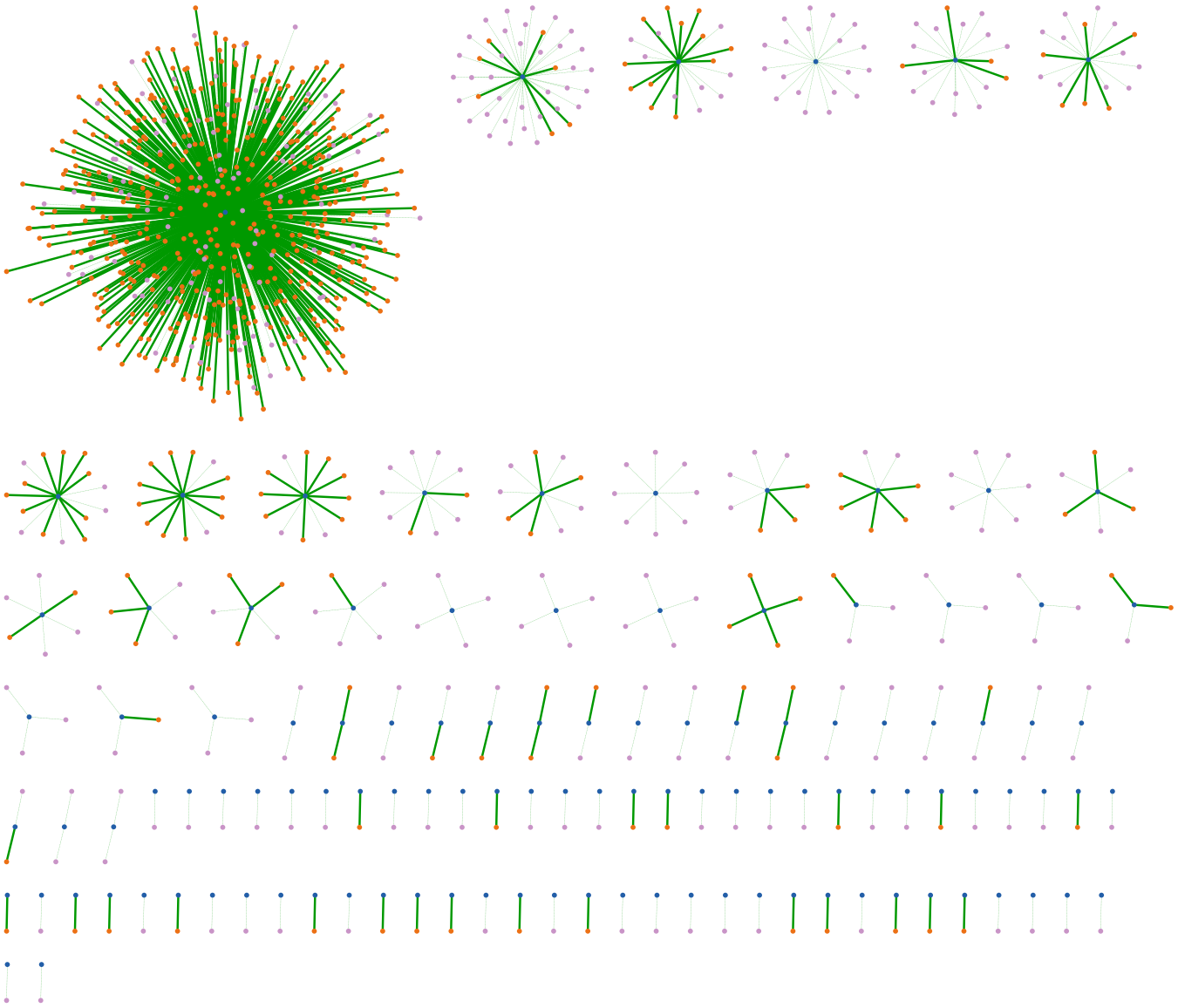
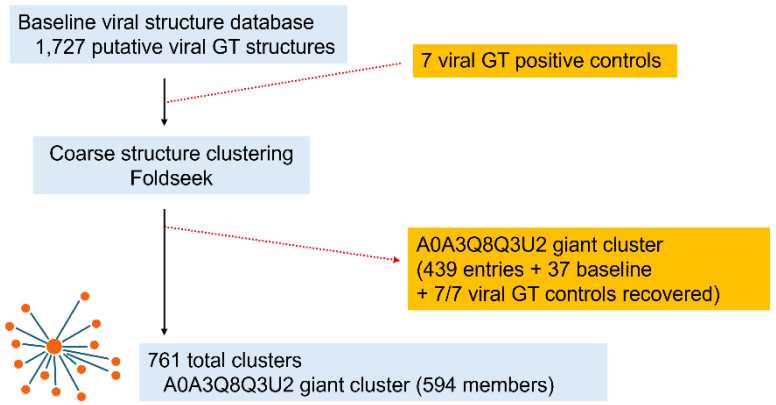
Supplementary Figures & Methods



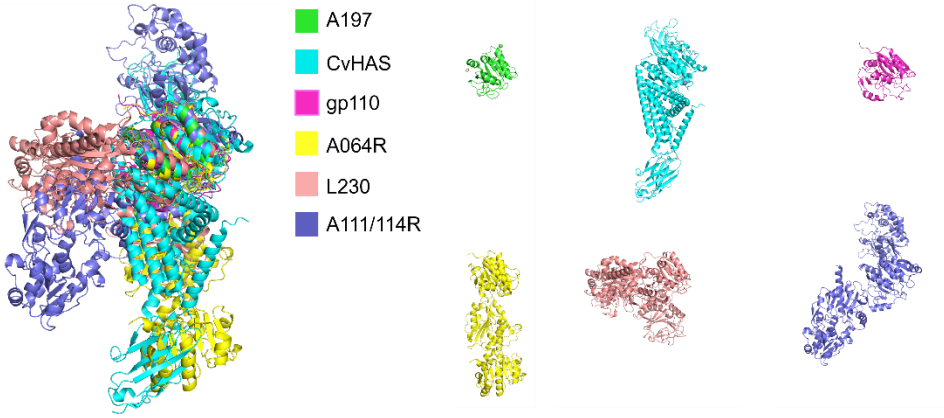
Supplementary Figure 1. Structural clustering of viral GT candidates.

The 1,081 viral proteins assigned to 115 multi-member clusters are shown as nodes, with each connected component representing one final Foldseek cluster. Blue nodes denote the 115 final cluster representatives, orange nodes denote 620 direct members with an explicit non-self representative–member alignment in the Foldseek m8 output, and purple nodes denote 346 indirect members assigned to the same final cluster but lacking an explicit alignment to the final representative in the current alignment output. Solid edges represent explicit representative–member alignments, whereas dashed edges were for indirect members and do not represent additional pairwise structural alignments. The dominant cluster contained 594 proteins. The 646 singleton proteins are not shown. Node positions and edge lengths reflect the Cytoscape layout.

A.



B.



C **Direct alignment extent and structural support**

versus A0A3Q8Q3U2 | versus best non-representative baseline giant member

Control	Direct comparison	Target	q-span/qlen	qcov	tcov	minTM
A197 2C0N · PDB	Representative	A0A3Q8Q3U2	170/191	0.89	0.88	0.38
	Best giant member	A0A346LU90	163/191	0.85	0.92	0.67
CvHAS 7SP6_A · PDB	Representative	A0A3Q8Q3U2	246/493	0.50	0.85	0.23
	Best giant member	UPI001FBC14AE	493/493	1.00	0.90	0.84
gp110 Che8_gp110 · AF3	Representative	A0A3Q8Q3U2	166/211	0.79	0.84	0.35
	Best giant member	A0A8S5Q9G5	208/211	0.99	0.97	0.80
A111/114R Q84432 · AF3	Representative	A0A3Q8Q3U2	211/860	0.24	0.85	0.16
	Best giant member	M1HIA0	853/860	0.99	1.00	0.96
A064R pbcv1_a064r · AF3	Representative	A0A3Q8Q3U2	166/637	0.26	0.82	0.17
	Best giant member	M1IF74	637/637	1.00	1.00	0.62
L230 (apo) Q5UQC3_apo · AF3	Representative	A0A3Q8Q3U2	210/895	0.23	0.81	0.12
	Best giant member	K7YEL3	895/895	1.00	1.00	0.79

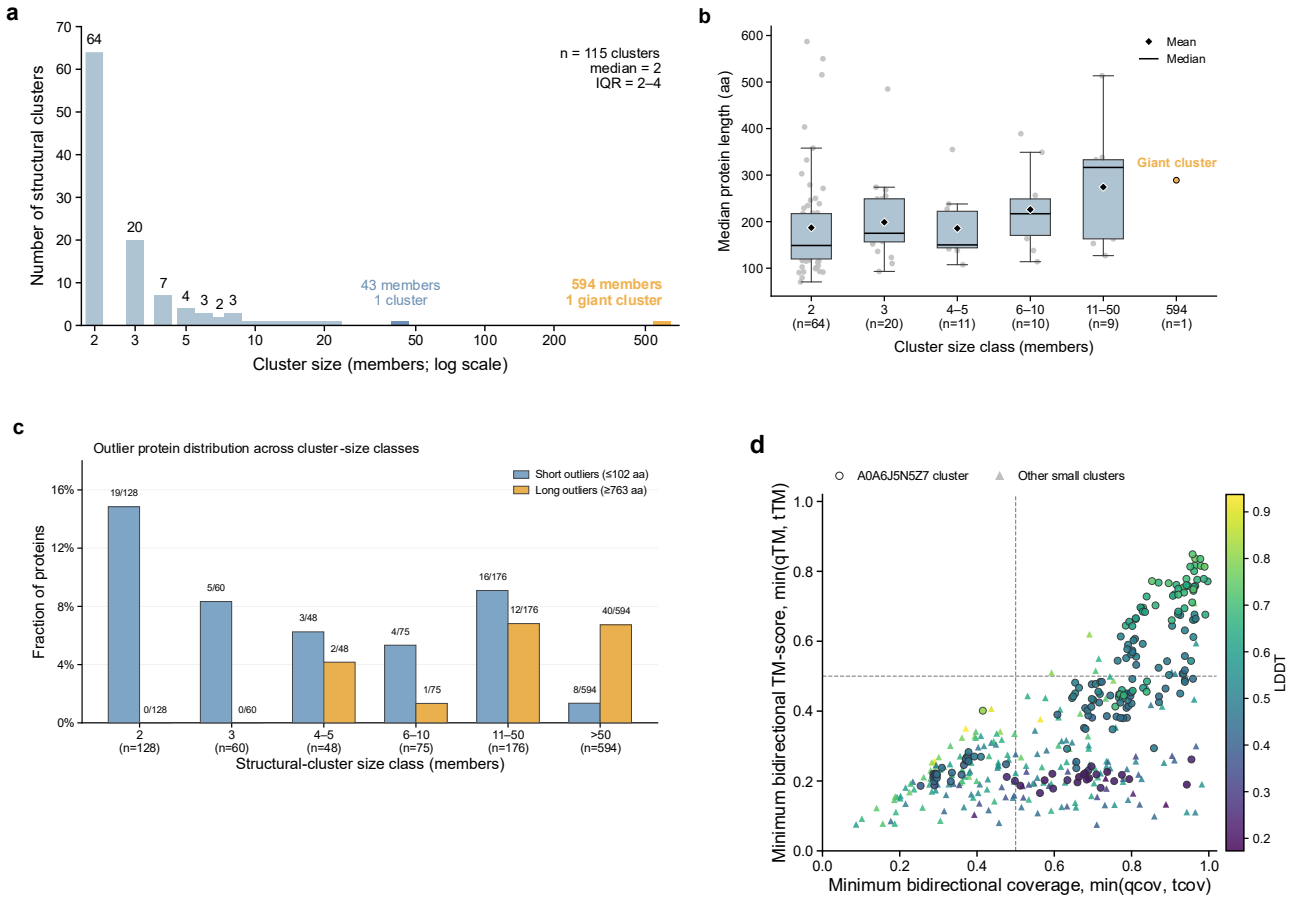
Colored bars show the aligned query-coordinate span; q-span = qend - qstart + 1. Best-member hits are ranked by min(qTM, tTM).

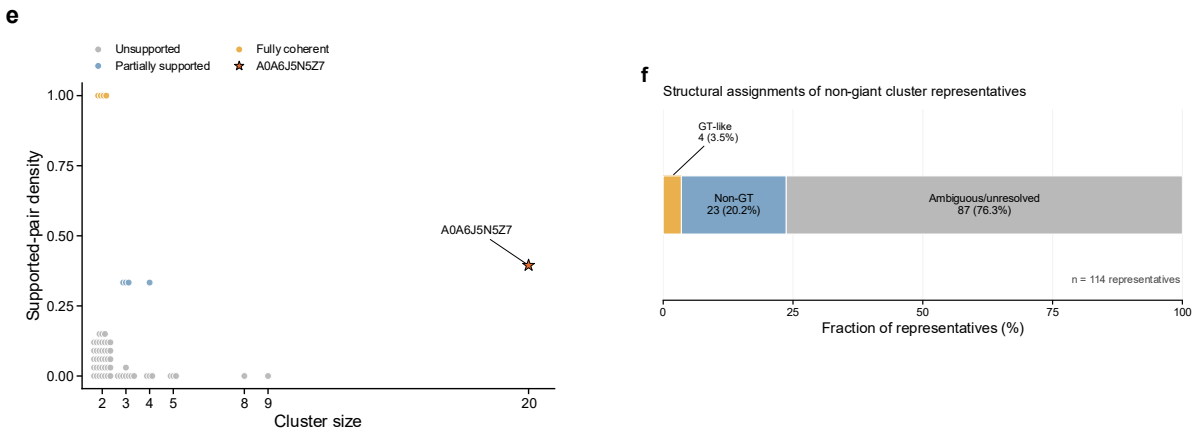
Supplementary Figure 2. Positive-control spike-in benchmark of permissive Foldseek coarse clustering. **A.** Foldseek clustering of 1,727 viral GT candidates produced 761 clusters, including a dominant cluster of 594 proteins represented by A0A3Q8Q3U2. After adding seven structural entries representing six viral GT controls,

clustering with unchanged parameters recovered all seven entries in the cluster represented by A0A3Q8Q3U2. The resulting cluster contained 483 entries, comprising 439 original dominant-cluster members, 37 proteins from other clusters and seven controls. Two nanobody chains co-deposited with CvHAS (PDB 7SP6) were excluded.

B. Structural superposition (left) and indidividual structures (right) of the six viral GT controls: A197 from *Sulfolobus turreted icosahedral virus 1* (green; PDB 2C0N); Chlorella virus hyaluronan synthase (CvHAS) from *Paramecium bursaria Chlorella virus CZ-2* (cyan; PDB 7SP6, chain A); the virion glycosylation-associated glycosyltransferase gp110 from *Mycobacterium phage Che8* (magenta; AlphaFold 3 model); the multidomain glycosyltransferase A111/114R from *Paramecium bursaria Chlorella virus 1* (yellow; UniProt Q84432; full-length AlphaFold 3 model); the multidomain rhamnosyltransferase/methyltransferase A064R from *Paramecium bursaria Chlorella virus 1* (salmon; full-length AlphaFold 3 model); and the bifunctional lysyl hydroxylase/glycosyltransferase L230 from *Acanthamoeba polyphaga mimivirus* (purple; UniProt Q5UQC3; full-length AlphaFold 3 models).

C. Direct structural alignments of each control with A0A3Q8Q3U2 (blue) and its best-matching non-representative member of the original dominant cluster (orange). Grey tracks represent full-length control proteins from N to C. Values indicate aligned query span, query coverage (qcov), target coverage (tcov) and the conservative structural similarity score min(qTM, tTM).

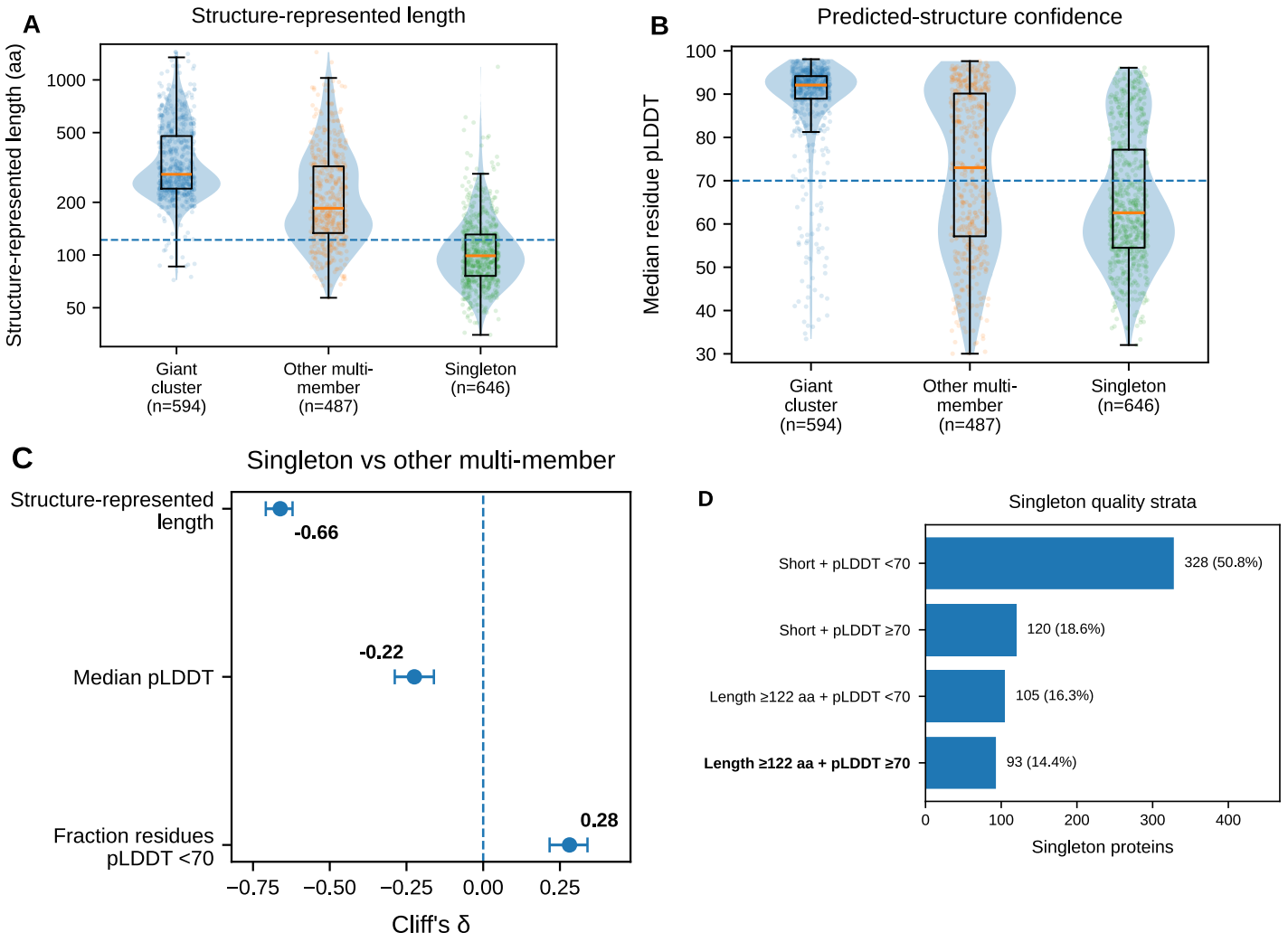




Supplementary Figure 3. Distribution and structural assessment of viral GT candidate clusters.

a. Exact size distribution of the 115 multi-member clusters containing 1,081 proteins. The 43-member and dominant 594-member clusters are highlighted. The median cluster size was 2 proteins (IQR, 2–4).
b. Distribution of cluster-level median protein lengths across cluster-size classes. Each point represents one cluster; boxes show the IQR, centre lines the median, whiskers 1.5 × IQR and diamonds the arithmetic mean. The 594-member cluster is shown as a single point.
c. Fractions of short (≤102 amino acids) and long (≥763 amino acids) proteins across cluster-size classes. Values indicate outlier counts relative to the total number of proteins in each class.
d. Pairwise structural comparisons of 383 protein pairs from 63 smaller clusters lacking direct representative–member alignments in the initial Foldseek output. Axes show the minimum query/target coverage, min(qcov, tcov), and minimum query/target-normalized TM-score, min(qTM, tTM); colour indicates LDDT. Comparisons from the 20-member A0A6J5N5Z7 cluster are shown as circles and all others as triangles. Dashed lines indicate thresholds of 0.50.

e. Supported-pair density for the same 63 clusters, defined as the fraction of all possible within-cluster pairs satisfying both $\min(qTM, tTM) \geq 0.50$ and $\min(qcov, tcov) \geq 0.50$. Coincident observations are shown as stacked dots; A0A6J5N5Z7 is highlighted.
f. Representative-level Foldseek assignments for all 114 non-giant clusters. Four representatives were classified as GT-like, 23 as non-GT and 87 as ambiguous or unresolved.



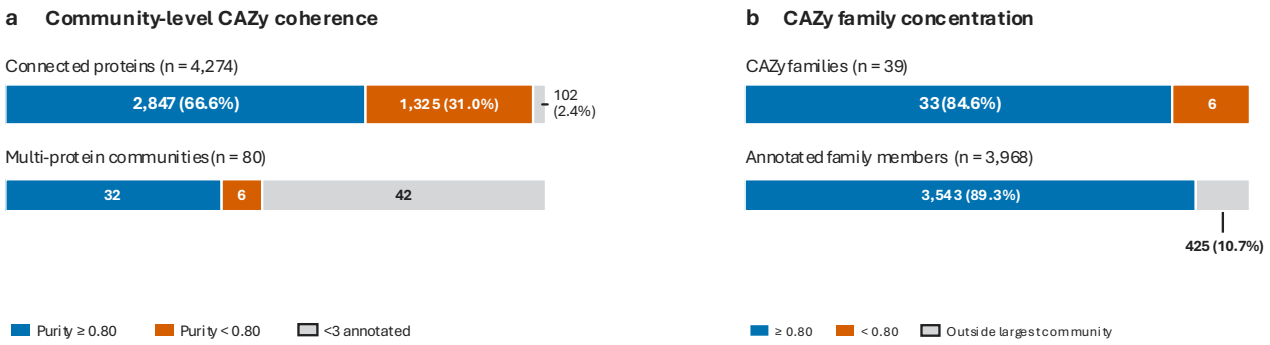
Supplementary Figure 4. Protein length and prediction confidence of viral GT candidate singletons.

A. Distribution of structure-represented protein lengths among the dominant cluster, other multi-member clusters and singletons. Numbers above the box plots indicate group medians. The dashed line marks the 10th percentile of the multi-member protein-length distribution (122 aa).

b. Distribution of median residue-level pLDDT for the same groups. The dashed line indicates pLDDT = 70. In a and b, boxes indicate the interquartile range (IQR), centre lines the median, whiskers 1.5 × IQR and individual points represent proteins.

c. Cliff's δ comparing singleton proteins with proteins from other non-giant multi-member clusters. Points show Cliff's δ and horizontal bars indicate 95% bootstrap confidence intervals. Negative values indicate lower values among singletons. Effect sizes were −0.661 for protein length (95% CI, −0.708 to −0.621), −0.224 for median pLDDT (95% CI, −0.288 to −0.161) and 0.281 for the fraction of residues with pLDDT below 70 (95% CI, 0.216 to 0.340). Statistical comparisons used two-sided Mann–Whitney U tests with Benjamini–Hochberg correction (adjusted P = 1.24×10^{-80}, 1.17×10^{-10} and 5.80×10^{-16}, respectively).

D. Grouping of 646 singleton proteins by structure-represented length (122 aa threshold) and median pLDDT (70 threshold). Of these, 328 (50.8%) were shorter than 122 aa with median pLDDT below 70, whereas 93 (14.4%) had lengths of at least 122 aa and median pLDDT of at least 70.

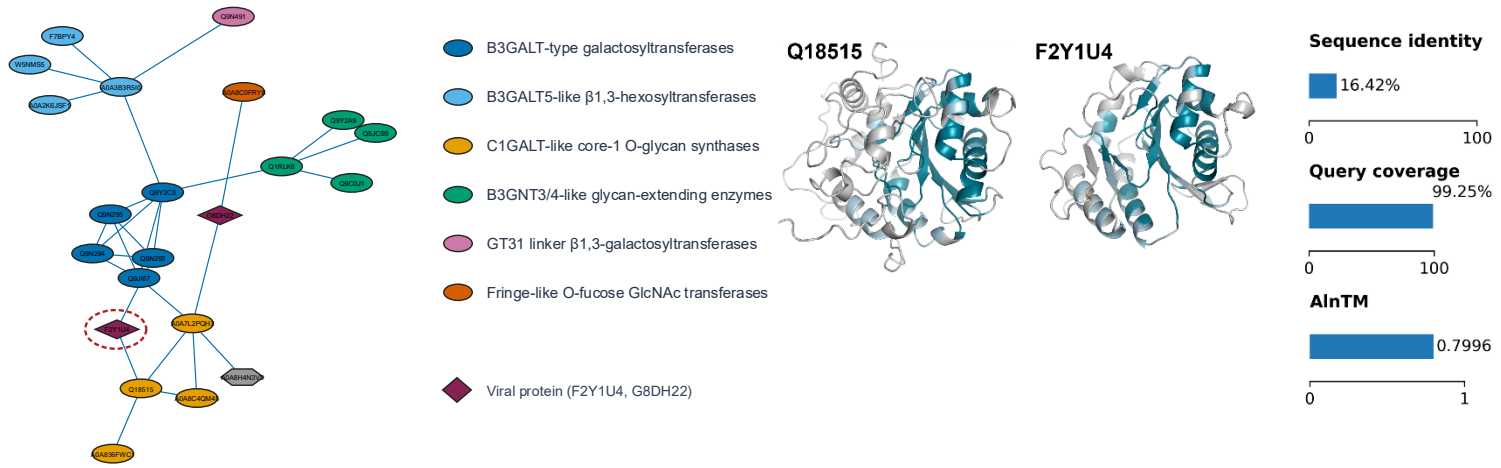


Supplementary Figure 5. Multi-protein structural communities substantially recapitulate CAZy family-level organization.

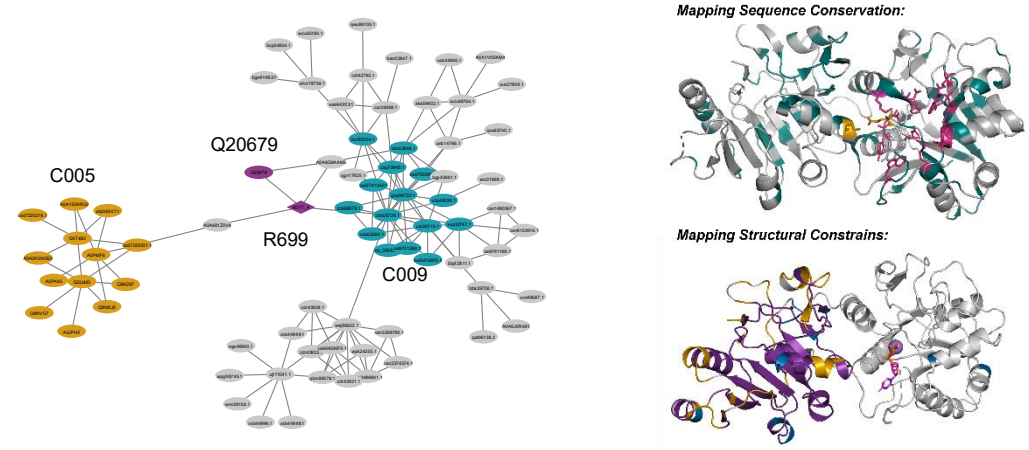
A. Community-level CAZy purity. The 80 multi-protein structural communities (4,274 proteins) were classified according to the proportion of CAZy-annotated members belonging to the most abundant family within each community. Communities containing at least three unambiguously annotated proteins were classified as family-coherent (purity ≥0.80; 32 communities) or mixed (purity <0.80; 6 communities). The remaining 42 communities were designated annotation-limited. The upper bar shows the number of communities in each category; the lower bar shows the corresponding protein counts, including proteins without CAZy assignments.

B. CAZy family concentration. For each of 39 CAZy families, concentration was defined as the fraction of its annotated members found in the structural community containing the largest number of members of that family. Thirty-three families (84.6%) had a concentration of at least 0.80 (upper bar). Across 3,968 unambiguously CAZy-annotated proteins, 3,543 (89.3%) occurred in their respective family's largest community, whereas 425 (10.7%) were distributed elsewhere (lower bar). The 580 singleton proteins were excluded from both analyses. CAZy annotations were applied after structural community detection and did not influence community assignment.

A.



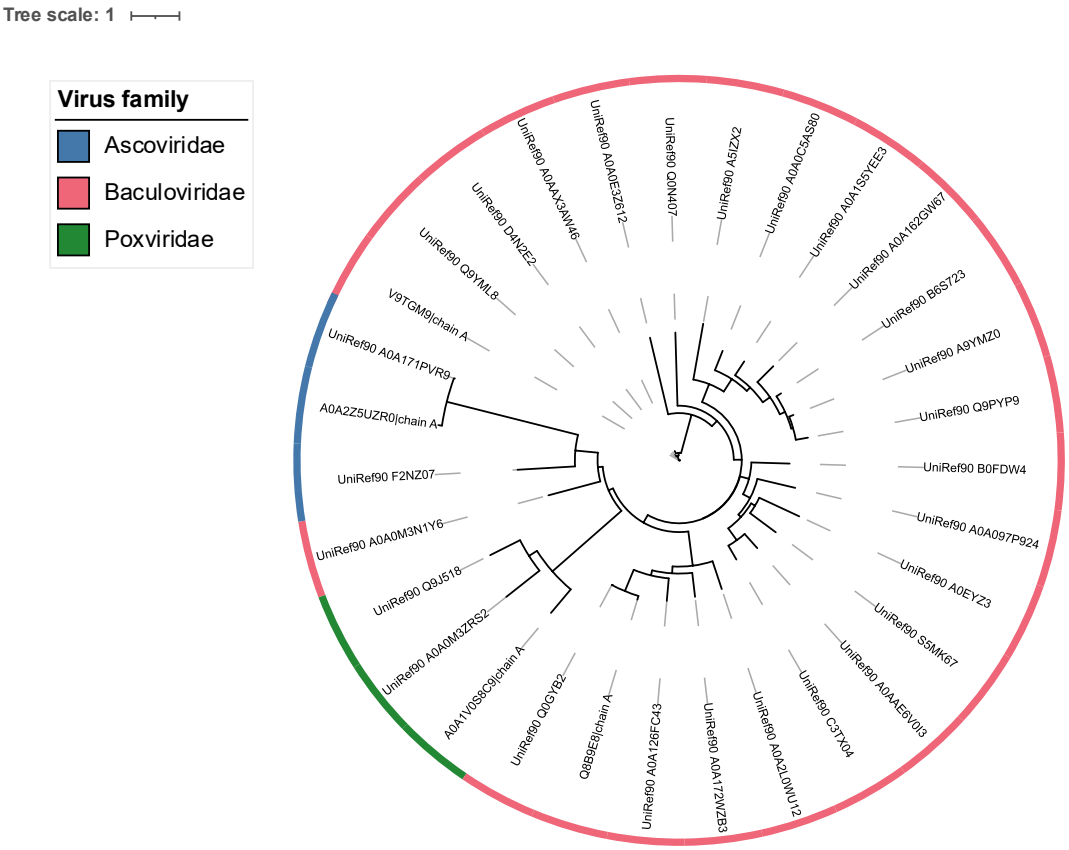
B.



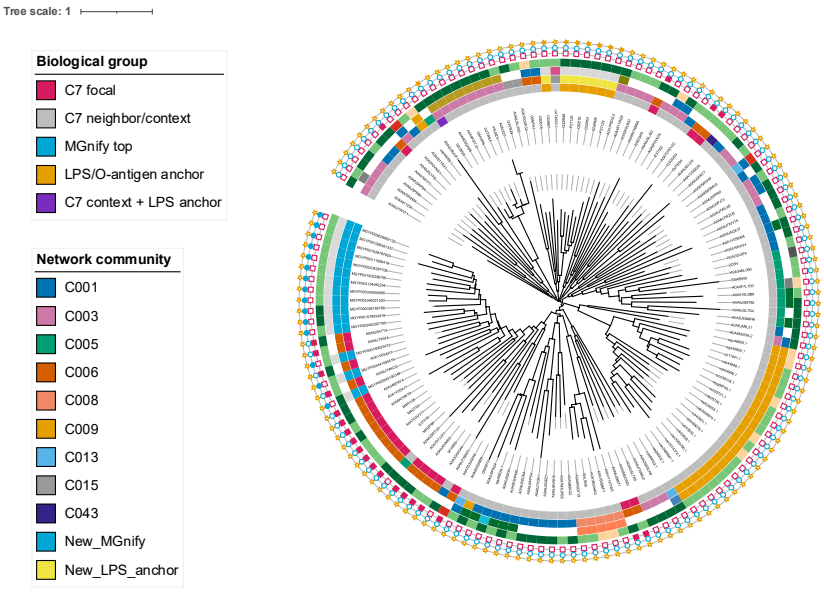
Supplementary Figure 6. Distinct structural relationships of viral glycosyltransferases within the GT-A landscape.

A. Structural placement of CAZy-unassigned viral GTs (F2YIU4 and G8DH22) within the GT31 family. Local structure-similarity network showing GT31 reference proteins coloured by functional subgroup, as indicated in the key. CAZy-unassigned viral proteins F2Y1U4 and G8DH22 are shown as purple diamonds, with F2Y1U4 highlighted by a dashed red circle. Structural comparison of the GT31 reference Q18515 and F2Y1U4 shows the strict pan-GT31 core (dark teal), additional structurally supported positions (light blue) and non-core regions (grey). Bars summarize structural comparisons of F2Y1U4 across four GT31 reference subclusters, showing mean subcluster-median sequence identity (16.42%), query coverage (99.25%) and alignment TM-score (AlnTM, 0.7996).

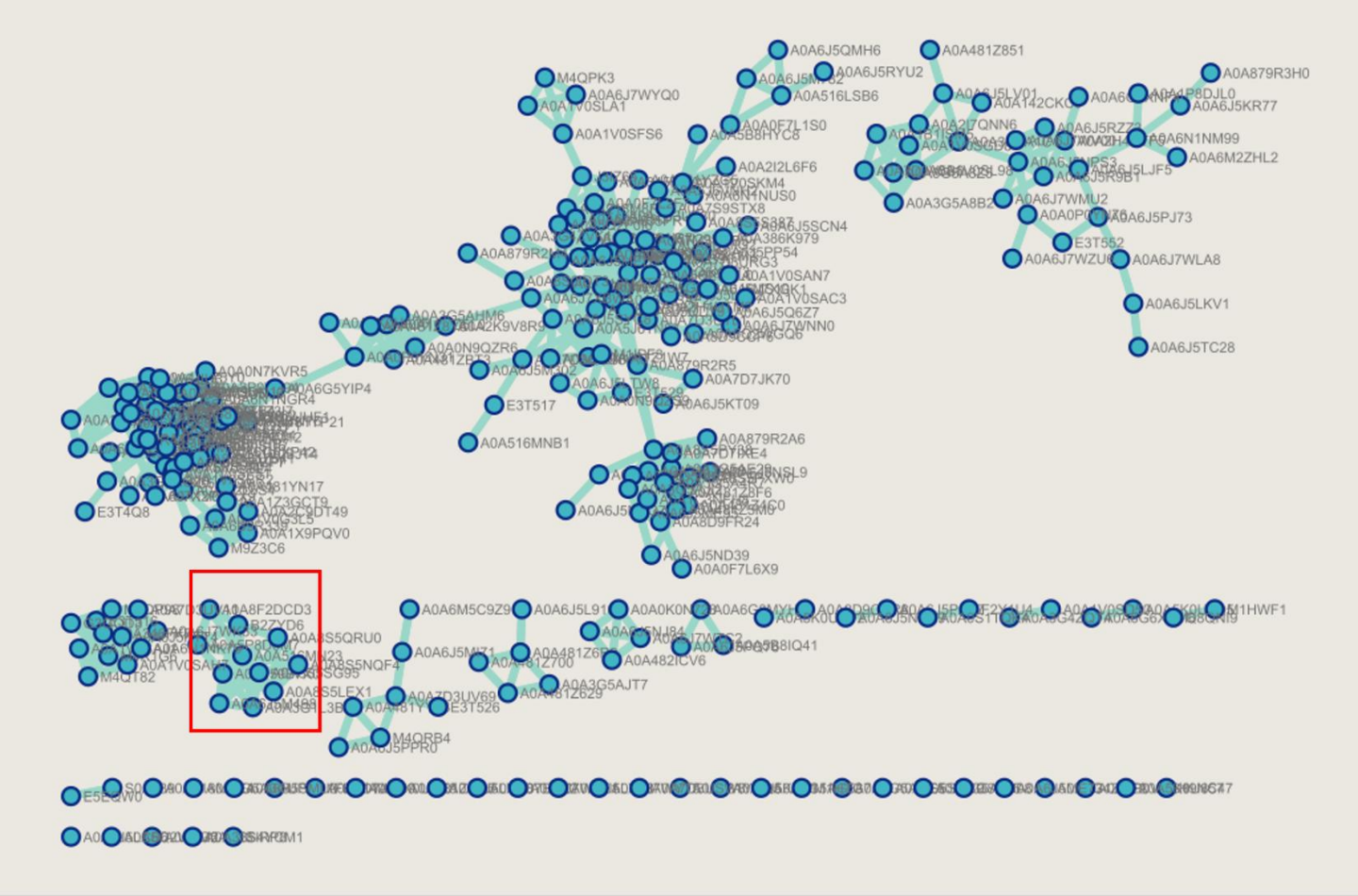
B. Domain-specific structural relationships of R699 across communities C005 and C009. Local structure-similarity network highlighting C005 (orange), C009 (turquoise), R699 (purple diamond) and the PLOD-like glycosyltransferase Q20679 (purple oval); remaining proteins are shown in grey. Upper, sequence identities between R699 and Q20679 mapped onto the R699 structure following FoldMason alignment. Identical residues are shown in teal, identical active-site residues in magenta, and non-identical active-site residues in orange. Other residues are shown in grey. Lower, conserved structural cores independently defined from C005 and C009 reference proteins mapped onto R699. Both communities showed extensive structural correspondence within the non-UDP-binding domain, whereas the UDP-binding catalytic domain showed limited structural support.



Supplementary Figure 7. Phylogenetic distribution of C027 homologues across viral families. Unrooted phylogenetic tree of 31 C027 homologues. The outer color strip denotes viral-family assignment: Ascoviridae (blue; *n* = 3), Baculoviridae (pink; *n* = 25), and Poxviridae (green; *n* = 3). Branch lengths indicate amino-acid substitutions per site;

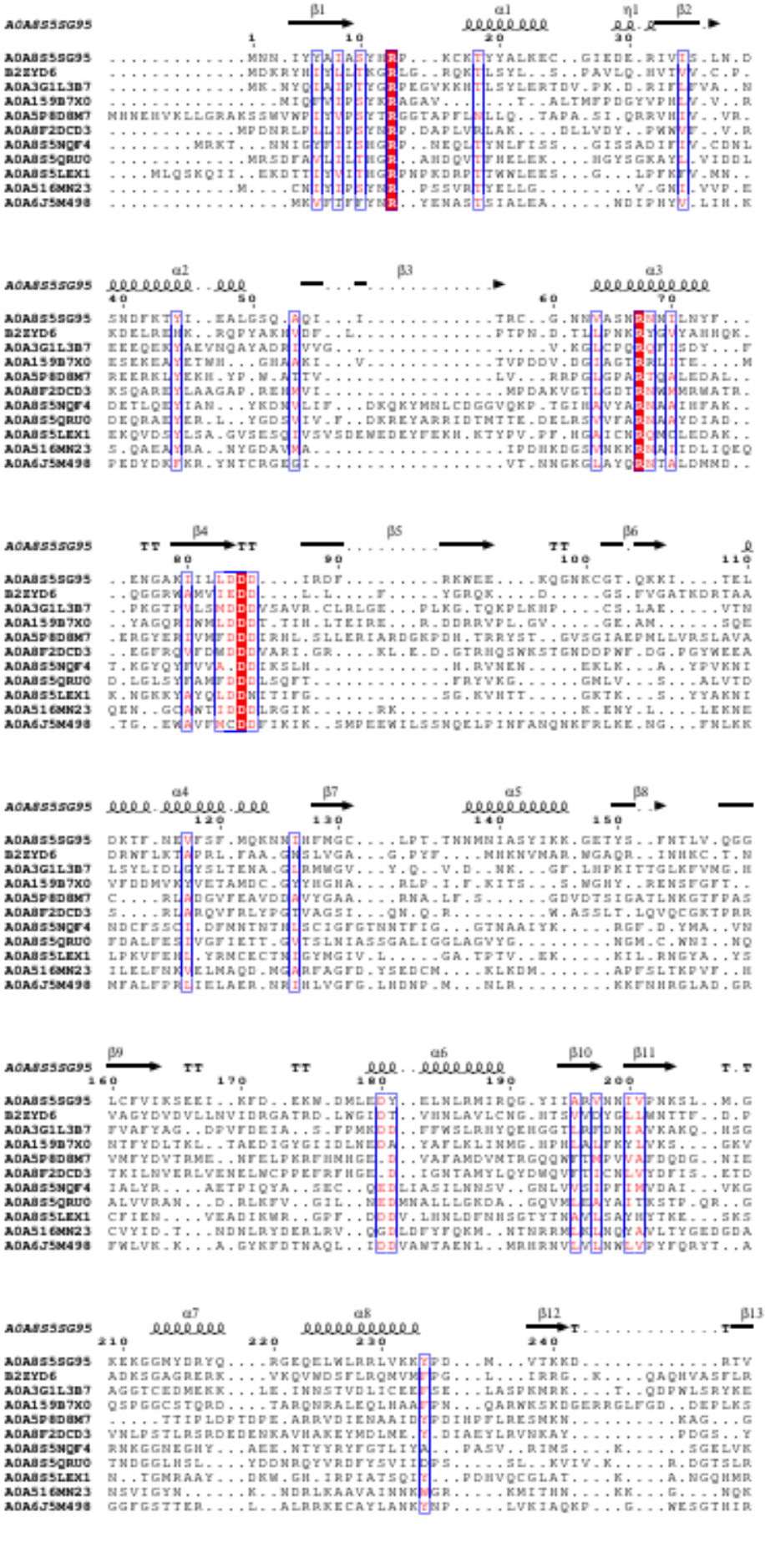


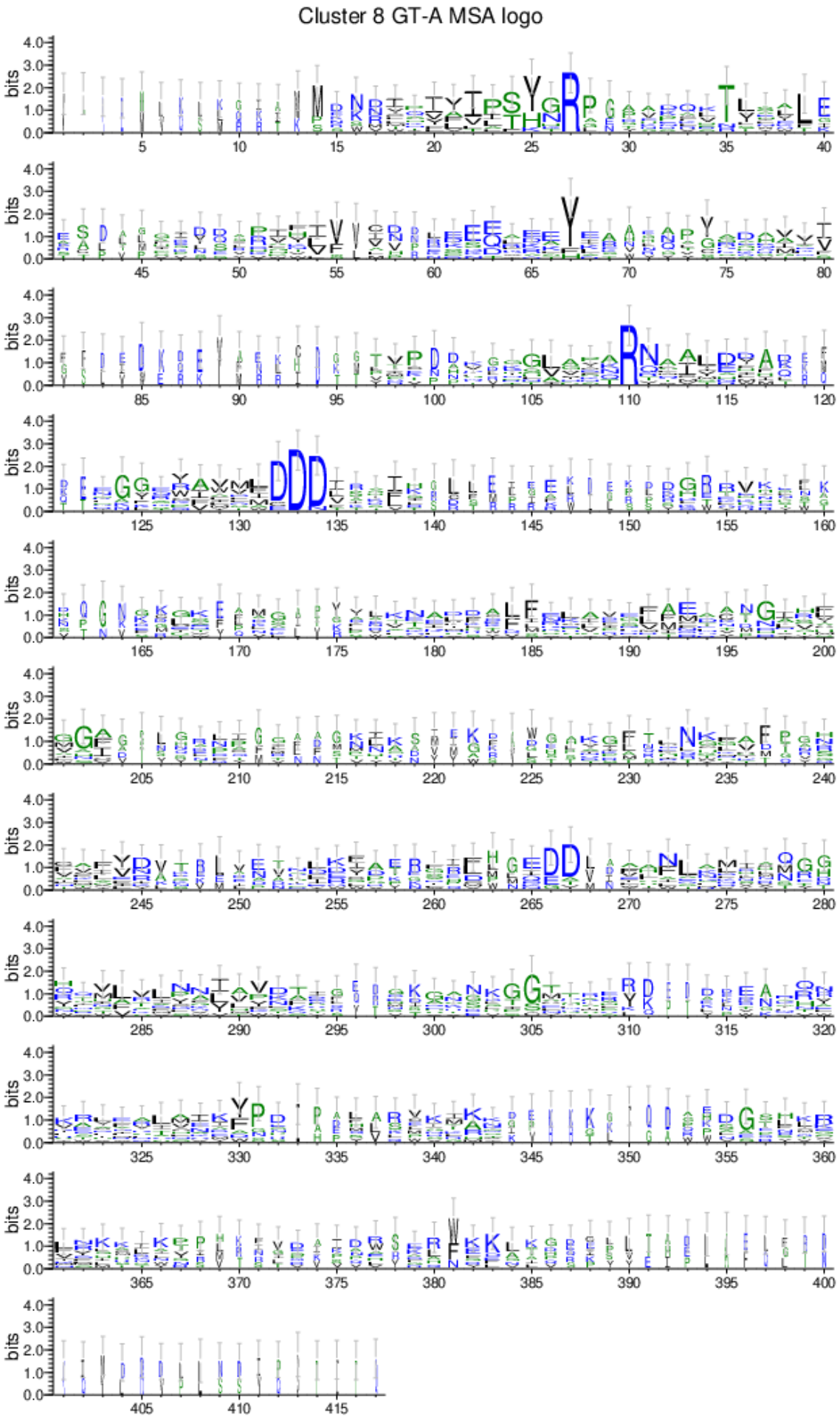
Supplementary Fig 8. Structural phylogeny of viral Cluster 7 and its extended neighbourhood. MAD-rooted FoldTree of 143 non-redundant structures spanning Cluster 7 focal proteins, network-neighbour proteins, MGnify structural hits and LPS/O-antigen-associated glycosyltransferase anchors. Branch lengths represent FoldTree-corrected structural distance. Concentric tracks indicate biological group, structure-similarity-network community, CAZy-family assignment and structure-confidence class; outer symbols mark Cluster 7 focal proteins, MGnify hits and LPS/O-antigen anchors.



Supplementary Figure 9. Iterative structural clustering identifies cluster 8 as a dark viral GT subnetwork.

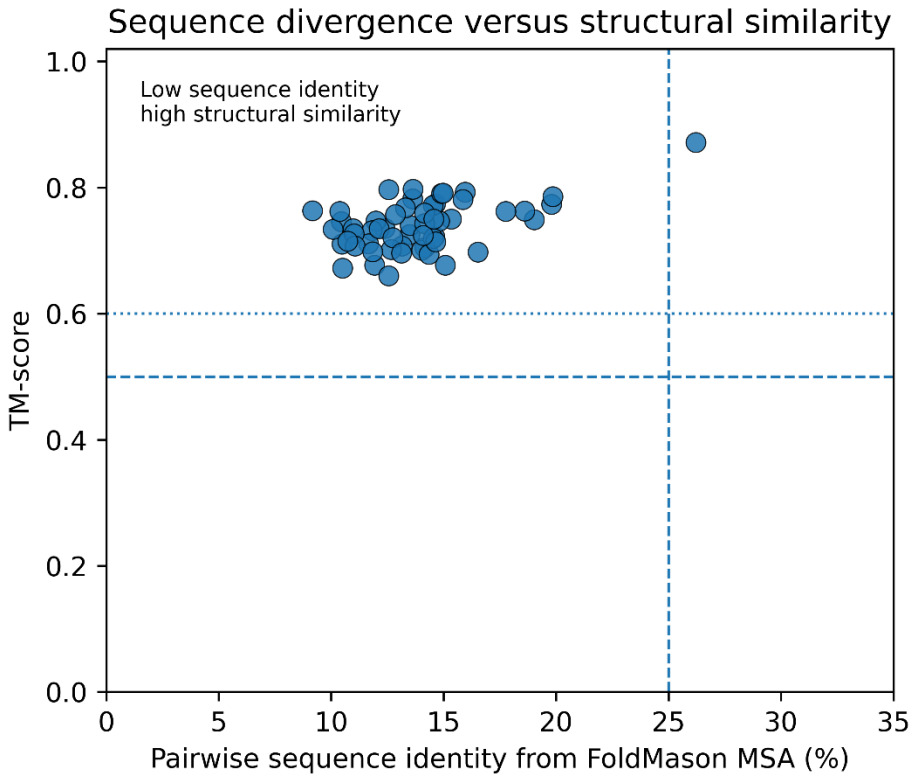
Structure similarity network of viral glycosyltransferase candidates generated from all-vs-all structural comparison. Nodes represent predicted viral GT-like proteins and edges represent pairwise structural similarity above the selected threshold. Iterative refinement of the initial clustering result revealed a small, distinct subnetwork, designated cluster 8 (red box). Cluster 8 was separated from the major viral GT communities and was prioritized for further analysis because its members lacked confident conventional sequence-based annotation but retained GT-like structural features and were subsequently found in DNA-modification-associated gene neighborhoods.





Supplementary Figure 10. Multiple sequence alignment of cluster 8

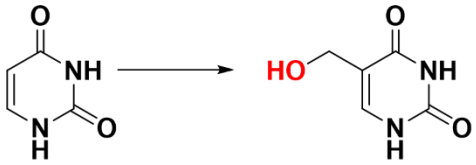
Multiple sequence alignment between cluster 8 members were generated by Foldmason. Visualization were performed by Esprit2 (ref). Sequence conservation were generated by Weblogo.



Supplementary Figure 11. Multiple sequence alignment of cluster 8

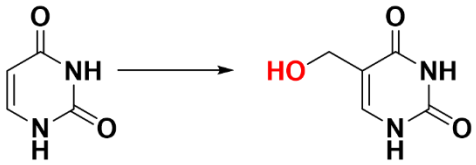
Each point represents one pairwise comparison between Cluster 8 proteins. The axis shows pairwise sequence identity derived from the FoldMason structure-guided amino-acid MSA, calculated over columns where both proteins contain residues. The y-axis shows the average TM-score from pairwise TM-align structural superposition.

Cytosine hydroxymethylase:



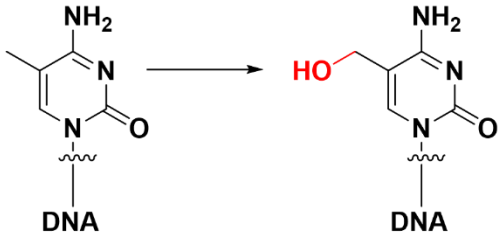
Type I & IV

Uracil hydroxymethylase:

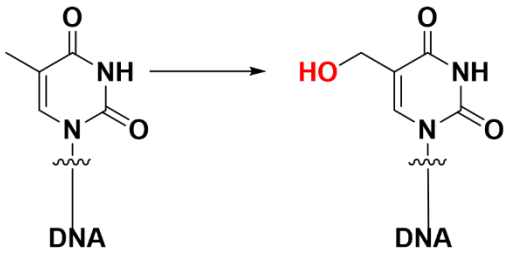


Type III

JBP/TET –like oxygenase on methylcytosine:

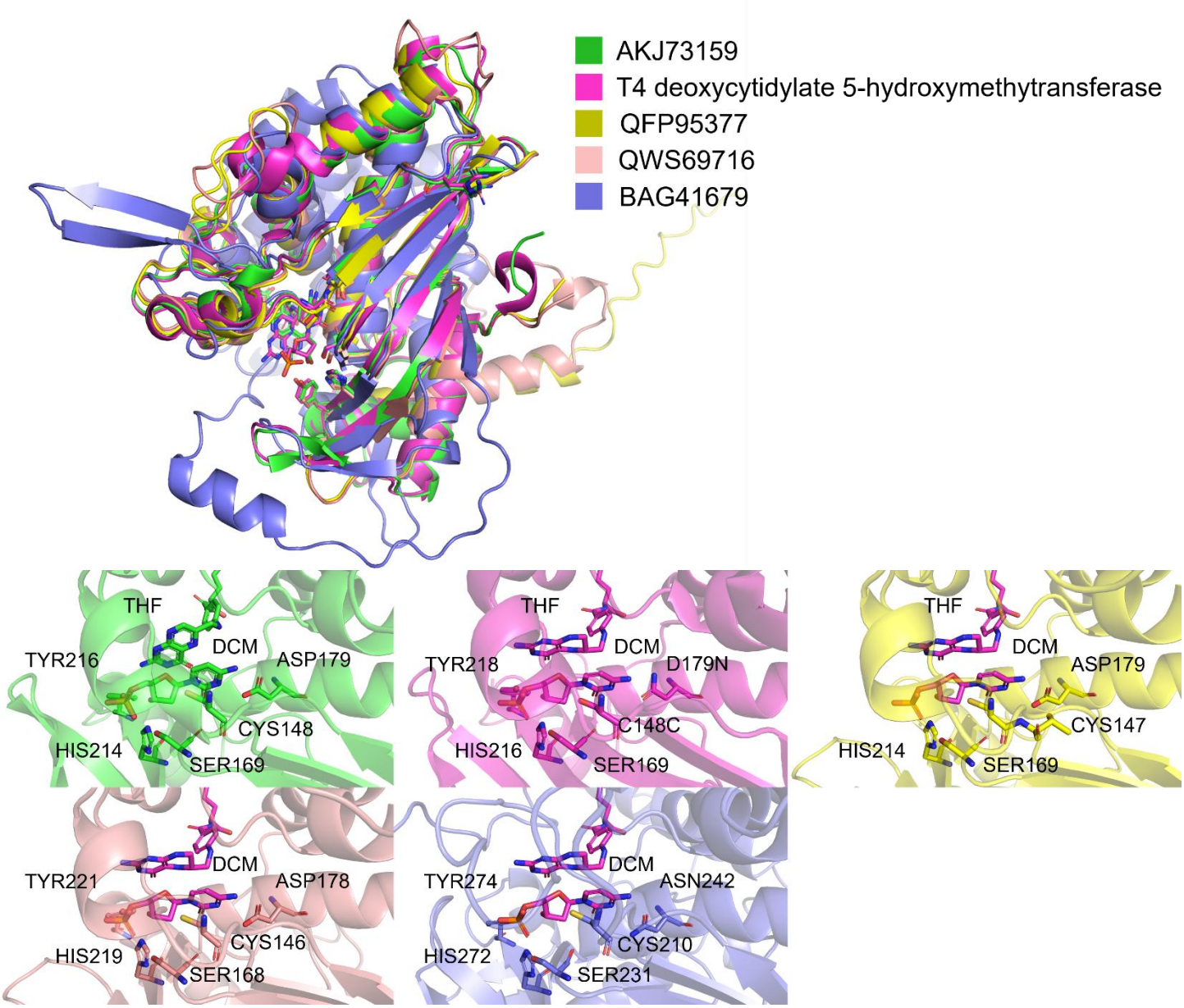


JBP/TET –like oxygenase on methylthymine:



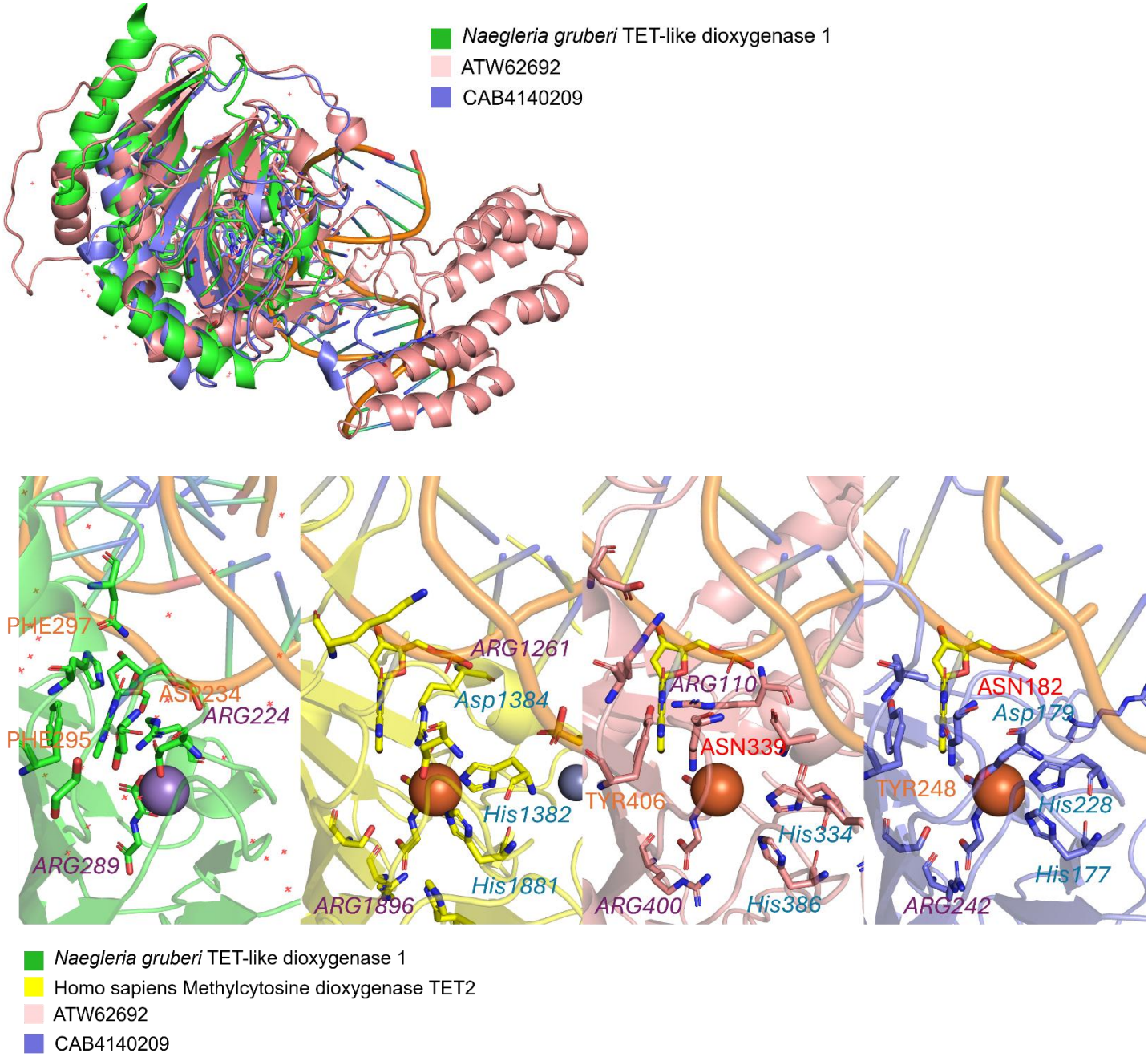
Type II

Supplementary Fig. 12. Hydroxymethyl-pyrimidine precursor chemistries associated with DNA-linked cluster 8 architectures. Schematic overview of the predicted precursor-generating reactions associated with the DNA-modification-linked cluster 8 locus types. Type I and Type IV architectures are predicted to use cytosine-derived nucleotide precursor routes, in which dCMP hydroxymethylase-like enzymes generate 5-hydroxymethyl-dCMP/5hmC-type precursors for downstream DNA glycosylation. Type II architectures are predicted to use a post-replicative TET/JBP-like 2-oxoglutarate/Fe(II)-dependent oxygenase route, in which thymine/5-methyluracil already embedded in DNA is oxidized to 5-hydroxymethyluracil. Type III architectures are predicted to use a uracil-derived nucleotide precursor route, in which a dUMP hydroxymethylase-like enzyme generates a 5-hydroxymethyl-dUMP/5hmdU-type precursor. Red highlights indicate the hydroxymethyl handle introduced or generated by each precursor-forming enzyme. These hydroxymethyl-pyrimidine handles provide the proposed nucleophilic acceptors for downstream cluster 3-associated DNA glycosylation.



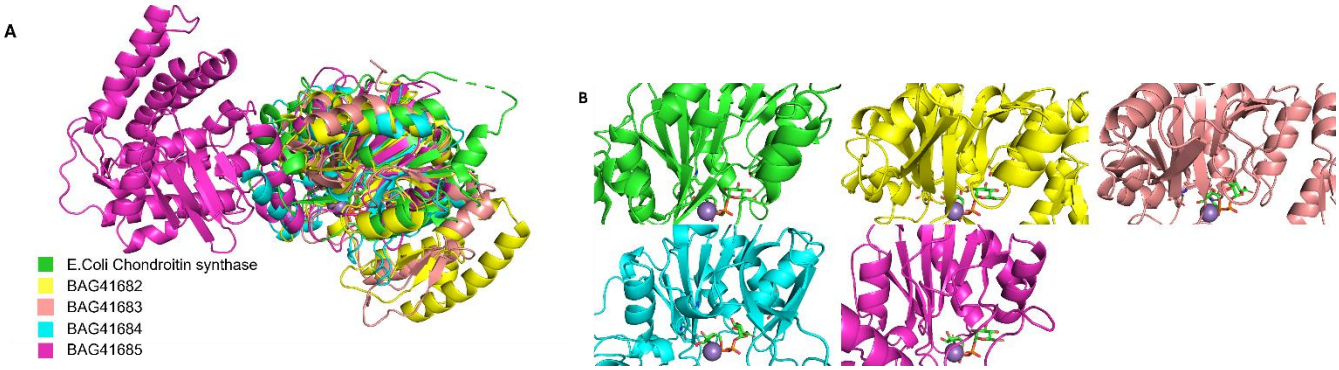
Supplementary Fig. 13. Structure-guided annotation of hydroxymethyl-pyrimidine precursor enzymes in DNA-associated cluster 8 loci. Structural and active-site comparison of predicted folate-dependent pyrimidine C5-hydroxymethyltransferases encoded near DNA-associated cluster 8 loci. The top panel shows structural superposition of representative cluster 3-neighborhood enzymes with the crystallographic reference T4 deoxycytidylate 5-hydroxymethyltransferase. The reference structure contains bound tetrahydrofolate (THF) and deoxycytidylate/dCMP-site ligand, providing an active-site template for substrate-recognition analysis. Predicted enzymes from Type I and Type IV loci, including AKJ73159, QFP95377, and QWS69716, preserve the cytosine-directed active-site architecture of T4-like dCMP hydroxymethyltransferases, including the catalytic cysteine, conserved histidine/serine residues, aromatic base-positioning residue, and an Asp residue positioned for cytosine recognition. These features support annotation of these proteins as dCMP C5-hydroxymethyltransferase-like enzymes that can generate 5hmC-type precursors for downstream DNA glycosylation. In contrast, the Type III enzyme BAG41679 preserves the same global fold and catalytic cysteine but contains Asn242 at the position equivalent to the cytosine-specific Asp residue in dCMP-directed enzymes. This Asp-to-Asn active-site substitution is consistent with altered pyrimidine specificity and supports annotation of BAG41679 as a dUMP hydroxymethyltransferase-like enzyme predicted to generate a 5-

hydroxymethyluracil-type precursor. Together, these comparisons justify the assignment of Type I/IV loci as 5hmC-linked pathways and the Type III locus as a distinct 5hmdU-linked hydroxymethyl-pyrimidine pathway.

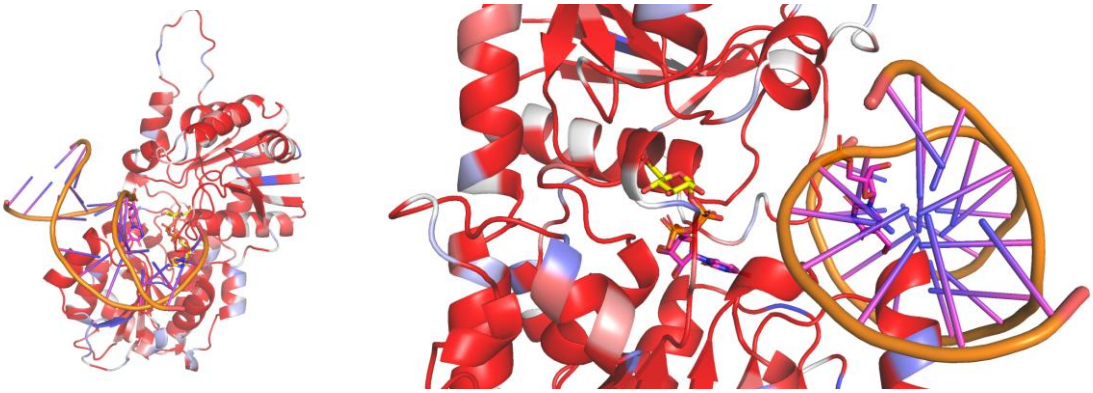


Supplementary Fig. 14. Structure-guided annotation of Type II TET/JBP-like oxygenases supports a thymine/5-methyluracil hydroxylation module. Predicted Type II oxygenases encoded in DNA-associated cluster 8 loci were compared with structurally characterized TET/JBP-like 2-oxoglutarate/Fe(II)-dependent dioxygenases. The left panel shows structural superposition of Type II oxygenases with reference Naegleria gruberi TET-like dioxygenase 1 and human TET-family oxygenase structures. The right panels show active-site comparisons centered on the catalytic metal, 2-oxoglutarate-binding region, and flipped-base

recognition pocket. Residues involved in Fe(II) coordination are highlighted in blue, whereas residues involved in 2-oxoglutarate binding are highlighted in magenta. These catalytic features are spatially conserved across the Type II oxygenases and reference enzymes, supporting annotation of the Type II proteins as TET/JBP-like 2OG/Fe(II) dioxygenases. The flipped-base recognition pocket is also conserved, including spatially aligned aromatic residues positioned to stack against or stabilize the extruded pyrimidine base. Notably, both Type II oxygenases contain an Asp-to-Asn substitution at the base-recognition position corresponding to the cytosine/5mC-specific Asp in TET-like enzymes. This substitution is consistent with a shift toward thymine/5-methyluracil recognition and supports a model in which the Type II oxygenases generate 5-hydroxymethyluracil-like DNA bases for downstream glycosylation by neighboring cluster 8 GTs.

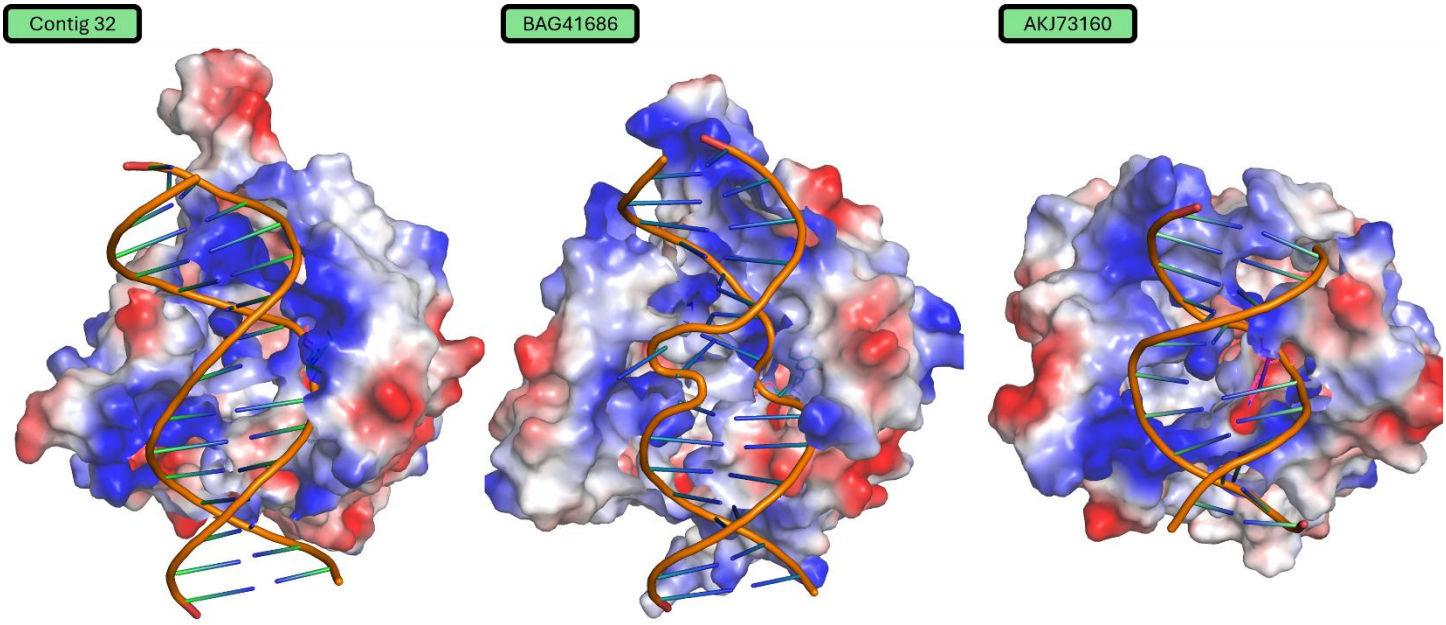


Supplementary Fig. 15. Structural annotation of additional GT-like enzymes in the B2ZYD6 Type III locus.
Additional proteins including BAG41682, 41683, 41684, 41685 encoded in the B2ZYD6 Type III locus were compared structurally with *E. coli* chondroitin synthase, a GT2-fold glycosyltransferase reference. Global structural superposition and active-site inspection show that several B2ZYD6 proteins adopt GT-like folds with conserved nucleotide-sugar donor-binding architecture, supporting their annotation as putative glycosyltransferases which contribute to downstream elaboration of a DNA-linked glycan.

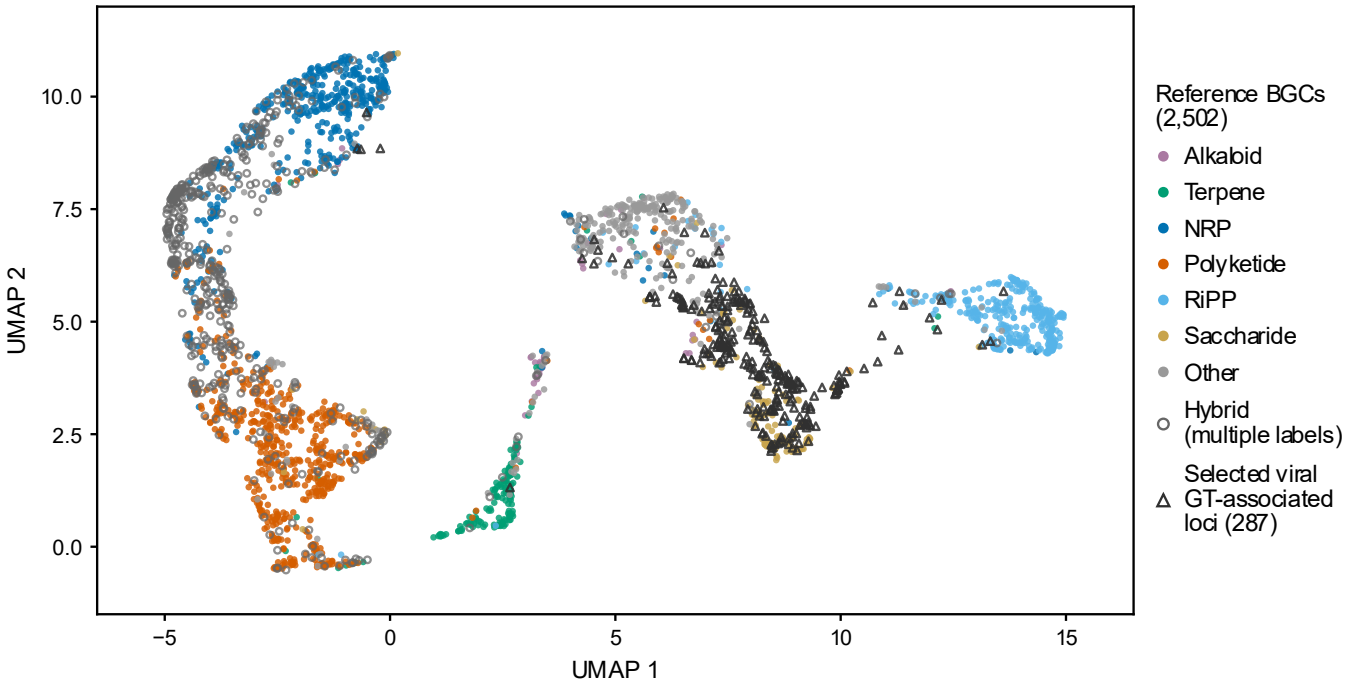


Supplementary Fig. 16 .AKJ73146 preserves conserved features of T4-like α-glucosyltransferases. Sequence alignment result between Enterobacteria phage T4 α-glucosyltransferases and AKJ73146 was mapped onto the predicted AKJ73146 structure. Conserved residues are shown in red and variable residues in blue. The global view (left) shows conservation across the catalytic cleft and DNA-binding surface, while the

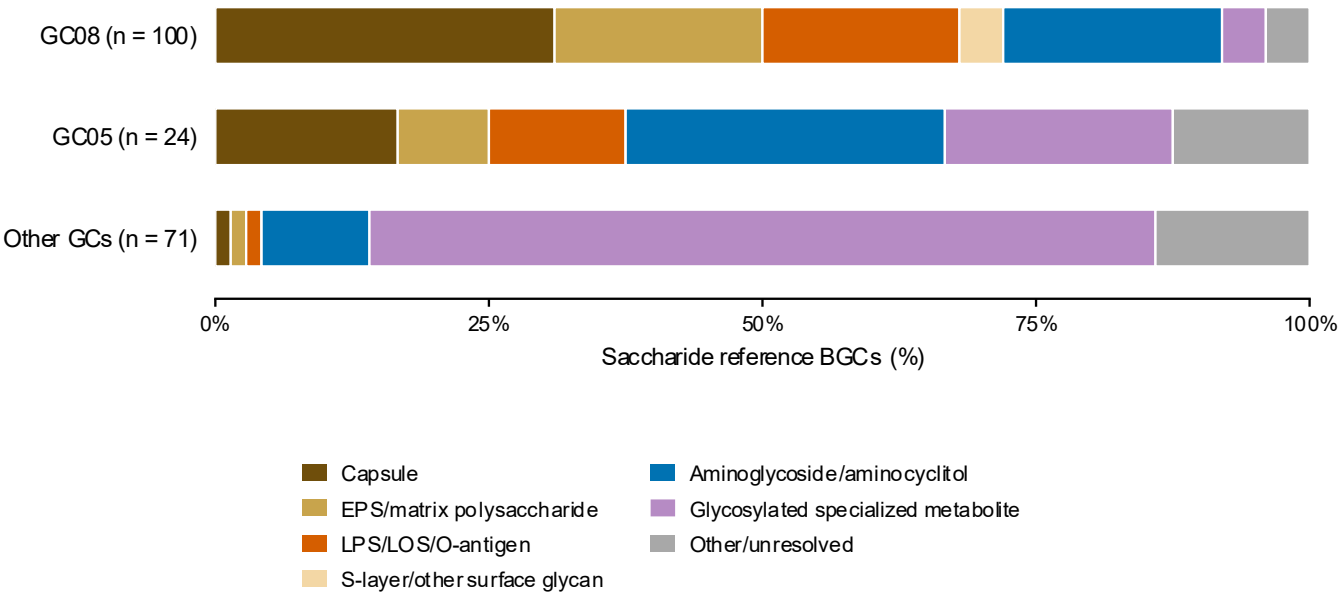
active-site zoom (right) highlights strong conservation around the predicted catalytic pocket. This pattern supports AKJ73146 as a T4-like α-glucosyltransferase that likely performs the first glucosylation step on 5hmC-DNA.



Supplementary Fig. 17. BAG41686 and AKJ73160 retain electropositive DNA-binding grooves. Electrostatic surface maps of contig 32, BAG41686, and AKJ73160 modeled with modified DNA acceptors. Blue indicates electropositive surface potential and red indicates electronegative potential. The reference DNA GT contig 32 contains a positively charged groove that accommodates the DNA phosphate backbone. BAG41686 preserves a comparable groove, supporting direct binding of hydroxymethylated DNA. AKJ73160 also retains an electropositive DNA-binding surface, consistent with DNA-associated substrate recognition, although its active site is remodeled to recognize Glc-5hmC rather than naked 5hmC. These results support distinct acceptor-recognition modes within a shared DNA-binding framework.



Supplementary Figure 18. Uniform manifold approximation and projection (UMAP) of 320-dimensional BGC-Prophet embeddings comprising 2,502 MIBiG reference biosynthetic gene clusters (BGCs) and 556 viral GT-associated loci. Reference BGCs are shown as circles coloured according to their annotated biosynthetic classes: alkaloid (n = 53, 2.1%), terpene (n = 150, 6.0%), non-ribosomal peptide (NRP; n = 468, 18.7%), polyketide (n = 606, 24.2%), ribosomally synthesized and post-translationally modified peptide (RiPP; n = 330, 13.2%), saccharide (n = 130, 5.2%) and other (n = 298, 11.9%). Open circles represent hybrid BGCs assigned to multiple biosynthetic classes (n = 467, 18.7%). Open triangles indicate 287 selected viral GT-associated loci, identified through Saccharide-class prediction and/or nearest-reference retrieval (P@10 ≥ 0.5). The remaining 269 viral loci were included in the global embedding but are not displayed. Percentages are calculated relative to the 2,502 reference BGCs.



Supplementary Figure 19. Chemotype composition of Saccharide-associated reference BGCs across global Leiden communities. Stacked bars show the relative proportions of annotated saccharide-related chemotypes among MIBiG reference biosynthetic gene clusters (BGCs) assigned to GC08 (n = 100), GC05 (n = 24) and the remaining global communities combined (Other GCs; n = 71). Global communities were identified by Leiden clustering of the joint embedding space comprising reference BGCs and viral GT-associated loci. Reference BGCs were categorized as capsule, extracellular polysaccharide (EPS)/matrix polysaccharide, lipopolysaccharide/lipooligosaccharide/O-antigen (LPS/LOS), S-layer/other surface glycan, aminoglycoside/aminocyclitol, glycosylated specialized metabolite, or other/unresolved. Percentages were calculated within each group, and n denotes the number of Saccharide-associated reference BGCs, including those with hybrid class annotations.

Supplementary Methods

Definition and mapping of the pan-GT31 structural core

A common FoldMason alignment (Gilchrist et al., 2026) was generated for 17 GT31 reference structures assigned to four multi-member structural subclusters: SC01 ($n = 5$), SC02 ($n = 4$), SC03 ($n = 4$) and SC04 ($n = 4$). Candidate pan-GT31 core positions were evaluated at homologous alignment columns using subcluster occupancy, model confidence and medoid-anchored local structural conservation. A strict pan-GT31 core position was required to satisfy all prespecified criteria in each of the four subclusters: occupancy ≥0.70; medoid-residue pLDDT ≥70; within-subcluster local lDDT ≥0.75 with neighbourhood coverage ≥0.60; and cross-subcluster local lDDT ≥0.65 with neighbourhood coverage ≥0.40 and focal-position presence ≥0.50. These criteria identified 98 strict core columns. Supported core positions satisfied the same occupancy and confidence requirements but were permitted to retain structural support in three of the four subclusters while satisfying the median geometric-support criteria. The supported set contained 147 columns, comprising the 98 strict positions and 49 additional supported-only positions.

Reference-derived core columns were mapped to Q18515 through its coordinates in the reference-only FoldMason alignment. F2Y1U4 was excluded from core discovery and mapped subsequently through the query-inclusive alignment, using Q9N294 residue indices as the coordinate bridge. Q18515 contained all 98 strict and 49 supported-only positions; F2Y1U4 mapped to 93 strict and 41 supported-only positions. Group-level placement metrics were calculated separately for SC01–SC04. The displayed values represent the arithmetic mean of the four subcluster-level medians.

Structure-guided mapping of sequence identities between R699 and Q20679

For structure-guided comparison of R699 with a metazoan PLOD-like collagen glycosyltransferase, the glycosyltransferase-relevant region of Q20679 was isolated by removing the lysyl-hydroxylase domain prior to alignment. R699 and the trimmed Q20679 construct were then aligned using FoldMason (Gilchrist et al., 2026), and the resulting pairwise amino-acid alignment was used as the coordinate framework for residue classification.

Alignment columns were parsed sequentially and converted to ungapped residue indices for each protein. These indices were then mapped onto the corresponding structure coordinates of R699 and trimmed Q20679 to recover the residue numbering used for visualization. Exact sequence identity was assigned when the aligned residues in the two proteins were identical. R699 active-site residues were defined from the experimentally characterized donor-binding, metal-coordinating and acceptor-binding network of R699 and included residues 48, 54, 62, 83, 85, 86, 114, 118, 138, 162, 163, 164, 216, 218, 222 and 438.

Each aligned residue position was assigned to one of four visualization classes: (i) active-site identity, when an exact sequence identity overlapped an R699 active-site residue; (ii) sequence identity, for exact identities outside the R699 active-site set; (iii) active-site non-identity, for R699 active-site residues aligned to non-identical residues in Q20679; and (iv) aligned non-identity, for all other aligned non-identical positions. Unaligned positions and background residues were rendered separately as non-highlighted regions. Residue classes were displayed on the R699 and Q20679 structures in PyMOL, using magenta for active-site identity, teal for non-active-site sequence identity, orange for active-site non-identity, and grey tones for other aligned or background regions. For structural context, UDP from the R699 structure was shown as grey sticks.