AI-facilitated high-resolution cryo-EM analyses of tubular mastigonemes reveal the structural roles of N- and O-glycans

Junhao Huang[1,4,5], Hui Tao[1,4], Sheng Chen[1,4], Yahua Cui[1], Yiran Xu[1], Chuangye Yan[1,5], and Nieng Yan[1,2,3,5]

[1]Beijing Frontier Research Center for Biological Structures, State Key Laboratory of Membrane Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

[2]Institute of Bio-Architecture and Bio-Interactions (IBABI), Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen 518107, Guangdong Province, China

[3]Institute of Chemical Biology (ICB), Shenzhen Bay Laboratory, Guangming District, Shenzhen 518132, Guangdong Province, China

[4]These authors contributed equally.

[5]To whom correspondence should be addressed: Junhao Huang (huangjh22@tsinghua.org.cn); Chuangye Yan (yancy2019@tsinghua.edu.cn); Nieng Yan (nyan@tsinghua.edu.cn).

Summary

Glycans are essential mediators of multiple biological processes, yet their chemical complexity and non-template-driven biosynthesis pathways pose significant challenges for the establishment of the sequence-structure relationship. Here, we report cryo-EM structures of the tubular mastigoneme from a golden alga species. A large number of N- and O-glycans are observed at 1.8-2.2 Å resolutions. We developed EModelG, an AI-guided pipeline for automated atomic modeling of glycans. In addition to a number of canonical high-mannose or complex type N-glycans, we have discovered and confirmed a non-canonical N-glycan linked to Asn in the Ala-Asn-Asp (AND) motif. The surface spikes on the mastigoneme are made of dense O-glycans coating the stem of linear PSXX tetrapeptide repeats. The trihydroxylated proline anchor and sugar compositions, linkages, as well as acetylation, methylation, and sulfation of sugar residues, are unambiguously resolved from the > 2.0-Å EM map and further confirmed with LC-MS/MS analyses. Well-resolved water molecules, which account for more than 10% of the resolved volume, mediate extensive hydrogen bonding. Our integrative approach, combining cryo-EM, glycoproteomics, and AI modeling, establishes a scalable framework for deciphering glycan folding.

Introduction

Glycans, a collective name for all forms of carbohydrates, represent the most abundant biomolecules on Earth by mass. In addition to providing energy, glycans participate in broad biological functions, such as signal transduction, immune response, biosynthesis, and cell adhesion [1-10]. In contrast to their biological importance, high-resolution structural information on glycans in their native conformations remains scarce due to their inherent chemical complexity, conformational flexibility, and non-templated biosynthesis [11-15].

Current analytical techniques for structural analysis of glycans, such as nuclear magnetic resonance (NMR), mass spectrometry (MS), and atomic force microscopy (AFM), each have their unique strength but also encounter limitations to varying degrees [15-18]. None of these technologies can resolve high-resolution structures of native glycans with complex compositions and conformations. In light of these challenges, we have adapted a structure-first paradigm by introducing the CryoSeek strategy, which exploits cryo-EM as the observation tool to discover uncharacterized bio-entities from natural or native specimens at sub-atomic resolutions and beyond [19]. Quick progress has been made on structure determination of various glycans ever since [19-22]. Dozens of glycofibril structures have been determined, revealing diverse protein stems covered with dense glycan shells.

The first of such glycofibrils that we discovered comprises a single strand of linear tetrapeptide repeats. In each repeat, a dihydroxyproline (diHyp) and the ensuing Ser or Thr are modified with two and one branched glycan chains, respectively [20,21]. For simplicity, we

refer them as the PSXX type of glycofibrils. The most unexpected discovery concerns the protein-free fibrils that are made exclusively of sugar residues. With different sugar compositions, they fold to distinct structures, unambiguously demonstrating a sequence-structure relationship of glycans. In all these glycofibrils, with or without protein components, the glycan moieties assemble to high-order structures with specific intra- and inter-chain contacts [22,23].

Despite the advances in the CryoSeek identification and structural determination of glycofibrils, the following limitations have prevented us from a comprehensive understanding of the sequence-structure relationship of glycans. First, the resolutions for the majority of the glycofibrils are around 3 Å, insufficient to support accurate assignment of the sugar moieties. At this range of resolutions, neither chemical modifiers nor water molecules can be resolved. Second, the origins and the identities of these glycofibrils remain to be characterized. Such information is crucial for dissecting the biogenesis pathway and to probe the biological functions of these fibrils. It is therefore necessitated to establish appropriate model organisms and systems for glycan-related characterizations using integrative methodologies [24]. To address these challenges, we have focused on mastigonemes, which are lateral appendages on the cilia or flagella of protists and zoospores [25-28], to decipher the sequence-structure-function relationship of glycans.

Based on morphological characteristics, mastigonemes are classified into two categories: tubular and non-tubular [28]. Tubular mastigonemes (T-mastigonemes) are generally

longer, thicker, and more rigid than their non-tubular counterparts (NT-mastigonemes). Our structural investigations of the NT-mastigonemes from the green alga *Chlamydomonas reinhardtii* have yielded several serendipitous findings, including a large array of well-resolved arabinoglycans, the identification of ~ 8000 residue-containing glycoprotein whose transmembrane domain may interact with PKD2 (polycystic kidney disease), and, most interestingly, the 5',5'-phosphodiester bonds that connect the adjacent glycan chains to form a lattice [29,30].

These structural findings raise many interesting questions and establish the framework for functional and mechanistic characterizations. Meanwhile, we have attempted to identify additional systems for glycan studies and sought to determine the structure of T-mastigonemes (Fig. 1A). Experimental and mathematical modelling suggest that T-mastigonemes function in the direction and hydrodynamics of the locomotion of some unicellular protists [31-33]. Removal of T-mastigonemes on the anterior flagellum could alter the swimming direction, and changes in the length and rigidity of T-mastigonemes would affect the propulsion by flagella [34]. T-mastigonemes are known to consist of multiple glycoproteins, yet their assembly mechanism and the roles of glycans remain poorly understood [35].

Here we report the cryo-EM structural analysis of the T-mastigonemes isolated from *Ochromonas danica*, a golden alga species, at resolutions at 1.8-2.2 Å. The high-resolution 3D EM reconstruction unveils two previously uncharacterized glycoprotein components. More importantly, various types of N- and O-glycans, including a non-canonical N-linked

glycan and the PSXX glycofibrils, were resolved to near atomic resolutions. We developed EModelG, an AI-guided pipeline for automated glycan model building, to place sugar chains into the cryo-EM reconstructions. The high-resolution maps unambiguously and directly reveal the sugar identities and covalent modifications, which have been confirmed with mass spectrometric and chemical analysis. In addition, an extensive network of ordered water molecules that account for >10% of the modeled volume is resolved. Our integrative workflow establishes *Ochromonas danica* to be a potential model organism to study the sequence-structure-function relationship of structured glycans, and also demonstrates the feasibility of routine, scalable, and high precision structural resolution of complex native glycan assemblies that can be used in other systems.

Results

High-resolution cryo-EM analysis of the central tube of T-mastigoneme

T-mastigonemes were directly isolated and purified from the long flagella of *Ochromonas danica*. Details of the mastigoneme preparation and cryo-EM data acquisition and image processing are provided in Methods. A total of 45,080 micrographs were processed using cryoSPARC [36]. For high-resolution structural determination, we focus on the central tube region (Fig. 1A). Eventually, the 3D reconstruction was obtained at an overall resolution of 2.17 Å for the tube wall, with the highest local resolutions reaching 2.0 Å (Fig. 1B, Figs S1-S3). The tube surface is covered with bushes of thin spikes. Following a standard protocol for data processing of helical filaments, 3D reconstructions were obtained for these spikes at an overall resolution of 2.19 Å, with the highest local resolution reaching 1.8 Å (Fig. S2).

The outer and inner diameters of the central tube are approximately 20 nm and 9 nm, respectively (Fig. 1C). The shell is arranged as a helical assembly, with a rise of 2.82 nm and a twist of 105.4 degrees between the repeat units, which each consist of six protein subunits, known as the *Ochromonas* mastigoneme-related (OCM) proteins. For clarity of presentation, we display the structure in groups of three repeating units, corresponding to a cumulative rise of 8.5 nm (3 × 2.82 nm) and net rotation of 43.8°. Among these, OCM1-OCM4 have been reported as the main components of T-mastigonemes [35]. Two additional components were revealed in the high-resolution map, which we name OCM5 and OCM6 (Fig. 1D).

The six OCM proteins all contain four highly conserved epidermal growth factor (EGF)-like domains, each stabilized by three pairs of disulfide bonds and adopt a fold similar to the disulfide bond repeat (SSR) structure found in Mst1 of NT-mastigonemes (Fig. 1E) [29]. The EGF strings from adjacent OCM proteins, both within and between the repeat units, align in anti-parallel (Fig. 1F). In addition to the EGF string, OCMs 1, 2, 5 and 6 each contain different numbers of the glycan-rich region (GRR) domain, three in OCM1 and OCM6, and one in OCM2 and OCM5 (Fig. 1E). The EGF strings from the OCM proteins constitute the inner wall of the tube. The multiple disulfide bonds in EGF-like domains likely contribute to the stability and rigidity of the wall (Fig. 1F). The GRR domains attach to the outer surface of the EGF wall, positioning the glycans on the exterior to be exposed to the aqueous milieu (Fig. 1G).

EModelG for automated structural modeling of glycans

The high-resolution cryo-EM map affords an opportunity to resolve a large number of N- and O-glycans with detailed information on the sugar type, linkage, and stereochemistry. We developed a unified workflow that brings together high-resolution cryo-EM, glycoproteomics, and an AI-based modelling platform to perform near atomic-level modelling of complex glycans (Fig. 2A).

To facilitate model building of sugar residues, we have developed an AI-based method, EModelX-Glycan (EModelG), for carbohydrate-aware automated structural building from cryo-EM maps (Fig. 2B). EModelG is based on the prior framework EModelX [37], which was developed for automated protein structure modeling. Starting from a raw 3D map, EModelG automatically identifies protein and carbohydrate densities, models these components in a decoupled manner, and integrates them into a single, self-consistent atomic model. This strategy supports assemblies spanning a broad compositional range, from protein-only complexes to glycan-rich glycoproteins, in which carbohydrates account for >95% of the molecular mass. Protein sequences, structural templates, and monosaccharide types can be supplied as additional inputs to improve the modeling performance (Fig. 2B).

The pipeline begins with voxel-wise identification of protein and carbohydrate components from the cryo-EM map using multiple 3D U-Net networks (Fig. 2C). Protein components are then modeled following the EModelX (w/o seq) protocol, while amino acid-type predictions (e.g., Asn, Ser, Thr), together with protein-glycan contact information,

are used to infer glycosylation sites that define the starting points for glycans.

For glycan modeling, monosaccharide templates are placed and fitted into the density via SE(3) (special euclidean group in 3 dimensions) sampling and gradient-based density fitting, generating an initial set of "glycan seeds". These seeds subsequently undergo combinatorial optimization of glycosidic linkages and monosaccharide identities. To address the combinatorial complexity, a greedy-inspired heuristic is employed, enabling efficient convergence to a chemically plausible structure with good density fitness (Fig. 2C, *lower right*).

The high accuracy of the final models relies on the robust performance of the underlying neural networks. We evaluated monosaccharide-site detection using precision–recall analysis (Fig. 2D). Here, precision is defined as the fraction of predicted monosaccharide sites lying within 4 Å of a modeled monosaccharide, whereas recall measures the fraction of modeled monosaccharides that have at least one predicted site within 4 Å. The area under the precision–recall curve (AUPRC) reaches 0.891, and even at a recall of 0.97 the precision remains above 0.96, confirming a high robustness in identifying the sugar residues.

We further evaluated the performance of EModelG by comparing the auto- and manually-built glycan models. Across the full glycan chains, EModelG's correlation coefficient (CC) and RMSD values approached those of the manually constructed model (Fig.

S4B, S5). For example, at positions 13 and 14, EModelG correctly modelled rhamnose and fucose, respectively (Fig. S4B). The markedly enhanced modelling efficiency in EModelG has offered great help to our model building.

To validate the structural modelling, we employed glycoproteomic approach using liquid chromatography-tandem mass spectrometry (LC-MS/MS) [17,38]. The combination of high-resolution cryo-EM analysis and glycoproteomics has proven to be a highly effective approach for uncovering the precise composition and linkages of glycans in their physiological context (Fig. S6-S16). These results demonstrate that near-atomic resolution cryo-EM maps can fully resolve glycan branching architecture down to stereochemical details. Thus, our integrative approach, combining cryo-EM, glycoproteomics, AI-facilitated auto-modeling, provides a convenient framework for the discovery and structural characterization of complex glycans. In the following, we will present detailed analysis of the structural findings.

Identification of a non-canonical glycosylation site

The high-resolution EM analysis has not only revealed the composition of the branched N-glycan chains linked to a few canonical NXS/T motifs found in the GRRs, but, unexpectedly, elucidated a non-canonical glycosylation site that occurs on the loop between the EGF1 and EGF2 domains in OCM3 and OCM4 (Fig. 3A). The glycan is linked to the Asn residue in the sequence Asn-Ala-Asn-Asp-X-His (AND). OCM3 and OCM4 share a high degree of sequence similarity, except that OCM3 consists of an extra N-terminal sequence of

PSXX repeats (Fig. 3A). Unlike the canonical N-glycans, which contain the conserved $Man_3GlcNAc_2$ core pentasaccharide, the AND N-glycans have a relatively simple composition with a linear chain of D-mannose connected through α-1,2 glycosidic linkages following the starting D-xylose. There are three and six D-mannose residues in the AND N-glycans in OCM3 and OCM4, respectively (Fig. 3A, S6-S7).

The AND motif is on the turning point of a loop that adopts a distinctive ∞-shaped conformation. A His residue positioned at the point of intersection of the loop forms extensive H-bonds with surrounding residues as well as the π-π stacking with a nearby aromatic residue to stabilize the loop conformation (Fig. 3B). To examine if the AND glycosylation may exist in other species, we conducted a Foldseek search against experimental and predicted structures using the coordinates of OCM3 and OCM4 as templates [39], and identified AND motif-containing proteins in eleven additional species (Fig. S17). Among the Ochrophyta and Oomycota species, *Phytophthora nicotianae* is a broad-host-range plant pathogen [40-42]. While some sequences have subtle variations from AND, all possess the critical Asn, His, and Trp/Phe in the corresponding structural loci, suggesting potential structural conservation (Fig. 3C, Fig. S17).

Type-specific structural roles of the N-glycans

The linear AND N-glycans are clearly resolved in OCM3 and OCM4, suggesting stable interactions with surrounding components. Indeed, they contribute to the inter-protein assembly within a repeat unit. The D-xylose and the multiple D-mannose residues form a

helical structure that lines the interface of the two halves of a unit, OCM1-OCM3 and OCM4-OCM6 (Fig. 3D). In OCM1-OCM3, the helical AND N-glycan, maintained through direct and water-mediated hydrogen bonds (H-bonds), emanates from OCM3 and connects OCM1-2. The same structural conformation and interaction patterns mirror in OCM4-OCM6 (Fig. 3D, insets). We hence refer to this type of structural glycans as the "glycostitches". The linear helical D-mannose residues are connected through α-1,2 glycosidic linkages, resulting in maximized interactions between the glycan chains and adjacent proteins in a symmetrical manner.

In addition to the relatively simple, linear AND N-glycans, two types of canonical NXS/T glycans are observed in the GRRs, the high-mannose type and the complex type (Fig. 4A). Both types comprise the conserved core pentasaccharide $Man_3GlcNAc_2$ (Fig. 3A, 4B). Whereas only D-GlcNAc and D-mannose are found in the high-mannose type N-glycans, additional D-xylose and 6-phospho-D-mannose (6-Pi-Man) are present in the complex type N-glycans. Sugar linkages are substantially more diverse in the complex-type, featured with two core D-mannose residues, each with a unique tetrahedral configuration through bond linkages to four sugar residues (Fig. 4B, Fig. S18).

Complex type N-glycans form clusters on the protein surface, which we refer to as the "glyco-islets" (Fig. 4A, Fig. S18). These well resolved glyco-islets may acquire their structural stability from the core tetrahedral D-mannose and the 6-Pi-Man; the latter forms a stable chelation with metal ions together with amino acid residues from the protein scaffold

(Fig. S18A). The glyco-islet arrays distribute uniformly across the T-mastigoneme, increasing its surface area and hydrophilicity. Presence of these glyco-islets may also prevent potential protease attack of the OCM proteins.

The high-mannose-type N-glycans form "glycobridges" that not only stabilize the internal assembly of the repeat unit, but also establish extensive interactions between adjacent repeats (Fig. 4A, Fig. S19). For instance, the large and branched N411-glycan chain in OCM6 stretches out, forming extensive direct and water-mediated interactions with amino acid residues from OCM4 and OCM5 within the same repeat (Fig. S19, right inset). On the other side, the N240-glycan in OCM6 stacks against the GRR of OCM1' with a large number of direct and water-mediated H-bonds (Fig. S19, left inset).

Conservation of glycobridge-like structures across species

The high-mannose N-glycans represent a predominant form of N-glycosylation that is found in a wide range of biological systems. A structural similarity search, using the glycobridge structure N240-glycan in OCM6 as template, retrieved 91 entries from Protein Databank (PDB) (Fig. 4C-E, table S2). All these glycan structures, including the high-mannose glycobridges in the T-mastigoneme, can be superimposed with root-mean-square deviation (RMSD) values of less than 2.0 Å (Fig. 4C). Most of these high-mannose N-glycans are involved in interactions with proteins and play critical roles in recognition and immune responses. For example, N322-glycan in gp120 in HIV-1 is known to participate in antibody recognition [10]. It exhibits a spatial conformation nearly identical to that of N240-glycan in

OCM6, with the RMSD of 0.93 Å over 62 atoms (Fig. 4D).

Evolutionary analysis of the glycoproteins that host these 91 identified glycan structures reveals their widespread distribution across viruses, prokaryotes, fungi, plants, and mammals. These findings suggest that similar type of glycans may adopt relatively stable conformations for protein recognition. The recurring structural motifs, which are stabilized by extensive intra- and intermolecular H-bond networks, likely originated early in evolution, providing a potential framework for identifying generalizable principles that govern glycan–protein interactions (Fig. 4E).

In contrast to the large number of glycobridge structures, a search of the PDB did not yield any structures resembling the complex type of N-glycans. This result aligns with the absence of standalone N-glycan structures, as they are typically observed only in complex with their protein binders. Moreover, since the AND non-canonical N-glycans were discovered only in our present study, no analogous structure exists in PDB either.

The surface glycofibrils made of PSXX glycofibrils

The spikes emanating from OCM3 can exceed 150 nm in length (Fig. 5A). The high-resolution map supports unambiguous identification of PSXX tetrapeptide repeats that constitute these spikes. Similar to our previously reported glycofibrils TLP-4a/b [20,21], the protein portion remains a linear peptide chain, and the fibrillar structure is maintained by the packing of glycans, which account for more than 95% of the total molecular mass of the

glycofibril. The fibril is thinner at the proximal end. This arrangement avoids potential clash with the rest of the T-mast wall (Fig. 5A).

Glycans spiral around the peptide backbone with a helical rise of 12.4 Å and a helical twist of -138.4°. As each tetrapeptide repeat is approximately 1.2 nm long, a fibril comprises over 100 repeats (Fig. 5A, B, Fig. S1B). Thanks to the high resolution, a large-scale water-mediated H-bond network within a glycofibril can be clearly observed. Water-associated densities account for over 10 % of the total modeled volume and make substantive contributions to structural stability (Fig. 5B, right).

The map also reveals that in PSXX, Pro is in the form of (3S, 4S, 5S)-3,4,5-trihydroxyproline (THP) (Fig. 5C). Previous work has documented the presence of di-hydroxyproline (diHyp) in certain diatoms via mass spectrometry [43], but never THP. Using LC-MS/MS, we verified the existence of THP by detecting the corresponding tri-hydroxylated molecular ion peak. Our structural observation thus expands the knowledge on the diversity of proline modifications (Fig. S20).

THP and Ser are modified by two and one O-glycans, respectively. The excellent high-resolution EM maps reveal the precise identity of each and every sugar residue in the glycans. Compared to N-glycans, the sugar composition in these O-glycans is markedly more diverse. In total, we have observed and confirmed 8 types of sugars, including L-fucose, D-galactofuranose (D-Gal*f*), D-glucose, D-galactopyranose, L-rhamnose, D-xylose,

N-acetyl-D-glucosamine, and D-mannose. The glycans attached to Ser contain fewer sugar units, 15 for the repeats at the distal end, compared to 22 each in the two glycans attached to THP (Fig. 5D, E). Nevertheless, the Ser-linked glycans appear to define the overall structural outline for the glycofibril (Fig. 5B).

In each THP, the two hydroxyl groups project in opposite directions, placing their respectively linked glycans in complementary physical spaces (Fig. S21C, right). The glycans linked to THP4 interweave with the Ser-glycan in the same PSXX repeat via extensive H-bonds (Fig. S21C, left). The three D-Gal*f* residues in the Ser-glycan, two consecutive and one on a different branch, point to a different direction from the bulk of the Ser-glycan, resulting in a position to be cradled by the horse-shoe shaped THP-glycan (Fig. S21C, left). This extensive hydrophilic interaction network facilitates a compact folding of the entire glycofibril (Fig. S21B).

Diverse modifications of the sugar residues in the PSXX glycans

In addition to revealing the identity of regular sugar residues, the high-resolution EM density allows precise identification and localization of covalent modifications. We have observed the following covalent modifications: methylation at positions 2, 3 or 4; acetylation at positions 2, 3 or 4; mixed methylation and acetylation at positions 2 and 3; di-acetylation at positions 2 and 3; and sulfation at positions 3 (Fig. 6A, B).

Among these, acetyl groups extend the interaction range of the host sugar rings,

without severely compromising hydrophilicity. By bridging branches on the same glycan or between adjacent glycan chains, sugar acetylation enhances the structural continuity (Fig. 6A, 6C, top). Methylation appears to suppress H-bond formation at specific monosaccharide positions in addition to marking an end to glycan chain extension (Fig. 6A, 6C, middle). The presence of sulfation in the mastigonemes has been documented [44,45]. Our structure shows that all three core glycan chains harbor sulfate modifications within their central segments. This highly hydrophilic tetrahedral moiety substantially increases the potential for water-mediated H-bond networks. For instance, in THP4-glycan, the sulfate engages in stable H-bond networks with two neighboring chains; in Ser-glycan, it cooperates with the acetylated branches to form a more complex network of interactions (Fig. 6C, bottom).

Cations serve as coordinating centers to stabilize the structural core

In addition to the diverse forms of covalent glycan modifications and the extensive water-mediated interaction network, the high-resolution EM map also reveals densities corresponding to cations (metal ions) along the core of the PSXX strand (Fig. 7A). These ions are coordinated by surrounding sugars, amino acids, and water molecules. The highest resolvable coordination number reaches eight. We speculate that some of the ions might be K^+.

Notably, coordination of one cation can engage glycans from multiple PSXX repeats. For example, we name three repeats, from the N to C terminus of the peptide chain, n-1, n, and n+1. Coordination of one ion involves the first D-Man on the THP3-glycan of repeat n, a

D-Glc on the Ser-glycan of repeat n-1, and a D-Gal on the THP4-glycan of repeat n+1 (Fig. 7A, B). Therefore, at least three successive PSXX repeats and their associated glycans are required to complete the coordination of one ion. On the other hand, the cation serves as an organizing center to stabilize the surrounding groups from three repeats. The remaining chelating groups involve a water molecule, the 5-hydroxyl group of THP (n), and the backbone carbonyl oxygen and the glycosidic oxygen (Fig. 7B, C). In particular, without the 5-hydroxyl group of THP (n), stable octahedral chelation could not be completed.

Of note, the high-resolution map unambiguously shows that D-Man (n) does not adopt the stable chair conformation observed for other sugar residues in the assembly. Instead, its C2 and C3 atoms undergo a flip relative to the standard chair conformation, adopting a geometry resembling a twist-boat conformation (Fig. 7D). Only in this specific conformation can the geometry permit coordination of the cation, which in turn may partially offset the energetic penalty associated with deviation from the chair conformation (Fig. 7D). Therefore, the multitude of ions distributed along the core of the glycofibril serve as the organizing nuclei that strengthen the entire assembly.

Discussion

The sequence-structure projection of proteins has been relatively well established, culminating in the success of AlphaFold [46]. By contrast, there are still formidable challenges in sequencing polysaccharides and resolving their high-resolution structures. Our study presented here demonstrates that both sequences and structures of glycans can be elucidated

directly when the resolution of a 3D EM reconstruction is sufficiently high.

Compared to the NT-mastigoneme, in which the glycans comprise of D-arabinose and D-Gal*f* as revealed in the ~ 2.3Å-resolution structure [47], the diversity and complexity of the glycans on the T-mastigoneme is remarkable. In total we have observed three types of N-glycans and uncovered their type-specific structural roles. The linear and mannose-rich AND N-glycan, which has never been characterized, exhibits a helical structure that stitches neighboring protein domains, the high-mannose type tends to bridge long-range interactions, and the bulky complex type clusters on the surface like islets. These glycans not only increase the surface area and water solubility, but directly contribute to the stability of large macromolecular assemblies. N-glycans have been characterized to function in signaling in intracellular trafficking and surface recognition [48-51]. The findings presented in this study have expanded our knowledge on the diversity and structural roles of N-glycans.

The discovery of the PSXX type of glycofibrils on the surface of the T-mastigonemes was entirely serendipitous We have been struggling to identify the origins for TLP-4a and TLP-4b ever since they were found in the pond water from Tsinghua University [20,21]. The PSXX glycofibrils are different from TLP-4a/b in length and branching patterns. Nevertheless, the presence of this type of glycofibrils in *O. danica* offers, especially with the unambiguous resolution of the sugar identities, establishes a tractable system to study the sequence-structure-function relationship of O-glycans.

O-glycans range from the short, GalNAc-initiated core structures common in animal mucins to the extensive arabinogalactan polymers found in plant and algal glycoproteins [52-54]. The PSXX glycofibrils showcases how O-glycosylation can dominate higher-order assembly. An immediate question concerns the folding principles of such glycans. Comparison of the three PSXX glycofibrils indicates a conserved spatial core despite substantial variations in the distal branch number and extension geometry (Fig. 8A, B). This modularity implies that the PSXX tetrapeptide repeat provides a portable sequence framework that specifies a reproducible glycan packing arrangement, while peripheral elaborations diversify surface chemistry and potential interactions. Consistent with this view, phylogenetic analysis reveals PSXX repeats distributed across multiple ochrophyte lineages with variable organization (Fig. S22, Table S3), supporting PSXX-like scaffolds as an evolutionarily flexible solution for extracellular or surface assemblies.

In this study, we mainly focus on the structural illustration of the glycans, as the information embedded in the high-resolution EM maps of native glycans with complex sugar compositions is enormous. The structural findings have laid a foundation for future investigation from multiple perspectives. For example, elucidation of the covalent modifiers on the sugar rings, the structured water molecules and ions fills an important void in understanding glycan structural assembly. It provides the framework for computational modeling and molecular dynamics simulation (MDS). As glycans are highly hydrophilic, their interchain contacts are frequently mediated by transient H-bond networks. The contributions of water to glycan folding have largely been inferred from quantum chemistry

and MDS [55]. Our high-resolution maps, which allow for direct visualization of hundreds of structural water molecules, along with the cations, provide an important ground truth to support future characterizations of glycan folding. Glycans can form non-conventional H-bonds and hydrophobic interactions mediated by alkyl hydrogen atoms, both of which play critical roles in glycan folding and stability [56]. Combining quantum chemical computations on the basis of high-resolution experimental structures, the interactions involving these hydrogen atoms can be computationally characterized to advance our understanding of fundamental glycan folding principles.

The structural discoveries naturally lead to many outstanding questions. Prominent among these are the elucidation of the biogenesis pathway for these glycans and the identification of the enzymes responsible for catalyzing sugar residue modifications and synthesizing 3,4,5-trihydroxyproline. The identification of these enzymes, likely requiring a multifaceted methodological approach, is crucial for deciphering the biological functions and assembly principles of the glycans, both for the mastigonemes and in broader contexts.

Acknowledgements

We thank the Shenzhen Medical Academy of Research and Translation (SMART) for providing the cryo-EM facility support and the computational facility support. We thank Gang Fu and Xiaosong Pei for technical support during EM data collection. Data analysis is supported by the Computing Labware for Electron-microscopy Visualization and Experimental Research (CLEVER) of Shenzhen Medical Academy of Research and

Translation (SMART). We thank Dingfei Yan and Haiteng Deng in Center of Protein Analysis Technology, Tsinghua University, for MS analysis.

Funding:

This work was funded by the National Natural Science Foundation of China (92478205, 32330052, N. Y, 32341016, 32171204, C.Y., and 324B2042, J. H), the National Key R&D Program of China (2020YFA0509301, C.Y.), Tsinghua University Initiative Scientific Research Program, and Shuimu Tsinghua Scholar Program. N.Y. thanks the Parkland Foundation for the Mindray Endowed Professorship.

Author contributions:

Conceptualization: N.Y., J.H., C.Y.; Methodology: J.H., H.T., S.C., Y.C., Y.X.; Cryo-EM data collection: J.H.; Structural modelling and refinement: J.H., S.C.; Investigation: J.H., H.T., S.C., Y.C., Y.X.; Funding acquisition: N.Y., J.H., C.Y.; Writing – original draft: J.H., N.Y.

Competing interests:

The authors declare no competing interests.

Data, code, and materials availability:

All data are available in the main text or the supplementary materials. To request the EM maps and structural coordinates, please contact the corresponding authors. For access to the source code, please visit the following link: https://www.cryoseek.org.cn/EModelG_v0.zip

Figures and legends

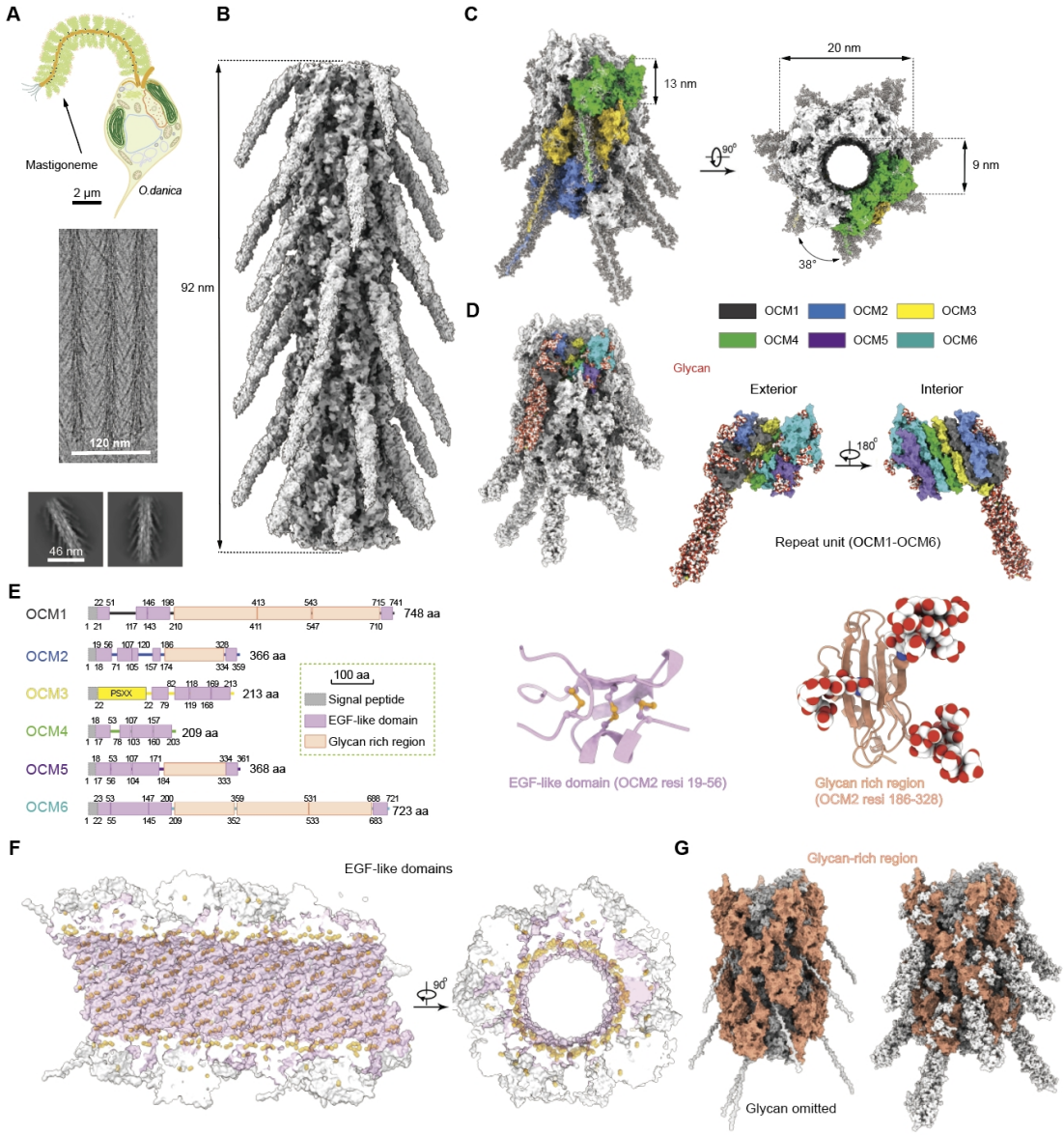

Fig. 1 | Cryo-EM structural analysis of the central tube of native tubular mastigonemes from the golden algae *Ochromonas danica*. **(A)** T-mastigonemes are lateral filaments lining the flagella of various protists. *Middle* and *bottom*: Representative EM micrograph and 2D class averages of native T-mastigonemes purified from *O. danica*. **(B)** 3D EM reconstruction of the central tube of the T-mastigoneme. Shown here is a representative tubular segment. **(C)** Helical assembly of the repeat units that constitute the tube shell. Representative repeat units in adjacent helical turns (~three repeats on turn) are colored green,

yellow and blue. **(D)** Six glycoproteins that constitute the shell for the central tube of the T-mastigoneme. The known OCM1-OCM4 and the EM map-revealed OCM5 and OCM6 form the basic repeat unit for the tube shell. The "spikes" that cover the tube surface are exclusively made by the heavily glycosylated, linear N-terminal fragment of OCM3. Glycans are shown as spheres. **(E)** Domain structures of OCM1-OCM6. The six OCM proteins each contain a signal peptide that is cleaved. They all share a similar core structure that is composed of four conserved Epidermal Growth Factor (EGF)-like domains, each spanning ~40 amino acids and characterized by three disulfide bonds. Shown on the top right is the first EGF-like domain in OCM2. Disulfide bonds are highlighted as orange ball and sticks. In OCMs 1, 2, 5, and 6, extensive glycosylation occurs to another conserved domain, which we name the Glycan Rich Region (GRR). Each GRR is featured with seven β-sheets and multiple glycosylation sites. A representative GRR from OCM2 is shown on the right. OCM1 and OCM6 each contain three GRRs, and OCM2 and OCM5 each have one. The N-terminal region of OCM3 features a unique linear and densely glycosylated fibril made of PSXX tetrapeptide repeats. As the number of repeats may vary, this segment is not counted for residue numbering of OCM3, the rest of which is highly similar to OCM4. **(F)** The EGF sheet from each repeat unit constitutes the inner wall of the central tube of T-mastigonemes. The regularly distributed disulfide bonds (orange spheres) strengthen the assembly of the tube shell. **(G)** The GRRs form patches on the outer wall of the central tube. The dense glycans, which are omitted in the left panel and shown as silver spheres in the right one, substantially enlarge the surface area. They may increase water solubility and protect the protein moieties from protease attack. All EM maps and structure figures were prepared in ChimeraX.

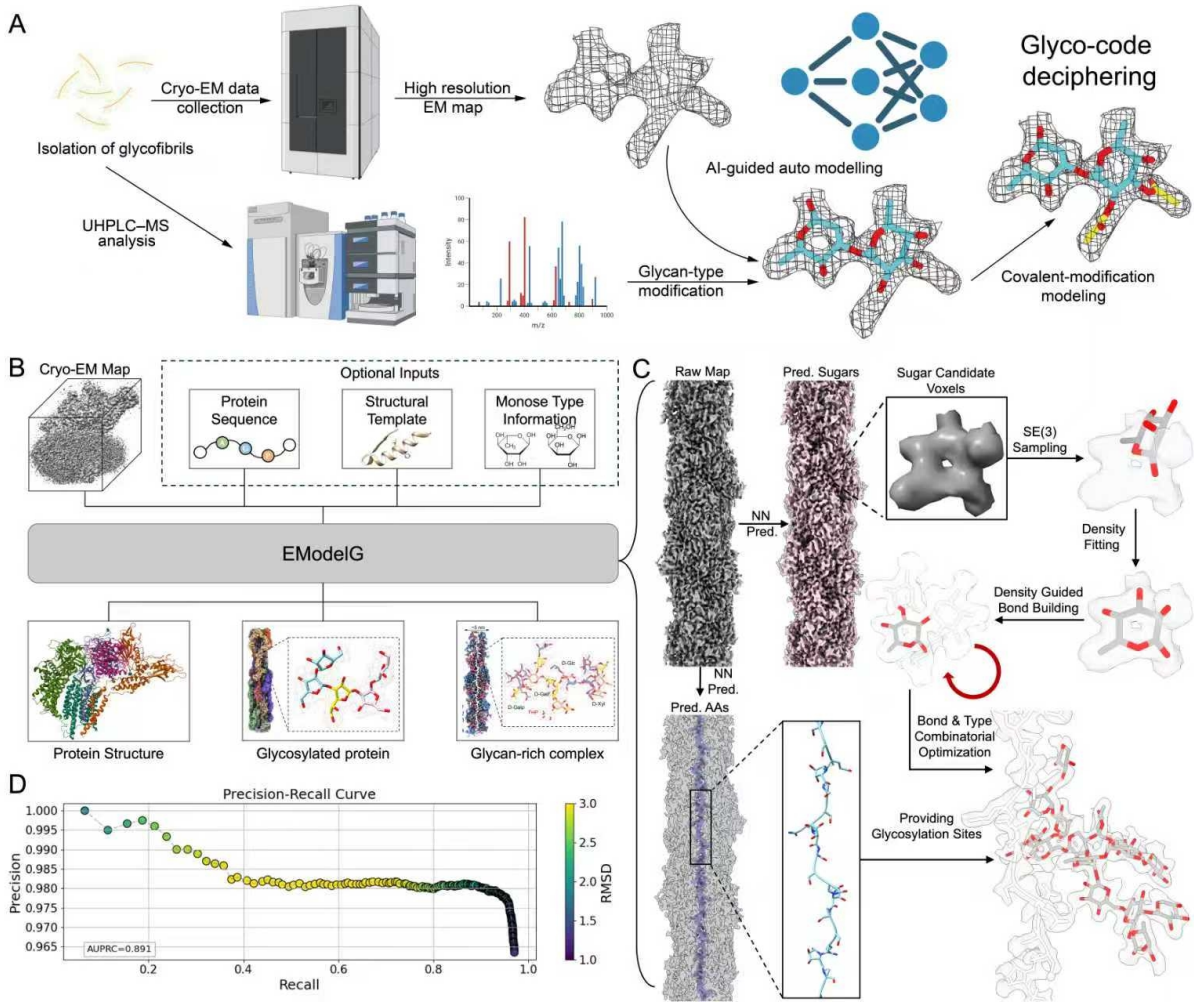

Fig. 2 | An integrative approach for structural determination and AI-driven automated modeling of glycans from cryo-EM and glycoproteomic data. **(A)** A strategy for accurate structural and chemical determination of glycans integrating cryo-EM and glycoproteomics. The panel was generated in BioRender. **(B)** EModelG, an AI-based cryo-EM glycoprotein auto-modeling program. With the input of a raw 3D density map, and optional structural templates, monosaccharide types, as well as protein sequences, EModelG efficiently builds atomic models for proteins or glycan-rich complexes with long, branched glycans. **(C)** Workflow for EModelG. Sites for amino acid and monosaccharide sites are predicted from the density map using neural networks (NN); the former are used for automatic protein model building and glycosylation-site identification, and the latter seed

glycan building. Iterative SE(3) (special euclidean group in 3 dimensions) sampling, density fitting, glycosidic-bond assignment and sugar-type optimization can yield glycoprotein structures with near-atomic-accuracy. **(D)** Evaluation of the precision for model building using EModelG. The precision-recall curve of neural-network is used to evaluate the precision of monosaccharide-site identification in glycofibril assemblies. An area under the precision-recall curve (AUPRC) of 0.891 suggests high accuracy of sugar site prediction. Each point is colored by the RMSD of the auto-built model against the final model.

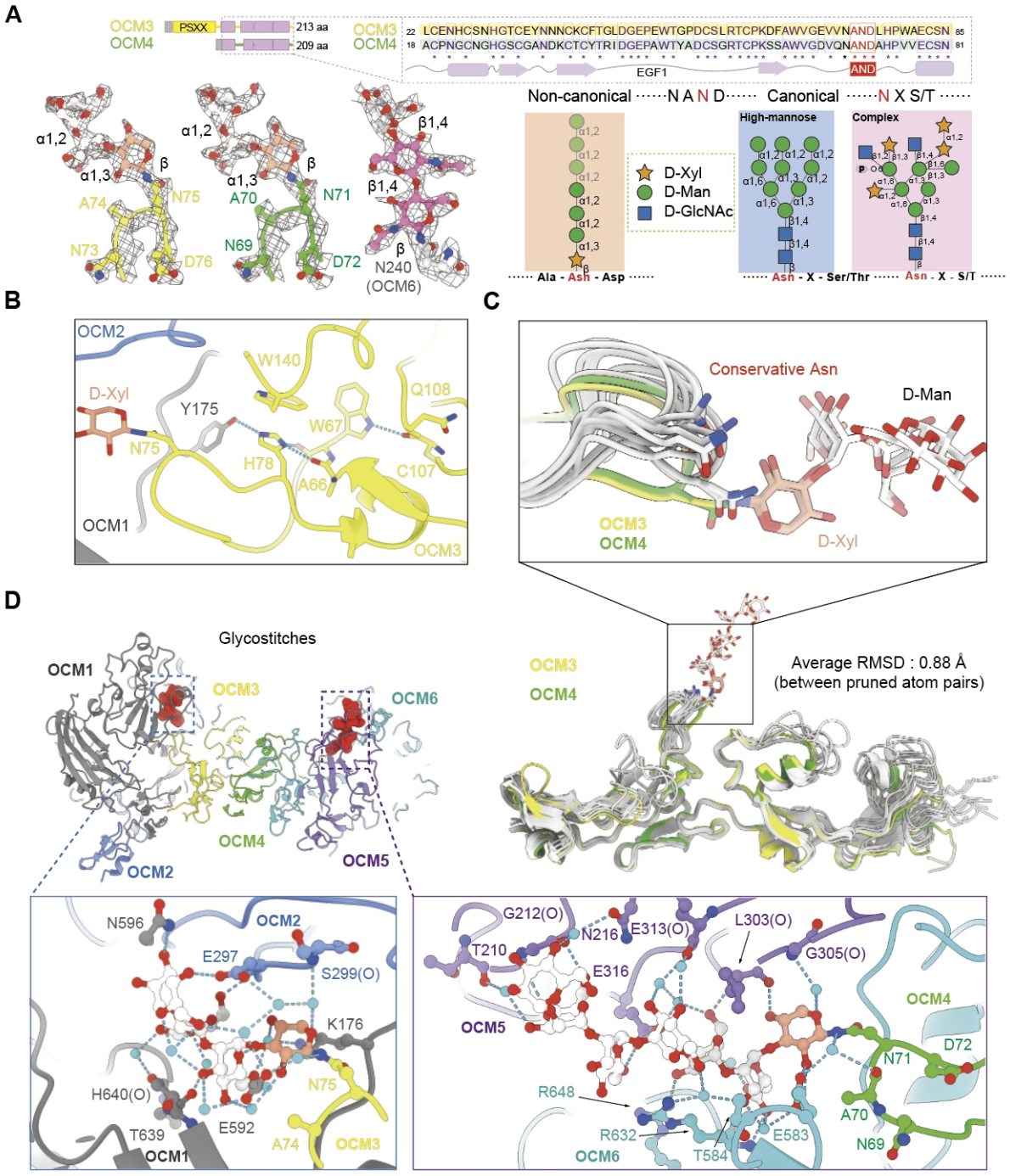

Fig. 3 | OCM3 and OCM4 each contain a non-canonical N-glycan linked to Ala-Asn-Asp.

(A) Characteristics of the non-canonical N-glycans. Glycosylation occurs to the Asn residue in Ala-Asn-Asp (AND), which is conserved in OCM3 and OCM4. Unlike the canonical N-glycan, which occurs to the NXS/T consensus motif and typically begins with a D-acetylglucosamine (D-GlcNAc), the non-canonical N-glycan starts with a D-xylose,

followed by a linear D-mannose chain composed of three and four residues in OCM3 and OCM4, respectively. The EM densities, shown as grey meshes, are contoured at 3.5 σ in ChimeraX. The N240-glycan, which is a canonical high-mannose type glycan, in OCM6 is shown for comparison. **(B)** Key structural characteristics of the AND motif. The AND motif is on a loop that adopts a shape of ∞. A conserved His residue positioned at the intersection point forms multiple H-bonds with surrounding amino acids and forms π-π stacking interaction with a neighboring Trp175, thus stabilizing the overall architecture of the ∞ loop. **(C)** Predicted structures that possess the AND motif. Using the structures of OCM3 and OCM4 as templates in Foldseek [39], nine homologous proteins, each of which can be superimposed with OCM3 with an average RMSD of 0.88 Å over 212 Cα atoms, can be identified. Detailed information of these proteins is listed in Fig. S17. *Inset*: Enlarged view of the AND-containing loops to highlight the conserved Asn residue. **(D)** The extended linear non-canonical N-glycans serve as glycostitches to strengthen the association of the neighboring proteins. *Insets*: The non-canonical N75-glycan in OCM3 (left) and N71-glycan in OCM4 (right) respectively stabilize the interactions between OCM1 and OCM2 and between OCM5 and OCM6.

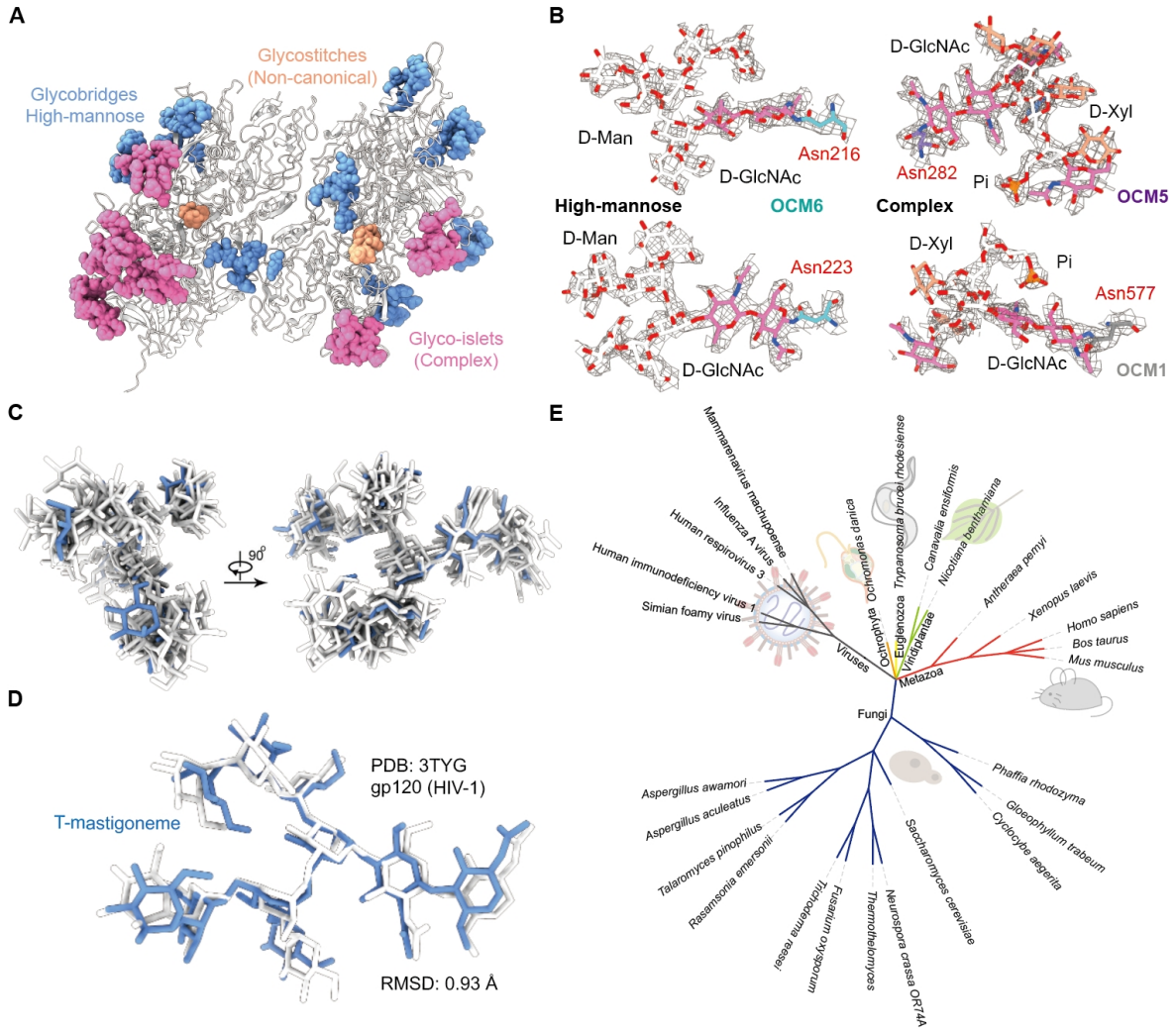

Fig. 4 | Widespread presence of glycobridge-like N-glycan structures across species. (A) Summary of the structural roles of N-glycans. Glycostitches by the non-canonical N-glycans, glycobridges by both types of canonical N-glycans, and glyco-islets by the complex N-glycans are colored orange, blue, and magenta, respectively. Whereas the glycostitches primarily mediate the intra-repeat interactions among different OCM proteins, glycobridges mainly strength the assembly between repeats. Glyco-islets decorate the surface of the T-mastigoneme, substantially increasing its surface areas and water solubility. **(B)** Representative densities and structures of the canonical high-mannose and complex N-glycans. The high-mannose type N-glycans are composed exclusively of D-mannose

residues on top of the conserved core pentasaccharide ($Man_3GlcNAc_2$). Shown here are representative high-mannose N-glycans in OCM6. The complex type N-glycans retain the conserved core pentasaccharide that is linked to NXS/T, but incorporate additional D-xylose (D-Xyl) and feature a conserved 6-phospho-D-mannose (6-Pi-Man). Shown here are two hetero-type N-glycans attached to Asn577 in OCM1 and Asn282 in OCM5. The densities are contoured at 3.5 σ. **(C)** Search of the protein databank (PDB) has retrieved 91 glycobridge-like structures of high-mannose type N-glycans. Shown here is the structural superimposition of N240-glycan in OCM6 (blue) with nine representative entries (colored white to dark grey). The PDB codes for these entries and glycosylation details are listed in Table S2. **(D)** Highly similar structures of the glycobridge-like N-glycans in T-mastigoneme and HIV-1. The glycans attached to Asn240 in OCM6 and Asn332 in gp120 (PDB code: 3TYG) [10] can be superimposed with the RMSD of 0.93 Å over 62 non-hydrogen atoms, indicating a high degree of structural similarity. **(E)** Glycobridge-like structures are distributed across species. Shown here is the phylogenic analysis of 91 selected species that contain glycobridge-like structures. Analysis details can be found in Methods.

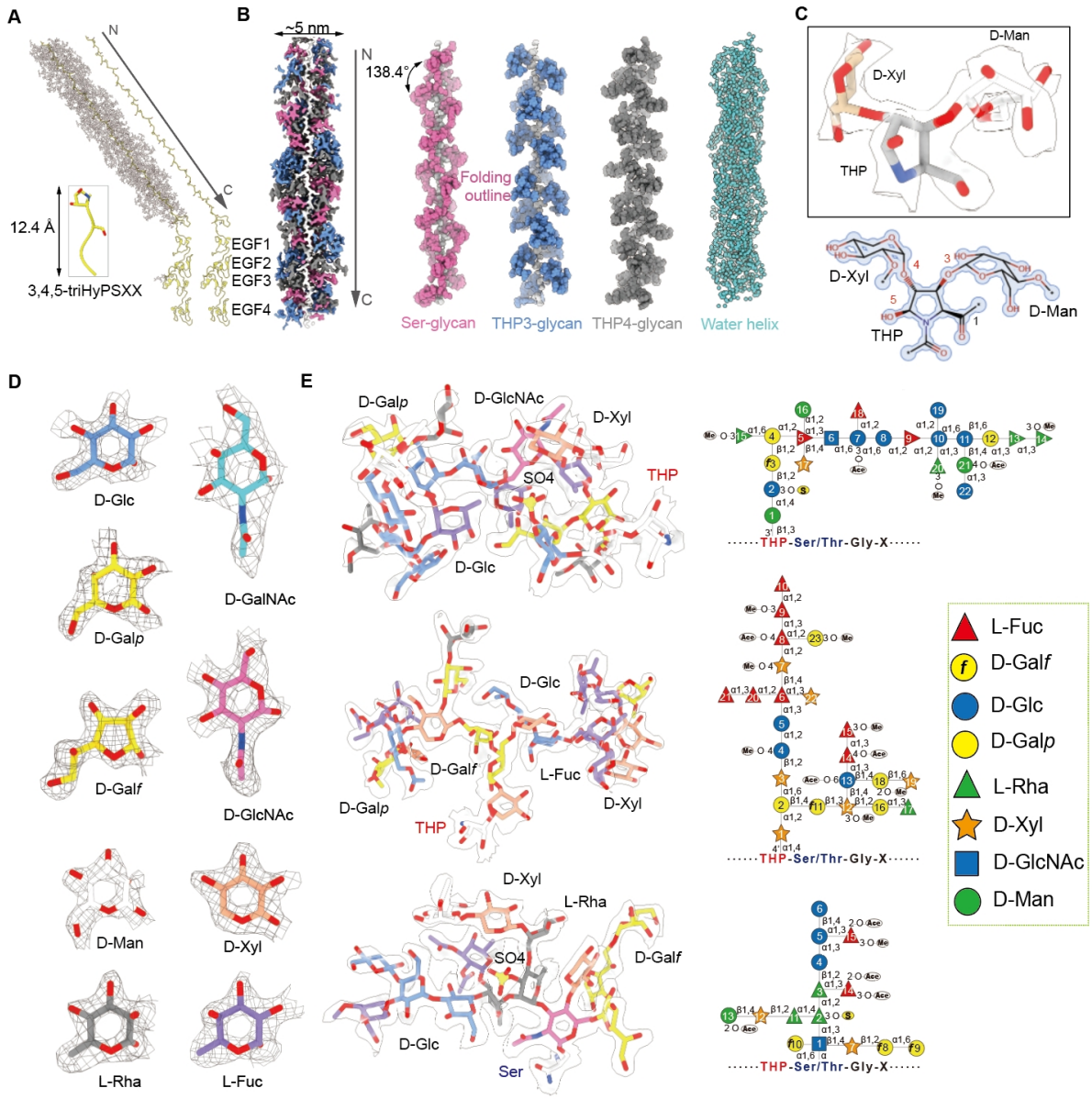

Fig. 5 | Structural assembly of the O-glycans around the linear PSXX tetrapeptide repeats. **(A)** O-glycan interactions mediate the structural assembly of the fibrils on the surface of the T-mastigonemes. The tetrapeptide repeats, which exists as linear chain coated with a dense glycoshell, are extended from the N-terminus OCM3. Each tetrapeptide repeat, containing the sequence of 3,4,5-trihydroxyproline (THP), Ser or Thr, followed by two amino acid residues, is approximately 12.4 Å in length. For simplicity, we still refer them as the PSXX glycofibrils. **(B)** Structural basis for the assembly of the PSXX glycofibrils. *Left*: The PSXX tetrapeptide repeats form the linear stem of the fibril. *Middle three*: The three

arrays of O-glycans, respectively linked to the hydroxy groups at the 3′- and 4′-position of THP (THP3 andTHP4) and Ser in the tetrapeptide repeat, each form a super spiral around the central linear peptide chain. Their extensive intra- and inter-glycan interactions support the structural assembly. *Right*: The well-resolved water molecules occupy ~ 11% of the molecular mass. They play critical roles in mediating sugar interactions. **(C)** 3,4,5-trihydroxyproline identified through high-resolution cryo-EM maps. Two glycans are O-linked to THP3 and THP4 via D-mannose and D-xylose, respectively. The local map and schematic illustration for THP-linked O-glycans are presented on the top and bottom, respectively. **(D)** High-resolution cryo-EM maps support accurate determination of various types of monosaccharides. **(E)** Complex sugar compositions and linkages of the three O-glycans in each repeat of the PSXX glycofibrils. Structural and linkages of the O-glycans attached to THP3, THP4, and Ser/Thr are presented on the top, middle, and bottom, respectively. All glycan linkages conform to the International Union of Pure and Applied Chemistry (IUPAC) standards. The EM maps are contoured at 8.0 σ.

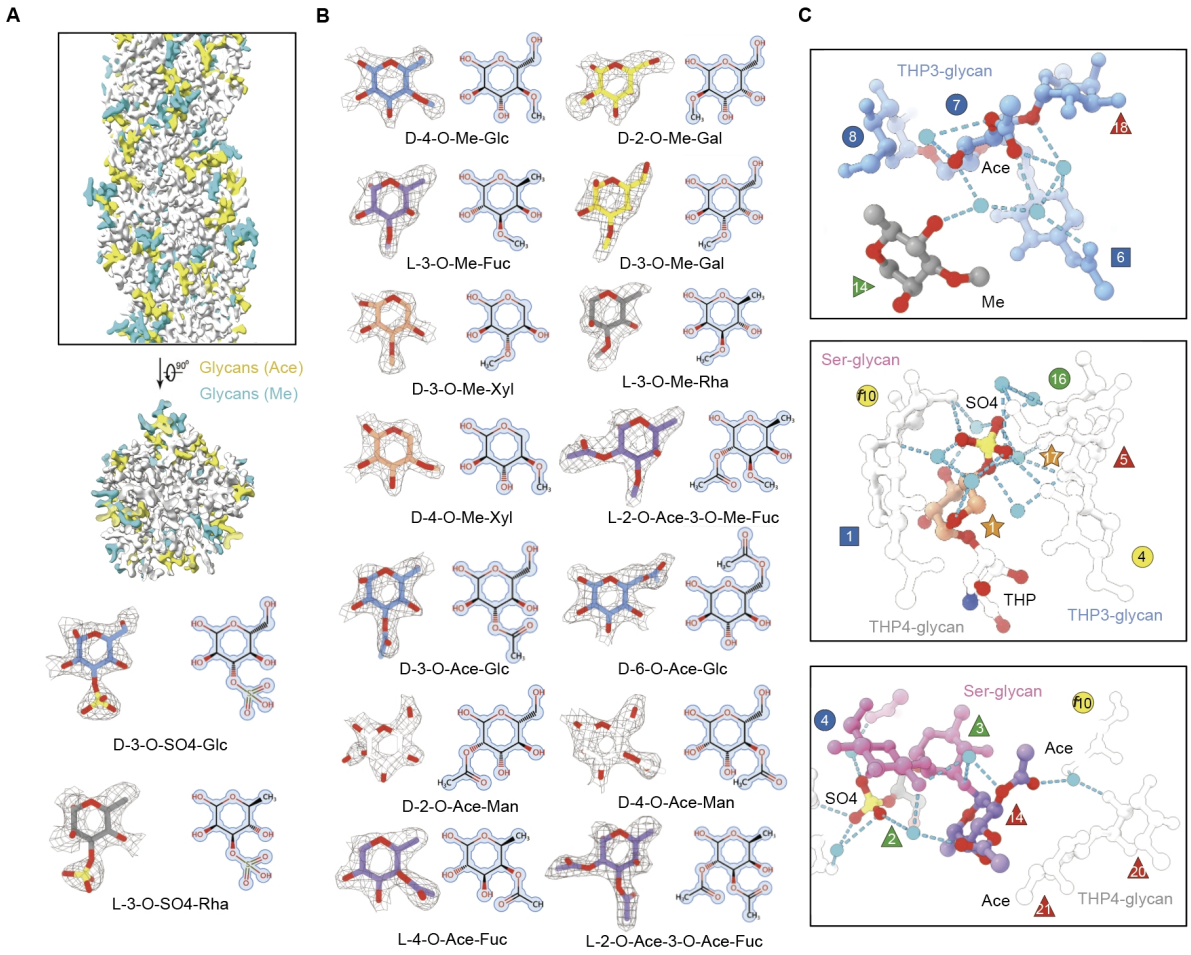

Fig. 6 | Localization and structural role of sugar modifiers in PSXX glycofibrils. **(A)** Distribution of acetyl and methyl groups that modify sugar residues in the PSXX glycofibrils. Acetylated and methylated monosaccharides are labeled gold and cyan, respectively. **(B)** High-resolution EM densities directly reveal the covalent modifiers on sugar residues. Linkages conform to the International Union of Pure and Applied Chemistry (IUPAC) standards. The EM maps are contoured at 8.0 σ. **(C)** Structural role of the covalent modifiers and water molecules in the assembly of PSXX glycofibrils. *Top*: A representative view of the hydrogen-bond (H-bond) network mediated by acetyl groups and water molecules of the THP3 O-glycan. *Middle*: A representative H-bond network mediated by the sulfate group that modifies a D-xylose residue in a THP4 O-glycan and water molecules. *Bottom*: A

representative H-bond network mediated by acetyl groups and water molecules of the O-glycan attached to Ser. Direct and water (cyan spheres)-mediated H-bonds are shown as blue, dashed lines.

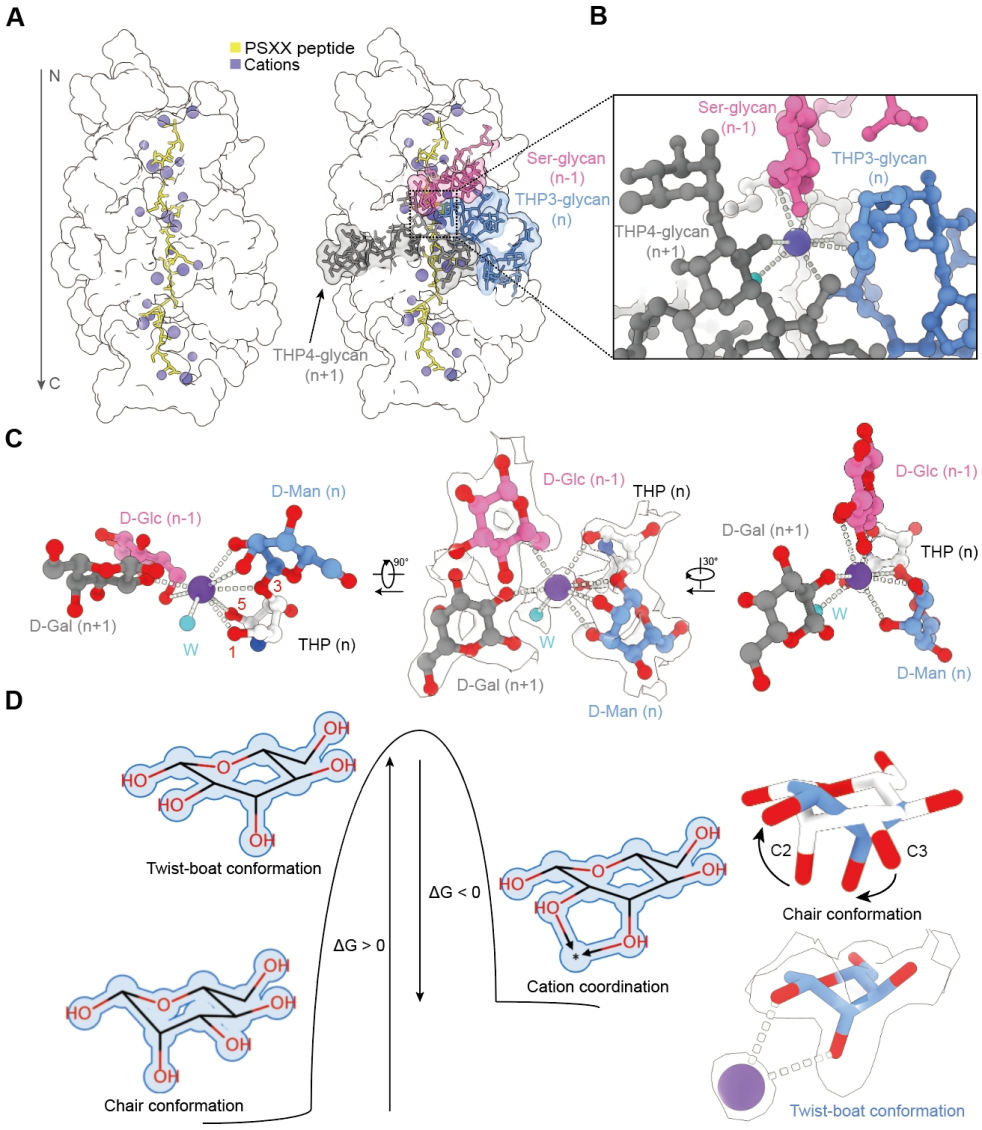

Fig. 7 | Well-coordinated cations stabilize the core assembly of the PSXX glycofibril.

(A) Well-ordered cations along the PSXX core. Ser, THP3, and THP4 glycans in one repeatr are highlighted pink, blue, and gray, respectively, and in other repeats are all simplified in white surface presentation. The strand of the PSXX tetrapeptide repeats is colored yellow, and ions are purple, **(B)** Octahedral coordination of a cation by polar groups from THP, glycans, and water. Ser, THP3, and THP4 glycans colored pink, blue, and gray, respectively. Water is shown as cyan sphere, and H-bonds are shown as gray dashed lines. **(C)** The cation serves as the organizing center for THP and three different glycans from three consecutive

repeats. THP is colored white, and monosaccharides from Ser-glycan in repeat n−1, THP3-glycan in repeat n, and THP4-glycan in repeat n+1 are colored pink, blue, and gray, respectively. Three views are shown. The EM maps are contoured at 10.0 σ. **(D)** A specific "twist-boat" conformation of a D-mannose to complete the cation coordination. Shown on the left is a free-energy diagram to illustrate the role of the cation in stabilizing the energetically unfavored twist-boat conformation of a D-Man. The EM maps are contoured at 10.0 σ.

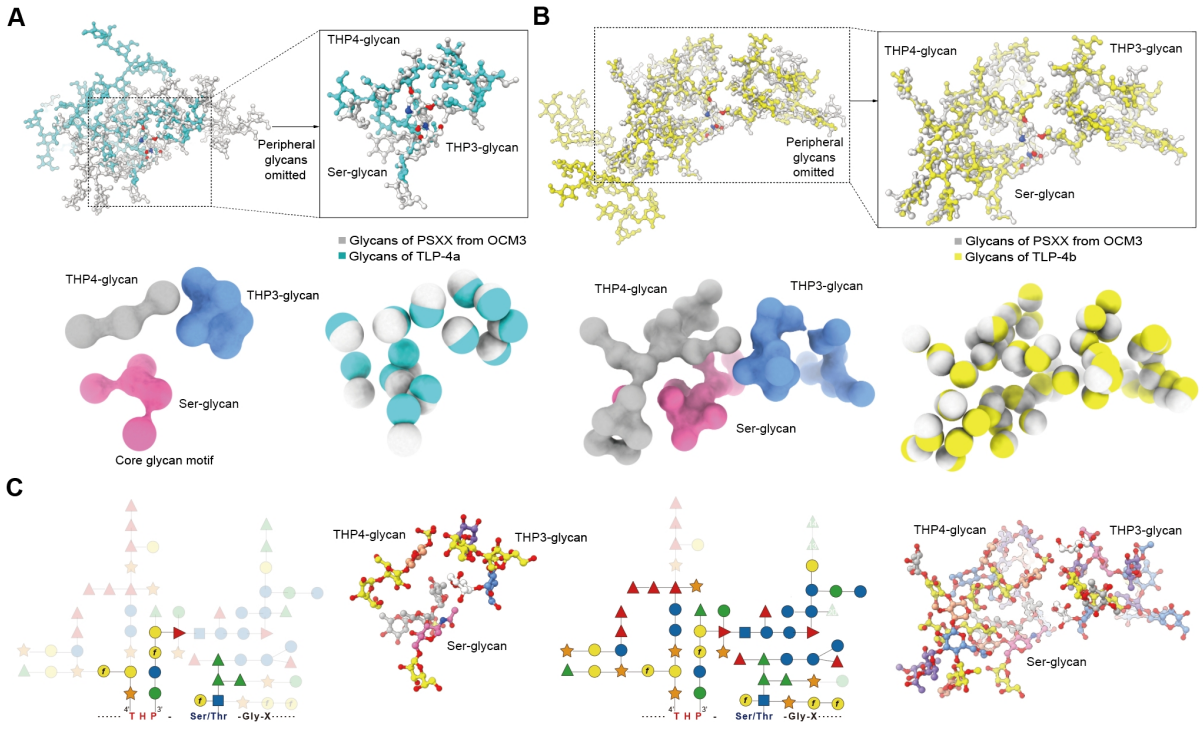

Fig. 8 | Structural characteristics and species distribution of the PSXX glycofibrils.

(A, B) Comparison of the PSXX glycan pattern on OCM3 with TLP-4a and TLP-4b. Each monosaccharide is simplified to a sphere to better compare their spatial arrangement. The sugar linkages that are identical in the two compared glycans are referred to as the core glycan motifs. The core glycan motifs linked to THP3, THP4, and Ser are colored blue, gray, and magenta, respectively. Glycans on OCM3, TLP-4a [20], TLP-4b are labeled gray, cyan, and yellow, respectively, when superimposed. **(C)** Composition of the sugar composition and linkages of the core glycan motifs among the three PSXX glycofibrils. It is evident that the OCM3-PSXX shares higher similarity with TLP-4b than with TLP-4a.

References

1 Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3 – 49 (2017). https://doi.org/10.1093/glycob/cww086

2 Schjoldager, K. T., Narimatsu, Y., Joshi, H. J. & Clausen, H. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol* **21**, 729 – 749 (2020). https://doi.org/10.1038/s41580-020-00294-x

3 Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3**, 97–130 (1993). https://doi.org/10.1093/glycob/3.2.97

4 in *Essentials of Glycobiology* (eds A. Varki *et al.*) (Cold Spring Harbor Laboratory Press
Copyright © 2022 by the Consortium of Glycobiology Editors, La Jolla, California. Published by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. All rights reserved., 2022).

5 Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330–333 (2020). https://doi.org/10.1126/science.abb9983

6 Sztain, T. *et al.* A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat Chem* **13**, 963–968 (2021). https://doi.org/10.1038/s41557-021-00758-3

7 Lundstrøm, J. & Bojar, D. Structural insights into host-microbe glycointeractions. *Curr Opin Struct Biol* **73**, 102337 (2022). https://doi.org/10.1016/j.sbi.2022.102337

8 Angata, T., Nycholat, C. M. & Macauley, M. S. Therapeutic Targeting of Siglecs using Antibody- and Glycan-Based Approaches. *Trends Pharmacol Sci* **36**, 645 – 660 (2015). https://doi.org/10.1016/j.tips.2015.06.008

9 Tang, H. *et al.* Glycan motif profiling reveals plasma sialyl-lewis x elevations in pancreatic cancers that are negative for sialyl-lewis A. *Mol Cell Proteomics* **14**, 1323 – 1333 (2015). https://doi.org/10.1074/mcp.M114.047837

10 Pejchal, R. *et al.* A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* **334**, 1097–1103 (2011). https://doi.org/10.1126/science.1213256

11 Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730–738.e713 (2020). https://doi.org/10.1016/j.cell.2020.09.018

12 Molinos-Albert, L. M. *et al.* Anti-V1/V3-glycan broadly HIV-1 neutralizing antibodies in a post-treatment controller. *Cell Host Microbe* **31**, 1275 – 1287.e1278 (2023). https://doi.org/10.1016/j.chom.2023.06.006

13 Hu, X. *et al.* Structural and mechanistic insights into fungal β-1,3-glucan synthase FKS1. *Nature* **616**, 190–198 (2023). https://doi.org/10.1038/s41586-023-05856-5

14 Pronker, M. F. *et al.* Sialoglycan binding triggers spike opening in a human coronavirus. *Nature* **624**, 201–206 (2023). https://doi.org/10.1038/s41586-023-06599-z

15 Cummings, R. D. & Pierce, J. M. The challenge and promise of glycomics. *Chem Biol* **21**, 1–15 (2014). https://doi.org/10.1016/j.chembiol.2013.12.010

16 Rudd, P. M. *et al.* in *Essentials of Glycobiology* (eds A. Varki *et al.*) 689–704 (Cold Spring Harbor Laboratory Press
Copyright © 2022 The Consortium of Glycobiology Editors, La Jolla, California; published by Cold Spring Harbor Laboratory Press; doi:10.1101/glycobiology.4e.51. All rights reserved., 2022).

17 Ashwood, C., Pratt, B., MacLean, B. X., Gundry, R. L. & Packer, N. H. Standardization of PGC-LC-MS-based glycomics for sample specific glycotyping. *Analyst* **144**, 3601–3612 (2019). https://doi.org/10.1039/c9an00486f

18 Both, P. *et al.* Discrimination of epimeric glycans and glycopeptides using IM-MS and its potential for carbohydrate sequencing. *Nat Chem* **6**, 65–74 (2014). https://doi.org/10.1038/nchem.1817

19 Wang, T. *et al.* CryoSeek: A strategy for bioentity discovery using cryoelectron microscopy. *Proceedings of the National Academy of Sciences of the United States of America* **121**, e2417046121 (2024). https://doi.org/10.1073/pnas.2417046121

20 Wang, T. *et al.* CryoSeek II: Cryo-EM analysis of glycofibrils from freshwater reveals well-structured glycans coating linear tetrapeptide repeats. *Proc Natl Acad Sci U S A* **122**, e2423943122 (2025). https://doi.org/10.1073/pnas.2423943122

21 Wang, T., Sun, Y., Li, Z. & Yan, N. The 8-nm spaghetti: well-structured glycans coating linear tetrapeptide repeats discovered from freshwater with CryoSeek. *bioRxiv*, 2024.2012. 2015.627649 (2024).

22 Li, Z. *et al.* CryoSeek identification of glycofibrils with diverse compositions and structural assemblies. *bioRxiv*, 2025.2009.2030.679562 (2025). https://doi.org/10.1101/2025.09.30.679562

23 Hu, M. *et al.* High-throughput cryo-EM characterization and automated model building of glycofibrils via CryoSeek. *LangTaoSha Preprint Server* (2025). https://doi.org/10.65215/bkvrt910

24 Atanasova, M., Bagdonas, H. & Agirre, J. Structural glycobiology in the age of electron cryo-microscopy. *Curr Opin Struct Biol* **62**, 70 – 78 (2020). https://doi.org/10.1016/j.sbi.2019.12.003

25 Witman, G. B., Carlson, K., Berliner, J. & Rosenbaum, J. L. Chlamydomonas flagella. I. Isolation and electrophoretic analysis of microtubules, matrix, membranes, and mastigonemes. *The Journal of cell biology* **54**, 507–539 (1972). https://doi.org/10.1083/jcb.54.3.507

26 Liu, P. *et al.* Chlamydomonas PKD2 organizes mastigonemes, hair-like glycoprotein polymers on cilia. *The Journal of cell biology* **219** (2020). https://doi.org/10.1083/jcb.202001122

27 Namdeo, S., Khaderi, S. N., den Toonder, J. M. & Onck, P. R. Swimming direction reversal of flagella through ciliary motion of mastigonemes. *Biomicrofluidics* **5**, 34108 – 3410815 (2011). https://doi.org/10.1063/1.3608240

28 Liu, P., Liu, Y. & Zhou, J. Ciliary mechanosensation - roles of polycystins and mastigonemes. *Journal of cell science* **136** (2023). https://doi.org/10.1242/jcs.260565

29 Huang, J. *et al.* Structure-guided discovery of protein and glycan components in native mastigonemes. *Cell* **187**, 1733–1744.e1712 (2024). https://doi.org/10.1016/j.cell.2024.02.037

30 Huang, J. *et al.* High-resolution mastigoneme structure reveals 5',5'-phosphodiesters stabilized glycan folding. *bioRxiv*, 2024.2012.2024.630281 (2024). https://doi.org/10.1101/2024.12.24.630281

31 Chen, L. L., Pousada, M. & Haines, T. H. The flagellar membrane of Ochromonas danica. Lipid composition. *J Biol Chem* **251**, 1835–1842 (1976).

32 Bouck, G. B. The structure, origin, isolation, and composition of the tubular mastigonemes of the Ochromas flagellum. *J Cell Biol* **50**, 362–384 (1971). https://doi.org/10.1083/jcb.50.2.362

33 Markey, D. R. & Bouck, G. B. Mastigoneme attachment in Ochromonas. *J Ultrastruct Res* **59**, 173–177 (1977). https://doi.org/10.1016/s0022-5320(77)80077-4

34 Holwill, M. E. & Peters, P. D. Dynamics of the hispid flagellum of Ochromonas danica. The role of mastigonemes. *The Journal of cell biology* **62**, 322 – 328 (1974). https://doi.org/10.1083/jcb.62.2.322

35 Yamagishi, T., Motomura, T., Nagasato, C. & Kawai, H. NOVEL PROTEINS COMPRISING THE STRAMENOPILE TRIPARTITE MASTIGONEME IN OCHROMONAS DANICA (CHRYSOPHYCEAE)(1).

J Phycol **45**, 1110–1115 (2009). https://doi.org/10.1111/j.1529-8817.2009.00722.x

36 Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290 – 296 (2017). https://doi.org/10.1038/nmeth.4169

37 Chen, S. *et al.* Protein complex structure modeling by cross-modal alignment between cryo-EM maps and protein sequences. *Nature Communications* **15**, 8808 (2024). https://doi.org/10.1038/s41467-024-53116-5

38 Veillon, L. *et al.* Characterization of isomeric glycan structures by LC-MS/MS. *Electrophoresis* **38**, 2100–2114 (2017). https://doi.org/10.1002/elps.201700042

39 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024). https://doi.org/10.1038/s41587-023-01773-0

40 Hansen, E. M., Reeser, P. W. & Sutton, W. Phytophthora beyond agriculture. *Annu Rev Phytopathol* **50**, 359–378 (2012). https://doi.org/10.1146/annurev-phyto-081211-172946

41 Baysal-Gurel, F., Bika, R., Simmons, T. & Avin, F. Identification and Management of Phytophthora Aerial Blight Caused by Phytophthora nicotianae on Catharanthus roseus. *Plant Dis* **106**, 1271–1277 (2022). https://doi.org/10.1094/pdis-06-21-1342-re

42 Midgley, K. A., van den Berg, N. & Swart, V. Unraveling Plant Cell Death during Phytophthora Infection. *Microorganisms* **10** (2022). https://doi.org/10.3390/microorganisms10061139

43 Nakajima, T. & Volcani, B. E. 3,4-dihydroxyproline: a new amino acid in diatom cell walls. *Science* **164**, 1400–1401 (1969). https://doi.org/10.1126/science.164.3886.1400

44 Chen, L. L. & Haines, T. H. The flagellar membrane of Ochromonas danica. Isolation and electrophoretic analysis of the flagellar membrane, axonemes, and mastigonemes. *The Journal of biological chemistry* **251**, 1828–1834 (1976).

45 Valaitis, A. P. *Carbohydrate analysis of mastigonemes from Euglena gracilis and Ochromonas danica*, University of Illinois at Chicago Circle, (1978).

46 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

47 Huang, J. *et al.* High-resolution mastigoneme structure reveals 5', 5'-phosphodiesters stabilized glycan folding. *bioRxiv*, 2024.2012. 2024.630281 (2024).

48 Coutinho, M. F., Prata, M. J. & Alves, S. Mannose-6-phosphate pathway: a review on its role in lysosomal function and dysfunction. *Mol Genet Metab* **105**, 542 – 550 (2012). https://doi.org/10.1016/j.ymgme.2011.12.012

49 Olson, L. J., Zhang, J., Lee, Y. C., Dahms, N. M. & Kim, J. J. Structural basis for recognition of phosphorylated high mannose oligosaccharides by the cation-dependent mannose 6-phosphate receptor. *J Biol Chem* **274**, 29889–29896 (1999). https://doi.org/10.1074/jbc.274.42.29889

50 Varki, A. & Kornfeld, S. Structural studies of phosphorylated high mannose-type oligosaccharides. *J Biol Chem* **255**, 10847–10858 (1980).

51 Moremen, K. W., Tiemeyer, M. & Nairn, A. V. Vertebrate protein glycosylation: diversity, synthesis and function. *Nat Rev Mol Cell Biol* **13**, 448–462 (2012). https://doi.org/10.1038/nrm3383

52 Shpak, E., Leykam, J. F. & Kieliszewski, M. J. Synthetic genes for glycoprotein design and the elucidation of hydroxyproline-O-glycosylation codes. *Proc Natl Acad Sci U S A* **96**, 14736–14741 (1999). https://doi.org/10.1073/pnas.96.26.14736

53 Basu, D. *et al.* Functional identification of a hydroxyproline-o-galactosyltransferase specific for arabinogalactan protein biosynthesis in Arabidopsis. *J Biol Chem* **288**, 10132 – 10143 (2013).

https://doi.org/10.1074/jbc.M112.432609

54 Ogawa-Ohnishi, M. & Matsubayashi, Y. Identification of three potent hydroxyproline O-galactosyltransferases in Arabidopsis. *Plant J* **81**, 736 – 746 (2015). https://doi.org/10.1111/tpj.12764

55 Shin, S., Mugnai, M. L. & Thirumalai, D. Water-Mediated Interactions between Glycans Are Weakly Repulsive and Unexpectedly Long-Ranged. *J Am Chem Soc* **147**, 17448–17458 (2025). https://doi.org/10.1021/jacs.5c04126

56 Zierke, M. *et al.* Stabilization of branched oligosaccharides: Lewis(x) benefits from a nonconventional C-H · · · O hydrogen bond. *J Am Chem Soc* **135**, 13464 – 13472 (2013). https://doi.org/10.1021/ja4054702