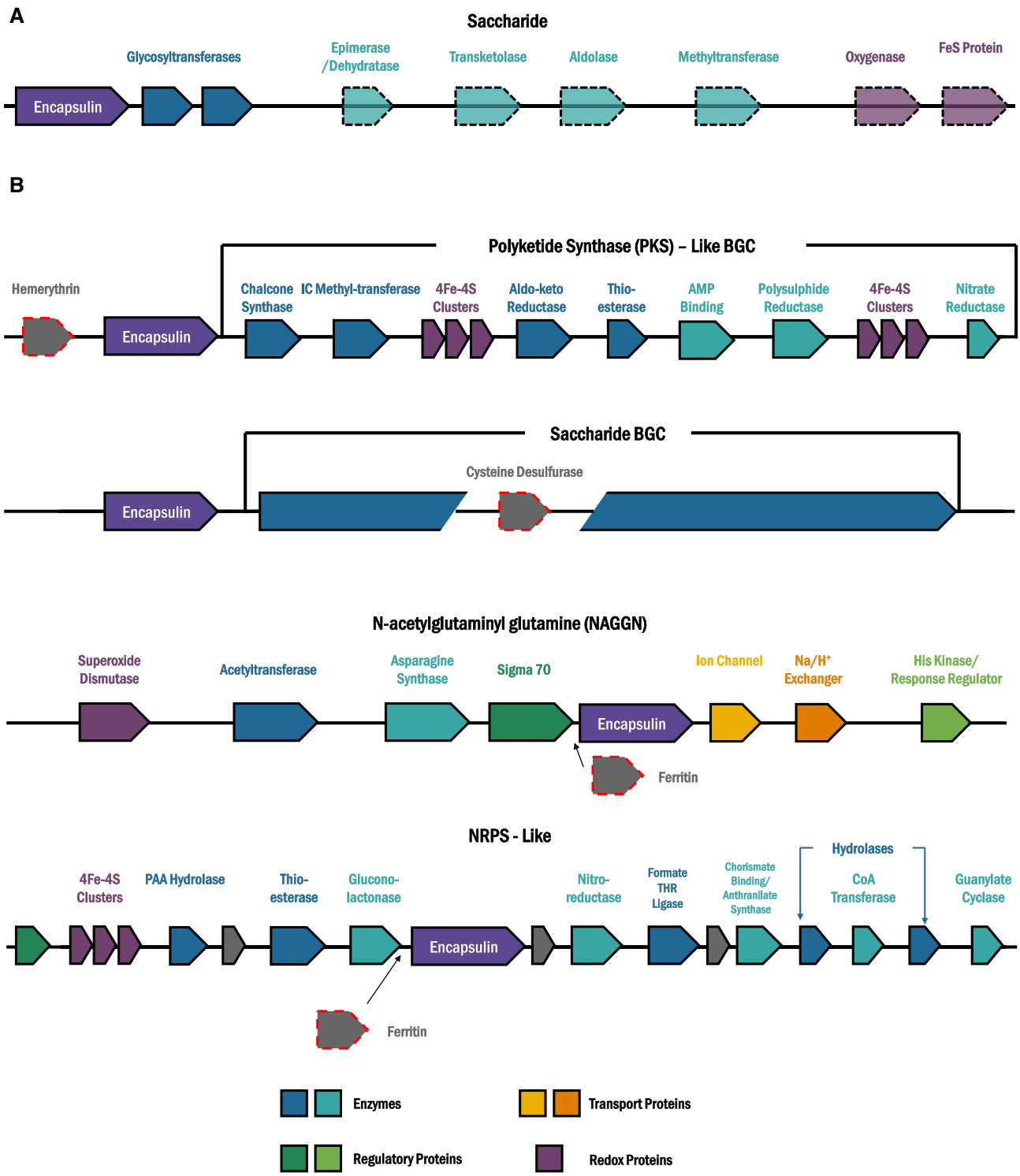

Figure 4. Putative encapsulin-associated biosynthetic gene clusters (BGCs). (**A**) 29 putative encapsulins are found in predicted saccharide BGCs, all containing at least one glycosyl transferase. These BGCs may also encode epimerases, aldolases, oxygenases, and redox proteins. (**B**) Some putative encapsulins are found in putative BGCs, and near known cargo proteins with loading peptides or domains (grey arrows with red dashed outline). In one example hemerythrin, a family 1 cargo protein, is seen upstream of the putative encapsulin and a polyketide synthase (PKS)-like BGC. Other encapsulin candidates are found upstream of saccharide BGCs containing cysteine desulfurase, a family 2 cargo. The *N*-acetylglutaminyl glutamine amide (NAGGN) BGC contains asparagine synthase and acetyltransferase enzymes, whilst non-ribosomal peptide synthetase (NRPS)-like clusters encode phosphate/AMP binding proteins. Both contain putative encapsulins with ferritin cargos. Enzymes not found in all operons are shown in dashed outline. Direction of arrows does not indicate gene orientation and is for schematic purposes only. PAA = phenylacetic acid, IC = isoprenylcysteine, TetR = tetracycline regulator

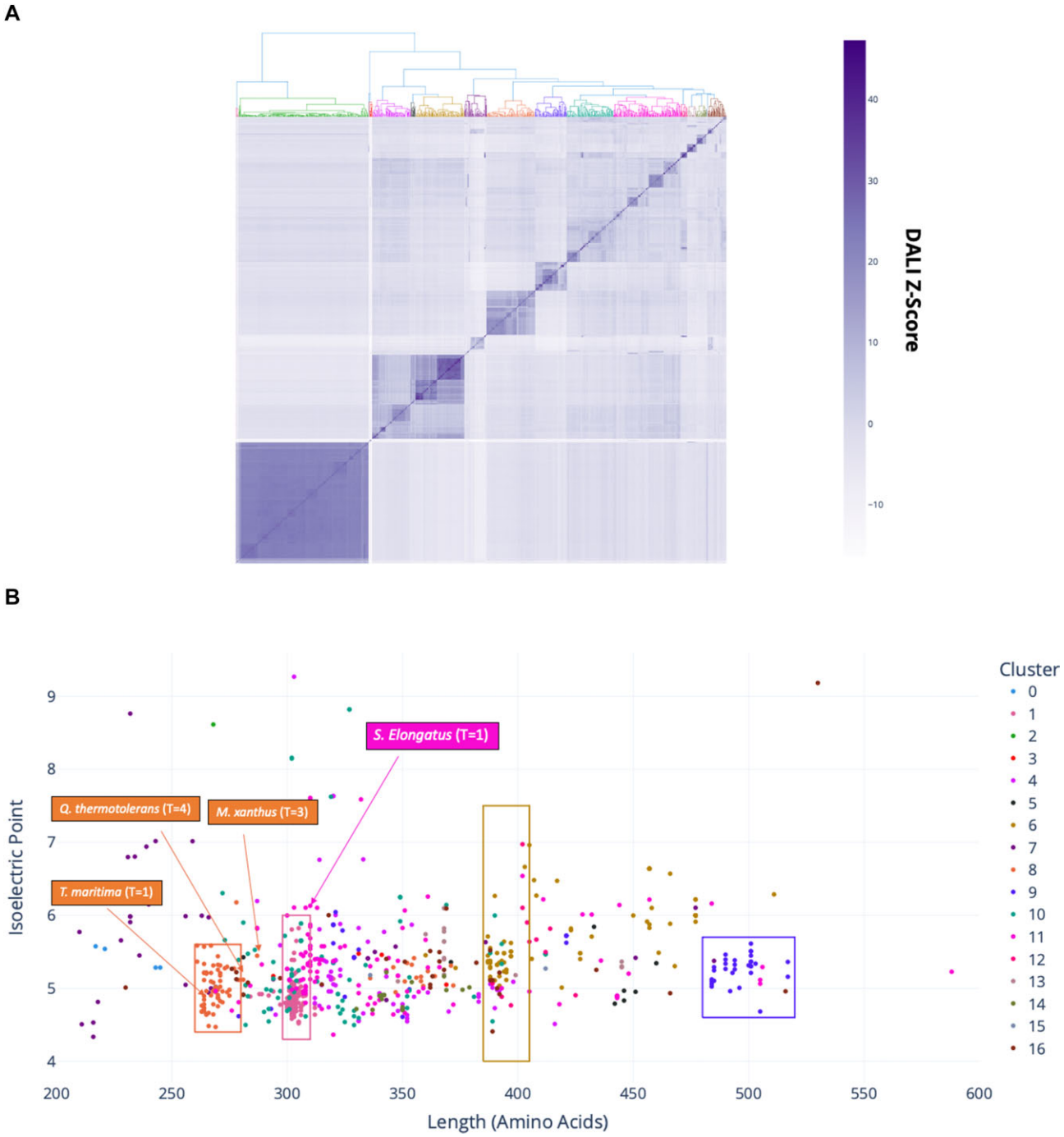

Figure 5. Predicted encapsulin structures form distinct clusters. (**A**) Clustered heatmap of DALI all-against-all pairwise similarity for high confidence encapsulin predicted structures. Clusters of structures sharing high similarity with each other can be seen visually and assigned using hierarchical clustering (see Methods). These clusters of similar structures are shown in a coloured dendrogram at the top of the heatmap. (**B**) Scatterplot of sequence length versus isoelectric point for predicted encapsulin structures, coloured by cluster. Whilst some clusters are relatively dispersed, there are some local regions where encapsulins of similar length and/or isoelectric point are clustered together (highlighted with coloured boxes). Experimentally solved encapsulin structures are shown, 3/4 of which fall within a single cluster.

for many candidate sequences due to missing metadata in the MGnify Protein Database (Supplementary Figures S1 and S3).

It must be noted that the Pfam annotations used in the initial search stage of this work were generated using ProtENN, a previously described deep learning model which has been demonstrated to be more accurate than conventional HMM-based approaches (33), particularly on sequences with remote homology to any existing annotated sequence. The authors of ProtENN suggest combining deep learning predictions with traditional HMM approaches for optimal performance and coverage. Unfortunately, re-annotating all 2.4 billion protein sequences using HMMs is far beyond the scope of this study, especially when ProtENN annotations are already provided, and are likely to be more accurate for the novel sequences sought after in this work. Aside from its use in the MGnify Protein Database, to our knowledge there are no examples

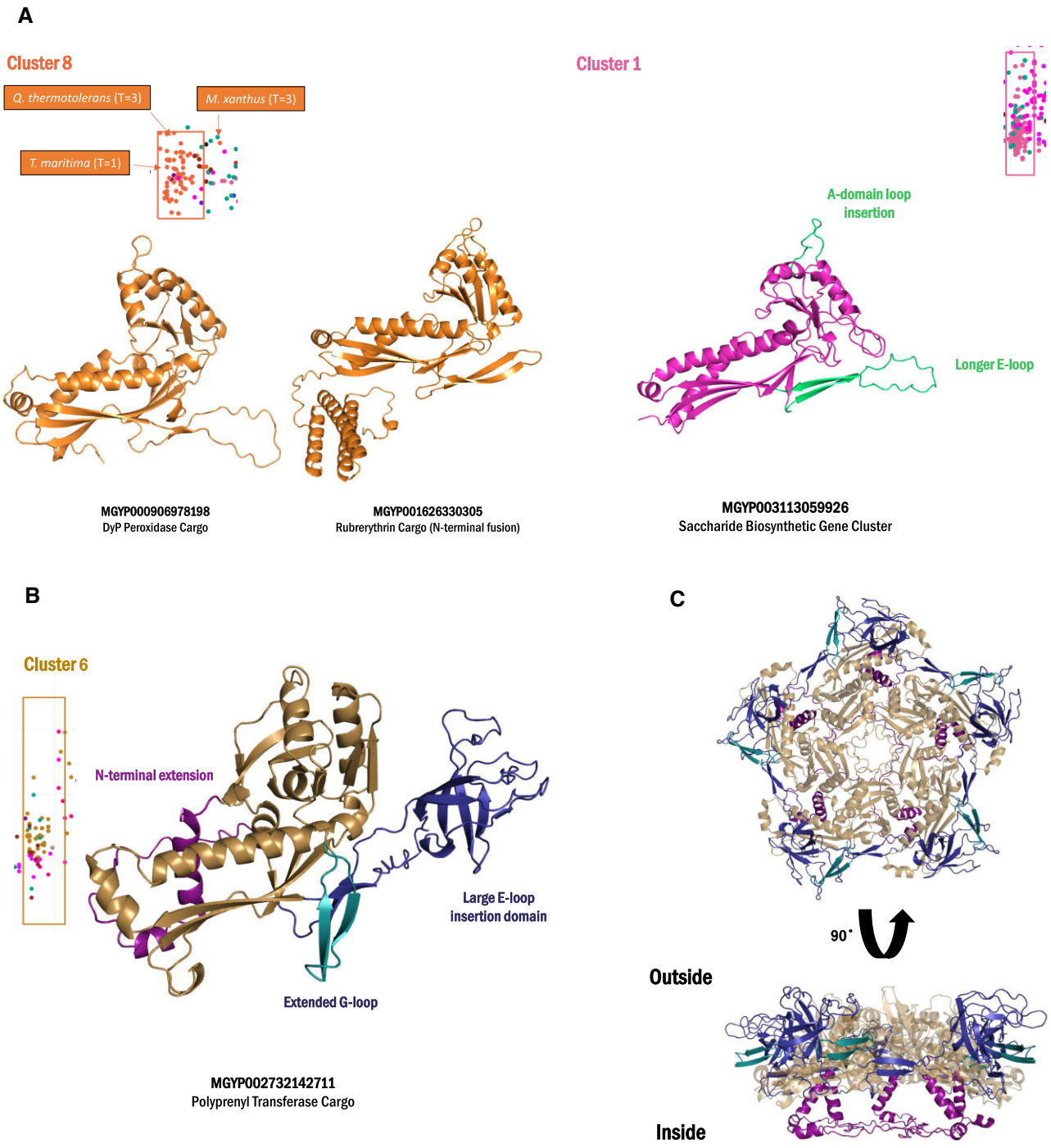

Figure 6. Predicted structures of putative encapsulins show novel features. (**A**) Cluster 8 contains predicted structures closest to experimentally solved encapsulin structures. Some resemble the *T. maritima* T = 1 encapsulin (left), while others have an E-loop angle closer to higher T-number encapsulins from *M. xanthus* and *Q. thermotolerans*. Cluster 1 (right) contains structures with minor variations, including insertion loops in the A-domain and a longer E-loop. (**B**) Cluster 6 contains novel encapsulin predicted structures with a large insertion domain in the E-loop (dark blue), and an extended G-loop (teal) not seen in experimental encapsulin structures. Some predicted structures contain N-terminal extensions or fusion domains (purple). (**C**) Alignment with the *T. maritima* encapsulin pentamer shows E-loop and G-loop extensions are predicted to decorate the outer surface of the capsid. Pentamer fitting is not energy minimized and shown for schematic purposes only.

in the literature of ProtENN being applied to metagenomics data. This work is the first study focusing on genome mining of the MGnify Protein Database, and as such the first detailed interrogation of ProtENN annotations applied to metagenomics data.

Despite the rigorous filtering strategy employed in this study, there is still a chance that some candidate encapsulins presented here are phage proteins and not encapsulin proteins of cellular origin. Encapsulins and phage capsid proteins from conventional databases can show as little as 20% sequence identity despite close homology in tertiary and quaternary structure (Supplementary Figure S4). The degree of uncertainty is compounded by the fact that the metagenomic candidate sequences presented here show low identity with sequences in conventional biological databases, for which species and functional annotations are available. This twofold issue of low sequence identity is then also complicated by the unclear evolutionary history of the HK97 fold: it is unknown whether this family of proteins originated in viruses or cellular organisms; both cases have been the subject of speculation in the literature (15,41). Due to this obscure evolutionary relationship, putative encapsulin hits may resemble phage capsid proteins in sequence or structure, however it is impossible to rule out the scenario that they are very primitive cellular proteins close to the HK97 fold's common ancestor. Genomic context can give clues as to a protein's origin, however in the metagenomic case this information is limited, and functional annotation of neighbouring genes is troublesome. Indeed, an ancient phage-like encapsulin sequence could be neighboured by primitive genes appearing viral in character, and with limited sequence identity and annotations it would be very difficult to decide whether these genes are cellular or viral in origin. These thought experiments serve to demonstrate the difficulty in distinguishing encapsulins from phage capsid proteins, and more broadly to discriminate between viral and cellular proteins. Notwithstanding the presence of such thought-provoking examples, this work strives to use all available information to rule out the presence of phage capsids where possible.

Biosynthetic gene cluster prediction revealed a potentially novel class of encapsulin-associated BGC, the Saccharide BGC. This may be an interesting new class of encapsulin system involved in producing cytotoxic or antimicrobial saccharides, as predicted by deep learning tools. The precise substrates and products of these saccharide pathways are not known. However, given the presence of putative encapsulins in these systems, and that the predicted product classes of these systems are antimicrobial/cytotoxic, it is assumed that these saccharide pathways produce toxic products or intermediates, hence the requirement for enzyme encapsulation. There are many known glycosylated cytotoxic natural products in bacteria (26), for example the substituted aminoglycoside pactamycin (47). However, it is important not to draw too strong a conclusion from the BGC prediction data; BGC prediction algorithms are notoriously error-prone and are known to produce many false positives (48). Indeed, a limitation was uncovered in this very study: antiSMASH falsely predicted around 80 'RiPP-like' BGCs (short for ribosomally synthesised and post-translationally modified product). The putative encapsulin genes in these BGCs are assigned the Pfam family PF04454 whose full name is 'Encapsulating protein for peroxidase'. However, antiSMASH incorrectly designates this Pfam family with the short name 'Linocin:M18' (49). Since linocin genes are usually found as part of real RiPP-like BGCs, antiSMASH falsely annotated these encapsulin-containing gene clusters as RiPP-like. This false annotation of encapsulin genes has been previously observed in the literature (16) and occurs in the Pfam database itself as well as in programs such as antiSMASH which make use of its functionality.

Putative cargo proteins from several Saccharide BGC examples returned no informative hits when searched using BLAST, or when predicted structures were searched against the PDB using Foldseek. The few significant (E-value $< 10^{-3}$) sequence hits from BLAST were all hypothetical, uncharacterized proteins with no annotated function.Structure hits only showed insignificant structural similarity over small regions (TM-scores and probabilities below 0.5). The number and accuracy of BGC predictions is limited in this case by the genomic context available in the contigs surrounding each candidate encapsulin, which explains the relatively few BGC predictions observed in this study. Such tools are usually intended to be run on full genome sequences. BGC prediction tools also make use of Pfam and other functional annotations, which have their own limitations with metagenomic proteins as previously mentioned. Given the limitations in the underlying data and the tools used, Saccharide BGCs remain a hypothetical new biological function for encapsulins until more rigorous experimental analysis can be carried out.

The same limited conclusions can be drawn from the observation of known family 1 or 2 cargos within other types of BGC, including Saccharide BGCs. The limitations of the data presented here indicate that this is simply a coincidence, however an encapsulin and its associated cargo forming part of a larger cluster of metabolic genes is an interesting possibility. It is speculated that if this were to be observed in a more significant number of genomes or metagenomic contigs, this could have implications for encapsulated ferritin or cysteine desulfurase function as part of a larger cluster of genes involved in secondary product metabolism. It is noted that BGC prediction in the context of encapsulins has not previously been carried out on as large a scale as in this work, and such approaches could be applied to the existing encapsulin datasets to potentially give new insights into biological function.

Predicted structures of putative encapsulin hits reveal some interesting new structural features. Whilst many of these predicted structures show similar topology to experimentally resolved encapsulin structures, the predicted structures from Cluster 6 display a set of novel structural features compared to known encapsulins. Insertions in the E-loop and A-domain may decorate the vertices of pentameric units in the capsid shell, and it is speculated that these insertion domains may lead to architectural differences in these putative capsids compared to experimental encapsulin structures. These architectural differences, if experimentally confirmed, could impart new physicochemical properties which may aid in encapsulin engineering. E-loop insertion domains may also confer new biological function to the encapsulin monomer, however sequence- or structure-based search failed to reveal any informative significant hits against this curious new structural region. Some encapsulins in this cluster also show N-terminal fusion domains, although again sequence- or structure-based search fails to shed light on the function of these domains. These could be cargo protein fusions, or they could be involved in protein–protein interactions with a separately encoded cargo protein. Lastly, E-loop and G-loop insertion regions could serve as useful functionalization sites for the ex-

terior of the capsid shell, which may have biotechnology applications in vaccine development.

To conclude, this study presents exploratory work towards discovering new encapsulin sequences in metagenomic databases, and the workflows required to filter and analyse these sequences. These new data may be useful in understanding encapsulin biology and/or in developing new engineering applications for encapsulins, however experimental characterization is now needed to further understand these potentially novel protein nanocompartments.

Data availability

All raw data and scripts needed to reproduce this work (including Python and bash scripts, Jupyter notebooks, novel putative encapsulin sequences, cargo protein sequences, MGnify Protein Database and European Nucleotide Archive accessions, and contig sequences) are available at Zenodo: https://doi.org/10.5281/zenodo.8183050 and GitHub: https://github.com/naailkhan28/encapsulin_metagenomics.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We warmly thank Dr Alexander Van De Steen and Ferdinando Sereno for helpful discussions and input on this work. We also thank Dr Ryan Payton, Mike Riley, James Fleming, Dave Fuller, and Dr Dave Houldershaw for their advice and support on computational aspects of this work. We gratefully acknowledge Prof Mark Williams and Dr Katherine Thompson for valuable oversight and guidance.

Funding

N.K.K. is supported by the Biotechnology and Biological Sciences Research Council [BB/T008709/1], with the London Interdisciplinary Biosciences Consortium Doctoral Training Partnership; Oracle Cloud credits and related resources provided by Oracle for Research via R.P., M.R., J.F. and D.F. (in part); EPSRC for funding SF [EP/R013756/1] through the Future Vaccine Manufacturing Research Hub (Vax-Hub).

Conflict of interest statement

None declared.

References

1. Sutter,M., Boehringer,D., Gutmann,S., Günther,S., Prangishvili,D., Loessner,M.J., Stetter,K.O., Weber-Ban,E. and Ban,N. (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat. Struct. Mol. Biol.*, **15**, 939–947.
2. Ross,J., McIver,Z., Lambert,T., Piergentili,C., Bird,J.E., Gallagher,K.J., Cruickshank,F.L., James,P., Zarazúa-Arvizu,E., Horsfall,L.E., *et al.* (2022) Pore dynamics and asymmetric cargo loading in an encapsulin nanocompartment. *Sci. Adv.*, **8**, eabj4461.
3. McHugh,C.A., Fontana,J., Nemecek,D., Cheng,N., Aksyuk,A.A., Heymann,J.B., Winkler,D.C., Lam,A.S., Wall,J.S., Steven,A.C., *et al.* (2014) A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J.*, **33**, 1896–1911.
4. Nichols,R.J., LaFrance,B., Phillips,N.R., Radford,D.R., Oltrogge,L.M., Valentin-Alvarado,L.E., Bischoff,A.J., Nogales,E. and Savage,D.F. (2021) Discovery and characterization of a novel family of prokaryotic nanocompartments involved in sulfur metabolism. *eLife*, **10**, e59288.
5. Andreas,M.P. and Giessen,T.W. (2021) Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nat. Commun.*, **12**, 4748.
6. Gorges,J., Panter,F., Kjaerulff,L., Hoffmann,T., Kazmaier,U. and Müller,R. (2018) Structure, total synthesis, and biosynthesis of chloromyxamides: myxobacterial tetrapeptides featuring an uncommon 6-chloromethyl-5-methoxypipecolic acid building block. *Angew. Chem. Int. Ed Engl.*, **57**, 14270–14275.
7. Sigmund,F., Berezin,O., Beliakova,S., Magerl,B., Drawitsch,M., Piovesan,A., Gonçalves,F., Bodea,S.-V., Winkler,S., Bousraou,Z., *et al.* (2023) Genetically encoded barcodes for correlative volume electron microscopy. *Nat. Biotechnol.*, **41**, 1734–1735.
8. Sigmund,F., Pettinger,S., Kube,M., Schneider,F., Schifferer,M., Schneider,S., Efremova,M.V., Pujol-Martí,J., Aichler,M., Walch,A., *et al.* (2019) Iron-sequestering nanocompartments as multiplexed electron microscopy gene reporters. *ACS Nano*, **13**, 8114–8123.
9. Lagoutte,P., Mignon,C., Stadthagen,G., Potisopon,S., Donnat,S., Mast,J., Lugari,A. and Werle,B. (2018) Simultaneous surface display and cargo loading of encapsulin nanocompartments and their use for rational vaccine design. *Vaccine*, **36**, 3622–3628.
10. Khaleeq,S., Sengupta,N., Kumar,S., Patel,U.R., Rajmani,R.S., Reddy,P., Pandey,S., Singh,R., Dutta,S., Ringe,R.P., *et al.* (2023) Neutralizing efficacy of encapsulin nanoparticles against SARS-CoV2 variants of concern. *Viruses*, **15**, 346.
11. Kwon,S. and Giessen,T.W. (2022) Engineered protein nanocages for concurrent RNA and protein packaging In Vivo. *ACS Synth. Biol.*, **11**, 3504–3515.
12. Van de Steen,A., Khalife,R., Colant,N., Mustafa Khan,H., Deveikis,M., Charalambous,S., Robinson,C.M., Dabas,R., Esteban Serna,S., Catana,D.A., *et al.* (2021) Bioengineering bacterial encapsulin nanocompartments as targeted drug delivery system. *Synth. Syst. Biotechnol.*, **6**, 231–241.
13. Lau,Y.H., Giessen,T.W., Altenburg,W.J. and Silver,P.A. (2018) Prokaryotic nanocompartments form synthetic organelles in a eukaryote. *Nat. Commun.*, **9**, 1311.
14. Altenburg,W.J., Rollins,N., Silver,P.A. and Giessen,T.W. (2021) Exploring targeting peptide-shell interactions in encapsulin nanocompartments. *Sci. Rep.*, **11**, 4951.
15. Krupovic,M. and Koonin,E.V. (2017) Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E2401–E2410.
16. Jones,J.A. and Giessen,T.W. (2021) Advances in encapsulin nanocompartment biology and engineering. *Biotechnol. Bioeng.*, **118**, 491–505.
17. Ovchinnikov,S., Park,H., Varghese,N., Huang,P.-S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. and Baker,D. (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.
18. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic level protein structure with a language model. *Science*, **379**, 1123–1130.
19. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Z'idek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
20. Bileschi,M., Belanger,D., Bryant,D., Sanderson,T., Carter,B., Sculley,D., DePristo,M. and Colwell,L. (2019) Using deep learning to annotate the protein universe. *Nat. Biotechnol.*, **40**, 932–937.
21. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D., *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, e110.
22. Richardson,L., Allen,B., Baldi,G., Beracochea,M., Bileschi,M.L., Burdett,T., Burgin,J., Caballero-Pérez,J., Cochrane,G., Colwell,L.J.,

et al. (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, **51**, D753–D759.
23. 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.
24. Release aria2 1.36.0. https://github.com/aria2/aria2, (1 March 2023, date last accessed).
25. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2023) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
26. Mirdita,M., Steinegger,M. and S"oding,J. (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, **35**, 2856–2858.
27. The UniProt Consortium (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
28. Google Colaboratory. https://colab.research.google.com/, (16 February 2023,date last accessed).
29. Holm,L. (2022) Dali server: structural unification of protein families. *Nucleic Acids Res.*, **50**, W210–W215.
30. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
31. Schrödinger,L.L.C. (2015) The PyMOL Molecular Graphics System. Version 1.8.
32. Plotly Technologies Inc. (2015) Collaborative data science. https://plot.ly (1 January 2024, date last accessed).
33. Bileschi,M.L. and Colwell,L.J. (2022) Using deep learning to annotate the protein universe. https://blog.research.google/2022/03/using-deep-learning-to-annotate-protein.html (1 January 2024,date last accessed).
34. Giessen,T.W. and Silver,P.A. (2017) Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nat. Microbiol.*, **2**, 17029.
35. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
36. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
37. 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.
38. Blin,K., Shaw,S., Kloosterman,A.M., Charlop-Powers,Z., van Wezel,G.P., Medema,M.H. and Weber,T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.*, **49**, W29–W35.
39. Richardson,L. (2023) beautifulsoup4: screen-scraping library. https://beautiful-soup-4.readthedocs.io/en/latest/, (16 February 2023, date last accessed).
40. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.*, **7**, 57.
41. Giessen,T.W. (2022) Encapsulins. *Annu. Rev. Biochem.*, **91**, 353–380.
42. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C., *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
43. Arulprakasam,K.R. and Dharumadurai,D. (2021) Genome mining of biosynthetic gene clusters intended for secondary metabolites conservation in actinobacteria. *Microb. Pathog.*, **161**, 105252.
44. Gulick,A.M. (2017) Nonribosomal peptide synthetase biosynthetic clusters of ESKAPE pathogens. *Nat. Prod. Rep.*, **34**, 981–1009.
45. Sagot,B., Gaysinski,M., Mehiri,M., Guigonis,J.-M., Le Rudulier,D. and Alloing,G. (2010) Osmotically induced synthesis of the dipeptide N-acetylglutaminylglutamine amide is mediated by a new pathway conserved among bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12652–12657.
46. Corpuz,J.C., Sanlley,J.O. and Burkart,M.D. (2022) Protein-protein interface analysis of the non-ribosomal peptide synthetase peptidyl carrier protein and enzymatic domains. *Synth. Syst. Biotechnol.*, **7**, 677–688.
47. Eida,A.A. and Mahmud,T. (2019) The secondary metabolite pactamycin with potential for pharmaceutical applications: biosynthesis and regulation. *Appl. Microbiol. Biotechnol.*, **103**, 4337–4345.
48. Prihoda,D., Maritz,J.M., Klempir,O., Dzamba,D., Woelk,C.H., Hazuda,D.J., Bitton,D.A. and Hannigan,G.D. (2021) The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Nat. Prod. Rep.*, **38**, 1100–1108.
49. 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.

Received: October 16, 2023. **Revised:** January 5, 2024. **Editorial Decision:** February 19, 2024. **Accepted:** February 21, 2024