Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

High-throughput cryo-EM characterization and automated model building of glycofibrils via CryoSeek

Mingxu Hu[1,6,7], Sheng Chen[2,6], Tongtong Wang[2,6], Lanju Qin[1,6], Qi Zhang[1,6], Yilin Zhang[1], Qijun Ge[1], Tiantian Chen[3], Meng Li[4], Caiwen Li[4], Guorui Xu[5], Qihui Gui[5], Zhangqiang Li[2,7], and Nieng Yan[1,2,7]

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

[2]Beijing Advanced Innovation Center for Structural Biology, State Key Laboratory of Membrane Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

[3]College of Environmental Science and Engineering, Ocean University of China, Qingdao 266100, Shandong Province, China

[4]Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, Shandong Province, China

[5]Yunnan Key Laboratory of Forest Ecosystem Stability and Global Change, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, Yunnan Province, China

[6]These authors contribute equally to this work.

[7]To whom correspondence should be addressed: M. Hu (humingxu@smart.org.cn), Z. Li (lizhangq@tsinghua.edu.cn), or N. Yan (nyan@tsinghua.edu.cn).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Abstract

With CryoSeek, a structure-first paradigm for discovery, we have determined high resolution 3D structures of a number of glycofibrils, in which well-ordered glycans either form a thick shell coating various protein cores or constitute the entire fibril. To improve the throughput of CryoSeek, we hereby report two methods. The recursive bisection clustering (RBC) strategy has been designed to enable high-throughput cryo-EM data processing of fibrils. EModelG is an AI-facilitated algorithm for automated model building of glycans. Using the RBC method, we have established a high-throughput workflow for CryoSeek and have reconstructed 3D EM maps for hundreds of fibrils that can be automatically modelled in EModelG. Based on their molecular compositions and structural features, we tentatively proposed a unified nomenclature scheme for the fibrils discovered via CryoSeek. These structures will lay the foundation for decoding the principles of glycan folding. Furthermore, to adapt to the high volume of cryo-EM structures quickly obtained with the CryoSeek strategy, we have established a namesake database for data archiving and sharing.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Introduction

CryoSeek is a structure-first research paradigm for biological discoveries (1, 2). In the conventional strategy of structural biology, three dimensional (3D) structures were determined for known objects using X-ray crystallography, NMR, or cryo-electron microscopy (cryo-EM). By contrast, cryo-EM is employed just as a microscope to image samples directly collected from natural environments or native tissues in CryoSeek. The resulting high-resolution 3D EM maps serve as the starting point for the identification and characterization of unknown bio-entities.

As proof-of-concept for CryoSeek, we started with water samples collected from multiple sources. After simple processing, such as filtration and concentration, the samples were subjected to cryo-sample preparation and cryo-EM imaging. Among a variety of particles observed in the micrographs, long fibrils easily stood out for their distinctive contours (1, 2). In our initial attempts with the water sample collected from the Tsinghua Lotus Pond, we obtained the 3D EM reconstructions of nine fibrils, with overall resolutions ranging from 3.0-3.5 Å (1-4). We named these fibrils TLP-x, where x stands for the groups on the basis of their molecular composition and structure features.

Among the structurally resolved fibrils, TLP-1a and -1b are pili-like fibrils that may belong to uncharacterized bacterial species. Another five are all coated with dense glycan shells. The highly ordered glycans on these glycofibrils afford unprecedented opportunities to determine complex glycan structures to high

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

resolutions. The mass ratio of sugar residues in these fibrils varies to different degrees. In TLP-IPT and TLP-12, each having a folded protein core, the thick glycoshells occupy at least 60% molecular mass. In TLP-3, whose stem comprises three intertwined linear peptide chains, the mass ratio of glycans is approximately 75%. In TLP-4a/4b and TLP-2, in which the protein core is a linear chain of polypeptide made of simple tetra- or dipeptide repeats, respectively, the folded glycans represent >90% of the mass. The most striking one, TLP-0, is completely protein-free (1-4). These findings have not only advanced our understanding of the folding principles of glycans, but also shed light on the design of biomaterials.

Accompanying the rapid progress in the structural determination of glycans using the CryoSeek strategy, several challenges have emerged. For instance, the absolute hand of glycofibrils lacking folded protein components was unclear in our early analyses. To address this issue, we have developed a cryo-EM data acquisition and processing method named Ahaha which allows for the determination of absolute hand of them (5). Despite these advances, two pressing technical challenges remain for CryoSeek. First, natural water samples typically comprise dozens, if not hundreds, of distinct fibril types, many of which are present at low abundances. It was nearly impossible to solve high resolution structures out of such high heterogeneity and low abundance challenges within a reasonable data processing timeframe using conventional image processing approaches. A high-throughput method is therefore necessitated. On the other hand, compared to proteins, whose structural models can now be conveniently predicted or auto-built in various

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

programs (6-10), structural modeling of sugar residues mainly relies on manual, empirical assignment. Therefore, cryo-EM data processing and model building of non-protein biomolecules have become the bottleneck for CryoSeek.

In the present study, we sought to solve the above technical challenges associated with CryoSeek. To improve the throughput of data processing, the recursive bisection clustering (RBC) strategy was introduced. RBC enables efficient, highly parallel processing and sensitive detection of low-abundance fibrils (down to 0.1‰). Using this method on micrographs derived from water samples at 13 sites, we generated >100 3D EM reconstructions of glycofibrils from ~236,000 micrographs, with a cumulative processing time of approximately 20 days.

Based on EModelX, we have developed EModelG, an AI-driven framework for the automatic model building of glycans. Trained on manually constructed high-order glycan structures, EModelG accurately identifies sugar sites in EM maps and performs density-guided assembly of glycosidic bonds. Our initial trials in EModelG have proved robust performance and broad applicability to both glycoproteins and protein-free glycofibrils.

Combining the two methods, a high-throughput workflow for glycofibril structure determination has been established, facilitating the rapid accumulation of such structural data. To systematically summarize and describe these structures, we have proposed a nomenclature scheme. To accommodate the quickly mounting

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

number of micrographs, 3D maps, and structures obtained via CryoSeek, we have established a namesake database (https://cryoseek.org) to facilitate data archiving and sharing.

Results

The RBC strategy for high-throughput CryoSeek image processing

As fibrils predominantly lie flat in vitreous ice, they exhibit only one view in the collected micrographs. Therefore, the significant heterogeneity of fibrils in water samples should be resolved by clustering picked particles, which are the selected segments along the fibrils. This method is known as 2D classification in cryo-EM. However, owing to the heterogeneity and low abundance of fibrils, in our previous studies, numerous rounds of 2D classification were required, each entailing extensive manual intervention, including input for particle selection and parameter configuration (1, 2). Such procedure was highly time-consuming, unable to keep pace with the relatively high speed of data collection. Therefore, a high-throughput image processing method is in urgent need.

RBC, shorthand for recursive bisection clustering (11), is a top-down hierarchical clustering algorithm, particularly suitable for handling large datasets or high-dimensional data. Distinct from conventional clustering, which adopts a bottom-up approach by merging similar samples, RBC employs a divisive strategy: starting with the entire dataset as a single cluster, it recursively splits one cluster into two child clusters based on a pre-defined criterion at each step. This process continues until a predefined stopping

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

criterion is met, such as when clusters reach a minimum size threshold or when the intra-cluster similarity exceeds a specified threshold.

One key advantage of RBC is that the number of clusters resulted can increase exponentially with recursion depth, while simultaneously maintaining high computational efficiency and being easily parallelizable. This parallelization capability, or scalability, stems from the characteristic that the subsequent computations for the two child clusters are completely independent after they are split; therefore, computational tasks can be highly concurrent as the depth of recursion increases. Accordingly, the RBC strategy, when applied in the 2D classification step, overcomes the inefficiency challenges posed by low abundance and high heterogeneity by virtue of its inherent distinctive characteristics (Fig. 1A). Herein, the capability of the RBC strategy is demonstrated through the processing of cryo-EM data derived from a water sample collected in a karst cave in Guilin. (110.3183°E, 25.2167°N).

Approximately 12.2 million particles (the red input node in Fig. 1B), or fibrils fragments, were picked from 22,272 micrographs. RBC with a maximum recursion depth of 12 was performed on these particles, constructing a top-down binary tree where each node has two child nodes (Fig. 1B). Blue nodes indicate intermediate clusters in RBC, whereas nodes with crosses denote early-stop clusters and green nodes represent final clusters, which are subsequently used to reconstruct atomic or near-atomic resolution density maps. Eighteen fibrils (corresponding to green final nodes) were isolated from 12.2 million particles, with their abundances ranging from 0.14‰ to 4.9‰ (Fig. 1C).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

RBC with a maximum recursion depth of 12, which yield a 12-depth binary tree, contribute to the following characteristics. For brevity, we define the splitting of one particle cluster (one node) into two child nodes as a "task". Firstly, from the perspective of computational scheduling, there are at most 12 tasks that exhibit sequential dependency (Fig. 1B), meaning each task relies on the completion of the preceding one. Therefore, the remaining tasks, which constitute the majority of all tasks, can be executed in parallel across computing resources. This enables processing of the dataset within one or two days when a maximum of 8 L40 GPU cards are allocated to each task. Second, the input particles are clustered into 4096 (2^{12}) clusters, corresponding to a detection sensitivity for abundances of approximately 0.24‰ (labeled on the left in Fig. 1B). Nevertheless, the vast majority of these clusters are not subjected to actual computation via the early-stopping mechanism (early-stop nodes) for the purpose of saving computing resources. As the task does not split the particle cluster evenly, 0.24‰ represents only an order-of-magnitude analysis of the detection sensitivity limit for abundance. In the demonstrated case, a fibril with an abundance of 0.14‰ can be detected (Fig. 1C).

Aside from the karst cave water sample presented in Figure 1, we provide supplementary demonstrations for two additional water samples (Fig. S1) to illustrate the general applicability of the RBC strategy.

Overview and categorization of fibrils structures

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Implementation of the RBC strategy relieved the data processing bottleneck in CryoSeek, thereby allowing the processing throughput to match the pace of data collection. Applying this strategy further facilitated the acquisition of 126 3D helical reconstructions from 13 distinct water samples (Table S1-3). As several reconstructions from different samples exhibited nearly identical structural features, a total of 50 characteristic and nonredundant fibril structures were ultimately identified, most of which were glycofibrils (Fig. 2).

Previously, nine fibril structures were resolved from the Tsinghua Lotus Pond and named TLP-1a/1b, TLP-IPT, TLP-12, TLP-4a/4b, TLP-3, TLP-2 and TLP-0 (1-4). The prefix "TLP" indicated their origin from the Tsinghua Lotus Pond, and the suffixes were assigned according to the structural features of their protein components. The only exception was TLP-1a/1b, in which "1" denoted the first fibril structures to be resolved. As CryoSeek now reveals a broader spectrum of fibril architectures from diverse natural environments, the TLP-based nomenclature is no longer sufficient to encompass all structural variants, necessitating the establishment of a more systematic and generalized naming scheme.

To establish a consistent scheme applicable to diverse fibrils, we adopted a composition- and structure-based system, in which the prefix defines the molecular composition and the suffix specifies structural features. Based on composition, the 50 nonredundant fibrils were divided into 13 protein fibrils and 37 glycofibrils, designated with the prefixes PF (protein fibril) and GF (glycofibril), respectively. Protein fibrils

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

contain minimal or no glycosyl modifications, whereas glycofibrils are predominantly composed of glycans, with glycan-mediated interactions playing an essential role in their assembly.

Rules for suffix assignment are summarized as follows. All currently resolved protein fibrils assemble through mechanisms resembling those of type IV and type V pili, either via N-terminal α-helical coiled-coil interactions or donor-strand exchange (DSE) (12-14). Thus, these 13 protein fibrils are assigned the suffix PLL (pilus-like). For glycofibrils, i) if the protein core is constructed from well-folded protein domains, the suffix will be directly derived from the domain name; ii) if the central stem consists of linear peptide repeats, the suffix follows the format xLyR, where *xL* denotes the number of linear peptide chains in the core and *yR* indicates the length of repeat residues; iii) glycofibrils composed entirely of glycans carry the suffix NP (no protein) to indicate the absence of protein components. Fibrils sharing similar structural features within a subclass are distinguished by sequential numbering. Therefore, this hierarchical nomenclature not only accommodates newly identified fibrils, but also allows systematic renaming of previously reported structures (Fig. 2).

Distinct structural properties of representative glycofibrils

Applying this nomenclature to previously reported TLP fibrils revealed that each of the six glycofibrils belongs to distinct subclasses. TLP-IPT, which contains consecutive IPT (immunoglobulin-like, plexins, transcription factors) domains in its central region (15), it therefore renamed GF-IPT (Fig. 3A).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

TLP-12 features a trimeric assembly of dodecapeptide repeats that intertwine into a β-helix, corresponding to the subclass GF-3BH, where "3" reflects its trimeric architecture and "BH" denotes the β-helical core (Fig. 3B). In this study, we identified three GF-3BH fibrils in total, which share similar stem architectures but differ in their glycan decorations (Fig. 2). Furthermore, we resolved two additional glycofibrils, termed GF-1BH1 and GF-1BH2. As their name indicate, the protein cores of GF-1BHs resemble those of GF-3BHs in folding into a β-helix, but differ in that GF-1BHs are composed of a single chain rather than a trimer.

TLP-2 and TLP-4a/4b, reported in our previous studies, represent glycofibrils with linear peptide repeats in the core. TLP-2 consists of dipeptide repeats, in which one conserved residue carries a glycan chain through glycosylation or phosphoglycosylation (Fig. 3C) (16, 17). TLP-4a/4b consists of tetrapeptide repeats containing a conserved dihydroxylproline (diHyP) (18) and an O-glycosylated Ser/Thr, with three glycan chains attached to each repeat (Fig. 3D). Additional fibrils with similar linear peptide architectures, but differing in glycan branching patterns, were also obtained. These fibrils were therefore renamed GF-1L2Rs and GF-1L4Rs, where "1L" denotes a single linear chain and "2R" or "4R" specify the di- or tetrapeptide repeat length, respectively (Fig. 2).

In the present study, two additional subclasses were identified, whose stems consist of hexapeptide or octapeptide linear repeats, named GF-1L6R and GF-1L8R, respectively (Fig. 2). The hexapeptide repeat contains two conserved O-glycosylated

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Ser/Thr residues at position 1 and 3 (Fig. 3E). Although the octapeptide repeat resembles the tetrapeptide repeat identified in GF-1L4Rs, the glycans attached to the first diHyp and Thr/Ser differ from those linked to the second pair, thereby defining the octapeptide periodicity (Fig. 3F).

Glycofibrils resembling TLP-3 contain trimeric assemblies of linear tripeptide repeats. These fibrils were renamed GF-3L3Rs, where "3L" represents three intertwined chains and "3Rs" indicates the tripeptide repeats (Fig. 3G). This convention can be extended to other glycofibrils with multiple linear peptide chains in their protein cores.

In addition to peptide-containing fibrils, we also identified several glycofibrils composed entirely of glycans, as revealed by high-resolution reconstructions and manual inspection (Fig. 3H). These fibrils, which lack any detectable protein components, were designated GF-NPs, with "NP" denoting no protein components (Fig. 2).

In summary, according to this nomenclature system, the 37 glycofibrils can be further classified into nine subclasses, and additional subclasses may be defined as more distinct glycofibrils are resolved (Fig. 2).

Automated cryo-EM glycan modeling with EModelG

AI-facilitated structural prediction and automated model-building tools have greatly advanced structural biology (19, 20), but these tools fail to predict or automatically build glycan models. Since we recently obtained several higher-order glycan structures, we

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

leveraged them, together with the limited glycan and ribonucleic acid structures available in the PDB, as training data to develop EModelG, an AI-driven glycan structure auto-building framework that enables fully automated modeling of glycoprotein and glycan structures directly from cryo-EM density maps.

Using GF-IPT as an example, which features well-folded protein domains and extensive glycan decorations, we demonstrate the framework's performance. Starting from the raw density map of GF-IPT, the framework performs neural-network–based density interpretation to assign voxel-wise probabilities to protein and carbohydrate regions. Predicted protein regions are subsequently modeled by EModelX, producing an initial atomic model that provides putative glycosylation sites for downstream carbohydrate modeling. Guided by these anchors, EModelG iteratively conducts SE(3) (Special Euclidean group in 3 dimensions) rotational sampling of monosaccharide templates, gradient-based density fitting, and density-guided glycosidic bond construction to extend glycan chains from the protein surface. The resulting model integrates both protein and glycan components into a coherent atomic representation aligned with the experimental map (Fig. 4A).

Another representative example is the glycofibril GF-1L2R5 (2.53 Å), whose stem is linear dipeptide repeats, with surrounding glycans occupying most of the molecular mass. Using EModelG, we accurately reconstructed the entire glycan branches (Fig. 4B). Automatically built monosaccharide rings (pink) closely match the experimental density (gray) and exhibit strong agreement with the independently built manual model (white). A

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

schematic comparison (Fig. 4C) further shows that nearly all monosaccharide residues and glycosidic linkages are recovered, with only a limited number of false positives arising in peripheral low-density regions.

Together, these observations establish EModelG as a general, interpretable, and fully automated framework for AI-driven, density-guided glycoprotein modeling. In contrast to traditional manual fitting workflows in Coot (21), EModelG operates without human intervention. To our knowledge, this represents the first automated method for glycan modeling directly from cryo-EM density.

To enable fully automated glycan modeling in the CryoSeek workflow, accurate structural autobuilding and glycosylation sites identification of the underlying protein scaffold are essential. We therefore applied the EModelG protein autobuilding module directly to the cryo-EM density maps without providing any prior sequence information. As an initial validation on a protein-only sample, the method was used to model the GF-PLL1 map at 2.87 Å resolution (Fig. S2A–C). Under these conditions, EModelG automatically generated continuous backbone and well-resolved side-chain models for individual protomers, with the auto-built traces closely following the experimental density.

For GF-PLL1, the de novo model was then subjected to sequence identification. The predicted sequence yielded a confident match to a UniProtKB entry, Q6FFR1, a major fimbrial subunit protein from Acinetobacter baylyi (BLAST E-value = 1.2×10^{-48}; Fig. S2C). The AlphaFold-predicted structure of Q6FFR1 showed strong agreement with the

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

EModelG auto-built model (TM-score = 0.87), providing orthogonal support for the correctness of the assigned sequence. Using this identified sequence, the final GF-PLL1 model was rebuilt and refined, resulting in excellent overall fit to the EM density.

Having established the accuracy of structural autobuilding and sequence identification on a pure-protein system, we next examined a glycoprotein target, GF-IPT1, which features a more complex protein architecture and multiple glycosylation sites (Fig. S2D–F). The EModelG module again produced a high-quality protein model for GF-IPT1 at 2.87 Å, and comparison with the manually built reference structure revealed close agreement in both the global fold and local side-chain conformations (Fig. S2E). In particular, the side chains at glycosylation sites were correctly placed and oriented toward the surrounding density, furnishing reliable Asn/Ser/Thr anchor points for subsequent glycan building. Using these automatically identified sites, EModelG initiated glycan chain growth and generated glycan models that recapitulate the connectivity and branching patterns of the manually built glycans (Fig. S2F). In the CryoSeek pipeline, this protein autobuilding and sequence identification step supplies the precise glycosylation anchors and structural context required for EModelG to perform fully automated glycan modeling.

To illustrate the performance of EModelG in protein-free glycan-modeling scenarios, we examined a representative glycofibril dataset (Fig. S3), for which both a manually-built glycan model and an EModelG auto-built glycan model were available. Across the entire glycan chain, the two models exhibit high structural concordance, with an average all-atom RMSD of 0.643 Å, indicating that EModelG accurately reproduces the stereochemical and conformational details determined by expert manual modeling.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

We next compared per-residue map–model correlation coefficients (CC) between the two glycan models, providing a direct measure of local density fitness. The auto-built model attains CC values that are nearly identical to those of the refined manual model, and both follow similar residue-wise trends along the chain, demonstrating that EModelG captures density-supported features with comparable precision. The global CC values are likewise similar, indicating that automated glycan modeling does not compromise overall map–model consistency.

Zoom-in views of representative monosaccharides further illustrate the quality and limitations of the automated modeling (Fig. S3D,E). In one region (Fig. S3D), the auto-built and manual models adopt different monosaccharide types but still superpose closely, yielding low local all-atom RMSD and very similar CC values; this highlights that, in regions of limited chemical distinguishability, alternative but geometrically plausible sugar assignments can fit the density equally well. In a second region (Fig. S3E), the auto-built and manual models use exactly the same monosaccharide type and display near-perfect superposition, showing that EModelG can recover both the correct sugar identity and its ring conformation and glycosidic geometry when the density is sufficiently informative. Together, these analyses demonstrate that, for well-resolved maps, automated EModelG glycan modeling achieves expert-level accuracy at both global and local structural scales.

The CryoSeek Database

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

As the RBC strategy and EModelG have overcome longstanding bottlenecks in the CryoSeek strategy for cryo-EM data processing and model building of non-protein biomolecules, we successfully determined hundreds of high-resolution glycan structures. To facilitate systematic investigation of the structural roles and folding principles of glycans, and to make these data easily accessible to the broader scientific community, we developed the CryoSeek database.

The CryoSeek database, designed and developed through the collaborative efforts of SMART (Shenzhen Medical Academy of Research and Translation), is a comprehensive repository that contains experimentally derived cryo-EM raw data, 3D density maps, and atomic models of glycans, nucleic acids, proteins, and other bioentities. As articulated by the CryoSeek strategy, the database embodies a structure-first paradigm for biological discoveries. Scientists worldwide can contribute cryo-EM-related datasets to the database and utilize its resources for scientific researches. Additionally, the CryoSeek database serves a broad community of researchers, educators and students across diverse scientific disciplines, fostering collaboration and accelerating structural biology innovation.

Users can sequentially upload their cryo-EM metadata files to the CryoSeek database through *Deposit* function on the website. The database then runs a suite of programs to identify potential links within the submitted dataset and perform further validation. All submitted entries are manually reviewed and evaluated by the administrator team, who complete the annotations and return the entries to users for

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

comment and approval. Once approved by the authors, the finalized entries are publicly released through the CryoSeek database (Fig. 5A).

Users can also search for information or structures of interest via the *Search* function on the website. In addition, the CryoSeek database provides graphical access to and interactive visualization of the 3D density maps and atomic models for the scientific community through web pages. These pages enable users to *Download* the corresponding files for further analysis. Moreover, the CryoSeek database integrates several additional functions, such as Ahaha (5), a tool that helps users determine the absolute hand of their fibril structures (Fig. 5B,C).

Discussion

In this study, we proposed two methods, the RCB strategy and EModelG, to respectively enable high-throughput cryo-EM data processing and automatic glycan model building for CryoSeek. Notably, EModelG is the first automatic modeling methods for glycans to date. The establishment of this high-throughput workflow allowed us to obtain hundreds of 3D helical reconstructions, while the automatic modeling tools yielded numerous high-order glycan structures. To better characterize these structures, we developed a systematic and generalized nomenclature system that enables distinction of their molecular compositions and structural features.

These structures provide an unprecedented foundation for investigating the structural roles and folding principles of glycans, offering new insights into their

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

functional relevance in biological systems. In addition, we developed the CryoSeek database to facilitate sharing of related 3D density maps, atomic models, and associated raw data, which may further advance structure-guided discoveries.

Among the 126 3D helical reconstructions obtained to date, 100 correspond to glycofibrils. The proportion of resolved glycofibrils is significantly higher than that of protein fibrils, which may be attributed to several factors. One plausible explanation is that glycofibrils exhibit higher stability than protein fibrils, as their thick glycan shell can protect the fragile protein core from enzymatic degradation, or other environmental factors. Additionally, glycans are highly hydrophilic, and leveraging these unique properties of glycan structures could facilitate the design of novel biomaterials with both high polarity and enhanced stability.

The origins and potential functions of these fibrils remain to be elucidated. Thus, complementary techniques such as mass spectrometry and metagenomic sequencing of sampling sites are required for their comprehensive characterization. These analytical data and corresponding results will be deposited in the CryoSeek database in the future, and systematic comparison of these experimental results with related cryo-EM results will facilitate understanding of the diversity and specificity of species and bioentities across different sampling sites.

Unlike the protein data bank (PDB) and the electron microscopy data bank (EMDB) (22, 23), the CryoSeek database is dedicated to non-protein biomolecules,

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

providing search, visualization functions tailored to researchers studying glycans or other bioentities. The current version of the database focuses on data deposition and sharing, with future updates planned to integrate more interactive analytical tools to enhance its utility.

In sum, our study further advances the CryoSeek strategy as a high-throughput workflow for the structural characterization of novel bioentities. Meanwhile, these cryo-EM related data of bioentities have been deposited to CryoSeek database, which supports data sharing and helps to explore the fundamental principles of uncharacterized biomolecules such as glycans. Given that the CryoSeek strategy can be applied to broader scenarios, these resources are expected to accelerated discoveries in the research of glycoscience, and related structural field.

Materials and methods

Collection and pretreatment of water samples

Supplementary Table 1 summarizes the locations where the water samples were collected. With the exception of the water sample from the Tsinghua lotus pond, whose pretreatment procedure has been described in our previous study (1, 2), all other samples (5 to 10 liters in volume) were subjected to filtration using a 0.22 µm pore size filter (JinTeng). The resulting filtrate was concentrated to a final volume of 200 mL via a tangential flow filtration system (Merck). This tangential flow filtration process employed a 0.11 m^2 membrane (Biomax) with a molecular weight cutoff (MWCO) of 50

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

kDa, operating under a constant transmembrane pressure of 0.7 bar at ambient temperature. Subsequently, the concentrated solution was further reduced to a final volume of approximately 50 μL using Centricon ultrafiltration devices (Millipore) with a 100 kDa MWCO.

Cryo-sample preparation and data acquisition

With the exception of the water sample from the Tsinghua Lotus Pond, whose cryo-sample preparation and data acquisition protocols have been described in our previous study (1, 2), cryo-samples were prepared using a Vitrobot Mark IV (Thermo Fisher Scientific). Aliquots of 4 μL of the concentrated aqueous solution were deposited onto glow-discharged holey carbon grids (gold, 300 mesh, R1.2/1.3, Quantifoil). After a blotting duration of 6 seconds, the grids were plunge-frozen in liquid ethane precooled with liquid nitrogen. Cryo-EM data acquisition was performed using two Krios G4 300 kV transmission electron microscopes, each equipped with a Falcon4i direct electron detector and a Selectris X energy filter (slit width: 10 eV). Movies were acquired in electron event representation (EER) mode using EPU software, with a total electron dose of 50 $e^{-}/Å^{2}$ and a defocus range of 1.2–1.6 μm.

Cryo-EM image processing

Following motion correction and CTF estimation performed using cryoSPARC, particles were identified through filament tracing. After their extraction, RBC strategy for 2D classification was performed. For each node (Fig. 1), the bisection of particles into two smaller clusters was based on the affinity matrix, was computed using cryoSPARC. This

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

affinity matrix was calculated among the class averages obtained from the cryoSPARC 2D classification module. The number of classes in the 2D classification module was set to max{30, min{100, N/100}}, where N denotes the number of particles to be split. To conserve computational resources, early-stops are conducted for demonstrably non-informative clusters (early-stop nodes in Fig. 1), thereby preventing further unnecessary bisections. Bisection clustering is concluded when homogeneous particle populations are obtained (final nodes in Fig. 1). Subsequent to obtaining suitable particles following implementation of the RBC strategy, helical symmetry was imposed during 3D reconstruction. Helical parameters were determined either via helical indexing in Fourier space using AI-HEAD.

Model building and refinement

For protein fibrils, atomic models were first automatically built using EModelX, while for glycan fibrils, atomic models were initially automatically built using EModelG. Both sets of atomic models were subsequently subjected to manual adjustment in Coot (21) and finally refined with Phenix (24). The automatic building of glycans using EModelG is described in detail as follows.

Glycan structure auto-building is implemented as a two-stage, density-guided pipeline that decouples monosaccharide pose detection from subsequent glycan chain growth. The overall strategy is to first identify high-confidence monosaccharide candidates directly from the cryo-EM map, independent of sugar identity, and then refine both the connectivity and monosaccharide type under joint density and stereochemical constraints.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Given an input cryo-EM density map $\rho(x)$, we first predict carbohydrate-specific density using a trained 3D UNet(25) f_θ, yielding a voxel-wise glycan probability map $p_{\mathrm{gly}}(x) = f_\theta(\rho)(x)$. Thresholding and connected-component analysis produce candidate glycan regions, within which we exhaustively sample poses of a D-Xylopyranose template (Supplementary Algorithm 1). Rotations are drawn from a Fibonacci-sphere–based sampling of SO(3), combined with local translations within 1 A° around each candidate center to account for discretization and minor registration errors. For each pose, we evaluate the density at all template atom positions via trilinear interpolation, or say, grid sample, and define a robust pose score as the mean density over ring atoms minus their standard deviation. This choice penalizes heterogeneous or noisy fits while favoring compact, consistently supported rings. A non-maximum suppression scheme (Supplementary Algorithm 2) is then applied within each density region to retain only non-overlapping, high-scoring poses, producing a sparse assembly of D-Xylopyranose units without explicit glycosidic bonds.

Next, we integrate the monosaccharide assembly with the automatically built protein model to infer candidate glycosylation sites and partition the glycan density into site-specific regions. Starting from each glycosylation site, we iteratively grow glycan chains by selecting the nearest unused monosaccharide candidates that are density- connected to the current sugar (Supplementary Algorithm 3). For each potential linkage, we exhaustively test a library of 22 monosaccharide templates. Each template is superimposed onto the local geometry such that the putative glycosidic bond length lies near a target value (ℓ_0) within a tolerance (δ). The resulting pose is scored using the same density-based metric, optionally combined with a penalty on bond-length deviation, and the best-scoring

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

monosaccharide type and pose are selected (Supplementary Algorithm 4). Accepted poses are added to the growing frontier and linked via glycosidic bonds; the process iterates until no further density-supported linkages can be established.

This design allows the pipeline to exploit the high information content of the cryo-EM map while maintaining explicit stereochemical control. Using a single generic pyranose template for initial pose detection avoids an early combinatorial explosion in sugar types, whereas the subsequent type-refinement step leverages both density and bond-length constraints to assign chemically plausible, heterogeneous glycoforms in a data-driven manner.

EModelG training and evaluation

For glycan-focused density interpretation, we reused the 3D residual U-Net architecture previously developed for our protein auto-building framework (20), but retrained it on a dataset specifically enriched for carbohydrate- and sugar-containing nucleic acid densities. All training maps were single-particle cryo-EM reconstructions with fitted PDB models (26), deposited before 2025-01-01 and with reported resolution better than 4 Å. After removing redundancy at the assembly level, the final corpus comprised 2319 protein/oligosaccharide assemblies and 3756 protein/nucleic acid assemblies. Including nucleic acid complexes was important because RNA and DNA provide abundant examples of furanose rings (ribose and deoxyribose), which act as a natural data augmentation source for learning five-membered sugar-ring densities that are otherwise under-represented in glycoprotein entries. For each map, we resampled the density to 1 Å voxel spacing, aligned

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

it to the fitted PDB frame, normalized voxel intensities to a fixed range, and computed Q-scores(27) for all atoms to quantify local map resolvability during label generation.

Supervision was designed as a coarse-to-fine hierarchy over density voxels. In the first stage, the network performs a coarse atomic-type classification on individual voxels with respect to nearby PDB atoms. Each voxel is labeled as (i) non-structural, (ii) generic atomic voxel, (iii) neighborhood voxel within a 3×3×3 window around ring atoms belonging to 4–7-membered rings, or (iv) ring-atom voxel for 4–7-membered rings. The corresponding output score for ring-atom voxels defines a scalar index s1 that is used both in the training loss and later as a proposal score for ring-atom candidates. In the second stage, the network is trained to recognize sugar-ring centroids in the vicinity of 4–7-membered ring atoms, yielding a centroid score s2. During inference, the product s1·s2 is used as a detection score for candidate sugar-ring centroids, and a 3D non-maximum suppression procedure is applied to obtain a sparse set of high-confidence glycan seeds. Both stages are trained jointly with class-balanced cross-entropy losses, with higher weights assigned to ring-atom and centroid classes to counteract strong class imbalance. Network parameters are optimized with the Adam optimizer (28) using mini-batches of randomly sampled 3D patches; a held-out validation set is used to monitor overfitting and select the final checkpoint.

We evaluated our method at both the neural network detection level and the final glycan model level using the following metrics:

1. Precision (sugar-ring centroid detection)

   For a given detection threshold, precision is defined as the fraction of predicted

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

sugar-ring centroids that fall within 4 Å of any true sugar-ring centroid in the reference model.

2. Recall (sugar-ring centroid detection)

   For the same threshold, recall is defined as the fraction of true sugar-ring centroids that have at least one predicted centroid within 4 Å. This measures how many true glycan seeds are successfully recovered.

3. Area under the precision–recall curve (AUPRC)

   By varying the detection threshold, we obtain a precision–recall curve. The area under this curve (AUPRC) summarizes the overall detection performance, which is especially informative for this highly imbalanced task where sugar-ring centroids are sparse.

4. All-atom RMSD for glycan structures

   To assess the accuracy of the auto-built glycan models, we compute an all-atom RMSD between the automatic model and a manual reference. First, monosaccharides are paired by bipartite matching to obtain one-to-one correspondences. For each matched pair, the RMSD of all heavy-atom coordinates is calculated, and the final all-atom RMSD is reported as the average over all matched monosaccharides.

5. Phenix CC per residue

   As a complementary map-based metric, we use the real-space correlation coefficient per residue calculated by Phenix(29). For each protein or glycan residue, Phenix measures the local correlation between the experimental density map and

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

the model-derived map, providing a residue-level indicator of how well the model is supported by the density.

Data availability

Micrographs, density maps, and atomic models are accessible via the CryoSeek database (https://cryoseek.org).

Author contributions: N.Y., M.H., and Z.L. conceived the project. M.H. developed the RBC strategy, while S.C. and M.H. developed EModelG. L.Q. collected the water samples, and L.Q. and Y.Z. prepared the cryo-samples and conducted cryo-EM data acquisition. M.H. performed cryo-EM image processing; Z.L., T.W., S.C., and Q.G. built atomic models; Z.L. and T.W. analyzed the atomic models; and Z.L., M.H., T.W., and N.Y. proposed the naming scheme of fibrils. M.H. and Q.Z. constructed CryoSeek database. N.Y., M.H., and Z.L. wrote the manuscript with the input of all other authors.

Acknowledgements

We gratefully acknowledge Dr. Gang Fu, Xiaosong Pei, Dr. Xiaomin Li, Dr. Fan Yang, and Dr. Jianlin Lei for their technical support during cryo-EM image acquisition. We also thank the Structural Biology Core Facility at the Bio-Tech Center of SMART for technical assistance. Additionally, we acknowledge the Computing Labware for Electron-microscopy Visualization and Experimental Research (CLEVER) at SMART for support with data analysis. We acknowledge Dr. Li Xu for her guidance in model builing. We are grateful to the Tsinghua University Branch of China National Center for Protein Sciences

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

(Beijing) for providing cryo-EM facility access and support. This work was supported by the National Natural Science Foundation of China (92478205 to N.Y.) and SMART (to M.H.).

Conflict of interest

The authors declare no competing interests.

References:

1. T. Wang *et al.*, CryoSeek: A strategy for bioentity discovery using cryoelectron microscopy. *Proc Natl Acad Sci U S A* **121**, e2417046121 (2024).
2. T. Wang *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).
3. T. Wang, Y. Sun, Z. Li, N. Yan, The 8-nm spaghetti: well-structured glycans coating linear tetrapeptide repeats discovered from freshwater with CryoSeek. *bioRxiv*, 2024.2012. 2015.627649 (2024).
4. Z. Li *et al.*, CryoSeek identification of glycofibrils with diverse compositions and structural assemblies. *bioRxiv*, 2025.2009. 2030.679562 (2025).
5. Q. Zhang *et al.*, Absolute hand determination of glycofibrils from natural sources in cryo-EM. *BioRxiv* (2025).
6. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
7. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
8. C. Rohl, C. Strauss, K. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93 (2004).
9. K. Xu, Z. Wang, J. Shi, H. Li, Q. Zhang (2019) A2-net: Molecular structure estimation from cryo-em density volumes. in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 1230-1237.
10. K. Jamali *et al.*, Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450-457 (2024).
11. H. Simon, S. Teng, How good is recursive bisection? *Siam J Sci Comput* **18**, 1436-1445 (1997).
12. M. Lukaszczyk, B. Pradhan, H. Remaut, The biosynthesis and structures of bacterial pili. *Bacterial cell walls and membranes*, 369-413 (2019).
13. L. Craig, M. Pique, J. Tainer, Type IV pilus structure and bacterial pathogenicity. *Nat Rev Microbiol* **2**, 363-378 (2004).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

14. H. Remaut *et al*., Donor-strand exchange in chaperone-assisted pilus assembly proceeds through a concerted beta strand displacement mechanism. *Mol Cell* **22**, 831-842 (2006).
15. P. Bork, T. Doerks, T. Springer, B. Snel, Domains in plexins: links to integrins and transcription factors. *Trends Biochem Sci* **24**, 261-263 (1999).
16. P. Haynes, Phosphoglycosylation: a new structural class of glycosylation? *Glycobiology* **8**, 1-5 (1998).
17. P. Van den Steen, P. Rudd, R. Dwek, G. Opdenakker, Concepts and principles of O-linked glycosylation. *Crit Rev Biochem Mol Biol* **33**, 151-208 (1998).
18. T. Nakajima, B. Volcani, 3,4-dihydroxyproline: a new amino acid in diatom cell walls. *Science* **164**, 1400-1401 (1969).
19. K. Jamali *et al*., Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450-457 (2024).
20. S. Chen *et al*., Protein complex structure modeling by cross-modal alignment between cryo-EM maps and protein sequences. *Nature Communications* **15**, 8808 (2024).
21. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Biological crystallography* **66**, 486-501 (2010).
22. H. Berman *et al*., The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
23. C. Lawson *et al*., EMDataBank unified data resource for 3DEM. *Nucleic Acids Res* **44**, D396-403 (2016).
24. P. D. Adams *et al*., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Biological crystallography* **66**, 213-221 (2010).
25. O. Ronneberger, P. Fischer, T. Brox (2015) U-net: Convolutional networks for biomedical image segmentation. in *International Conference on Medical image computing and computer-assisted intervention* (Springer), pp 234-241.
26. H. M. Berman *et al*., The protein data bank. *Nucleic acids research* **28**, 235-242 (2000).
27. G. Pintilie *et al*., Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nature methods* **17**, 328-334 (2020).
28. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
29. T. C. Terwilliger, O. V. Sobolev, P. V. Afonine, P. D. Adams, Automated map sharpening by maximization of detail and connectivity. *Biological Crystallography* **74**, 545-559 (2018).
30. E. Pettersen *et al*., UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Figures and Figure legends

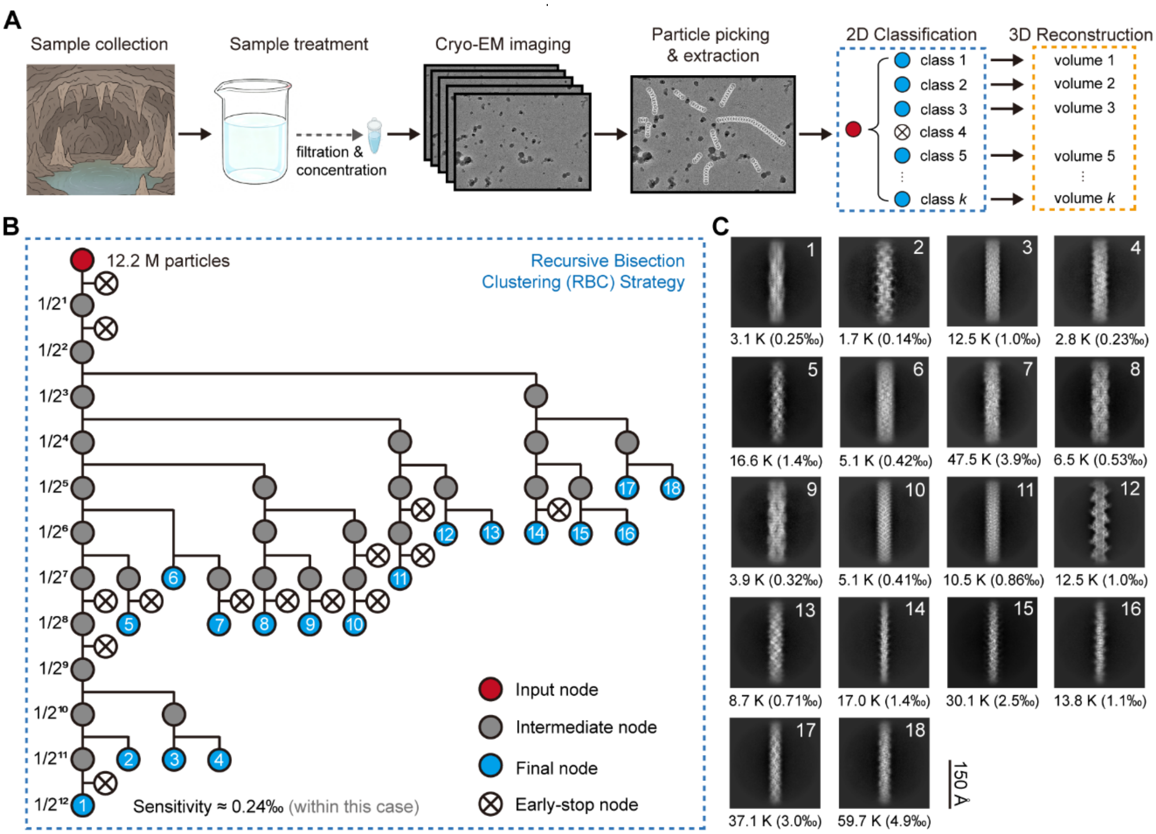
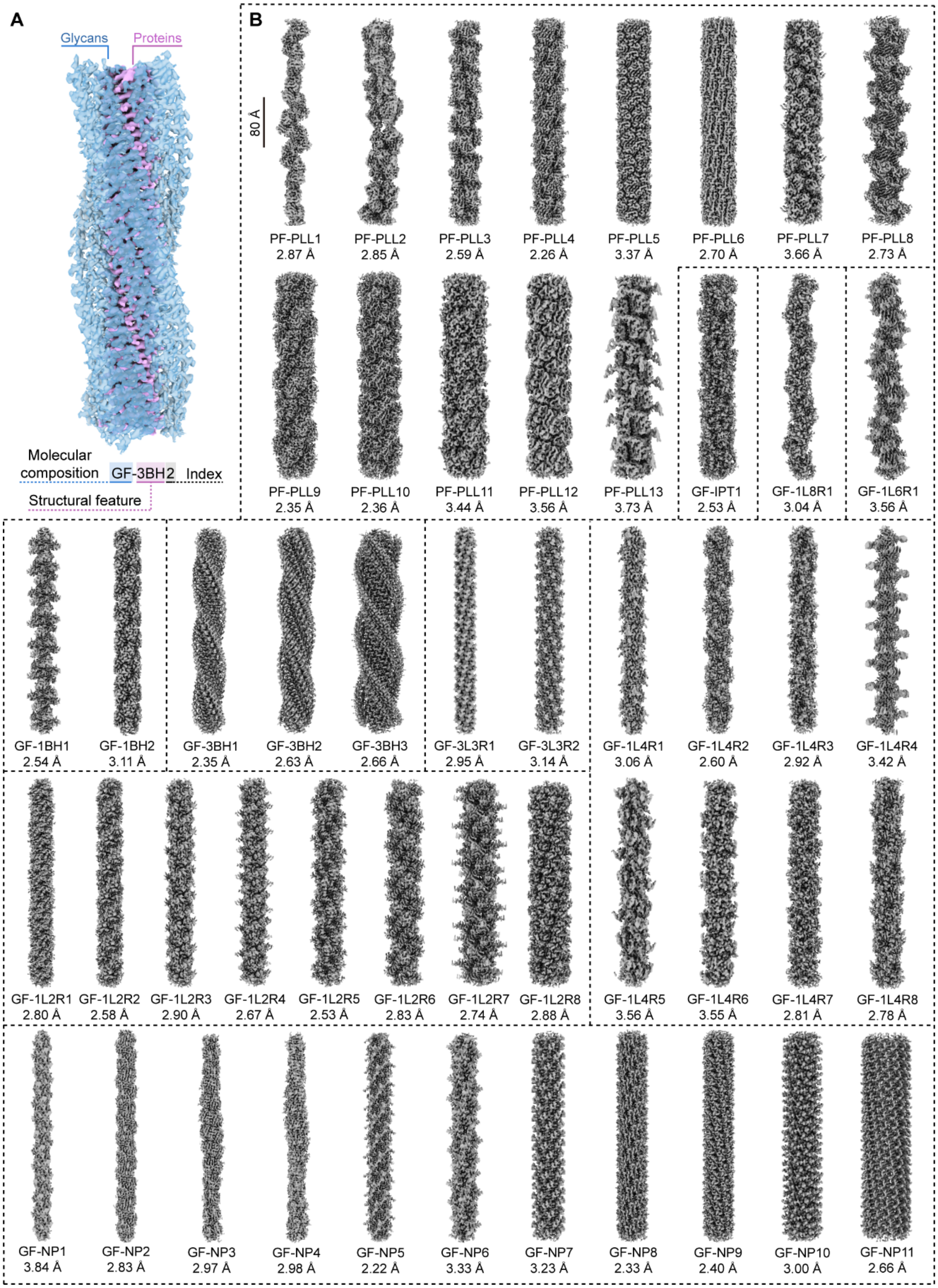


Fig. 1 | The recursive bisection clustering (RBC) image processing strategy for high throughput CryoSeek structure determination of heterogeneous fibrils. (**A**) Schematic diagram of the CryoSeek workflow improved with the RBC strategy for fibril structure determination. (**B**) Schematic diagram of the RBC image processing strategy. The binary tree was generated via a 12-layer RBC strategy for the 2D classification of picked particles along the heterogeneous fibrils. Detection sensitivity increases exponentially with the depth of bisection, as indicated on the left. The early-stop mechanism was implemented to save computational resources. (**C**) Representative 2D class averages resulted from the RBC data processing. Final nodes derived from the RBC strategy comprise particles that enable the reconstruction of high-resolution 3D EM

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

maps. The number of the particles and their proportion relative to the total input particles are annotated at the bottom.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.



Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

Fig. 2 | Nomenclature scheme for 50 distinctive nonredundant fibrils. (**A**) Rules of the nomenclature scheme. For each fibril structure, the name consists of three parts, representing its molecular composition, structural features, and index, respectively. (**B**) The 3D maps and corresponding names of the 50 characteristic nonredundant fibril structures are shown. These fibrils can be classified into 10 subclasses. PF, GF, and NP stand for protein fibril, and glycofibril, and no protein, respectively. The first number following GF refers to the oligomerization number of the protein strands. BH or L following the first number stand for β-helical core and linear peptide chain, respectively. In L subclass, the number between L and R refers to the residue number in each repeat, which represents the building block for the linear peptide chain. The last number indicates is a randomly assigned one to distinguish the glycofibril with each category. PLL: pili like. IPT: immunoglobulin-like, plexins, transcription factors. The black scale bar represents 8 nm. The resolutions are listed below each 3D EM reconstruction. All EM maps were contoured at 6 σ and prepared in ChimeraX (30).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

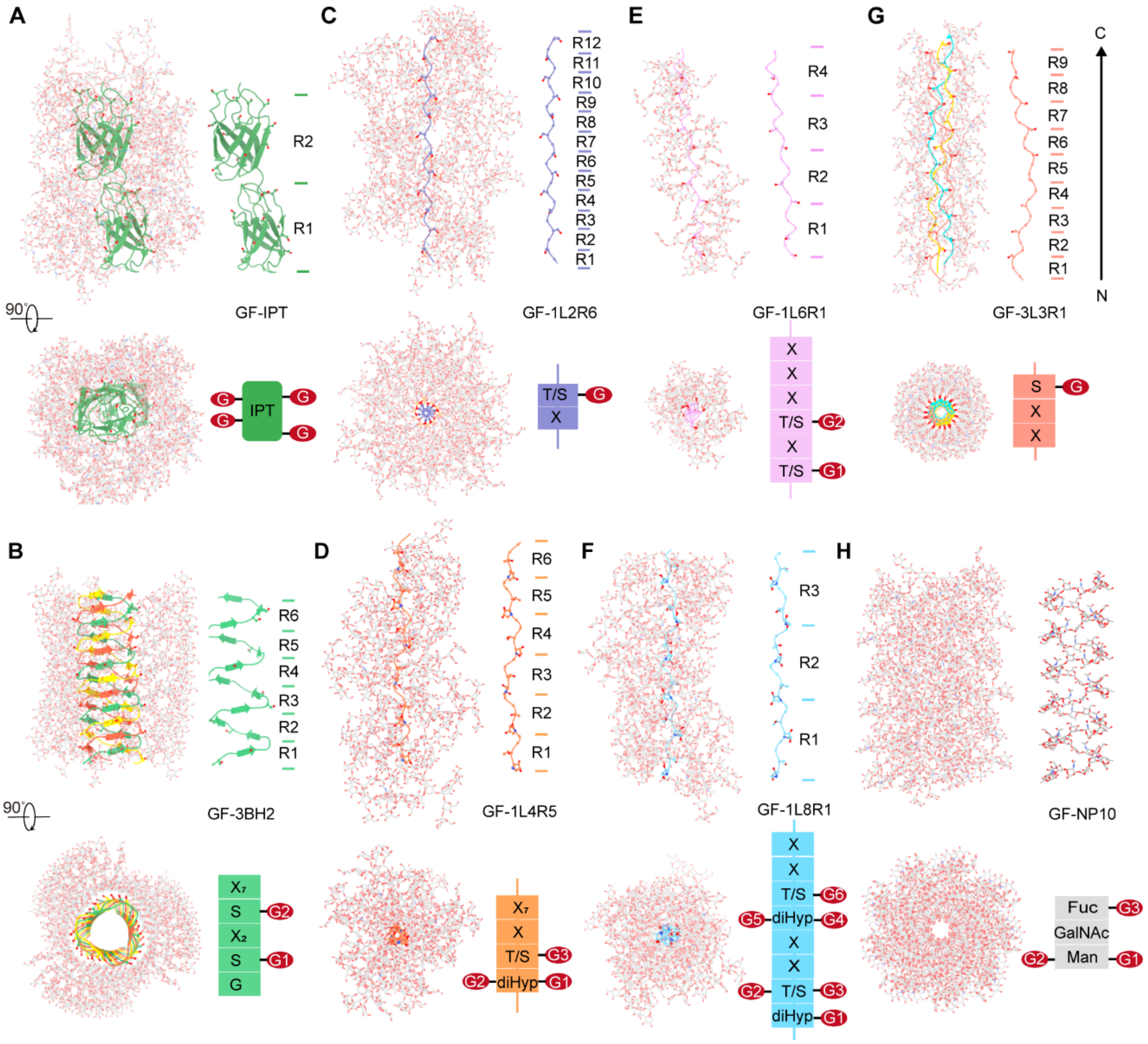


Fig. 3 | Structures of eight representative glycofibrils. Two perpendicular views of the structural models and the diagram of the protein cores are shown for GF-IPT (**A**), GF-3BH2 (**B**), GF-1L2R6 (**C**), GF-1L4R5 (**D**), GF-1L6R1 (**E**), GF-1L8R1 (**F**), GF-3L3R1 (**G**), GF-NP10 (**H**). The carbon and oxygen atoms of the coating glycans in each glycofibril are colored grey and red, respectively, and the protein cores are presented in different colors. The schematic illustration represents the asymmetric unit of each glycofibril. All structural figures were prepared in ChimeraX (30).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

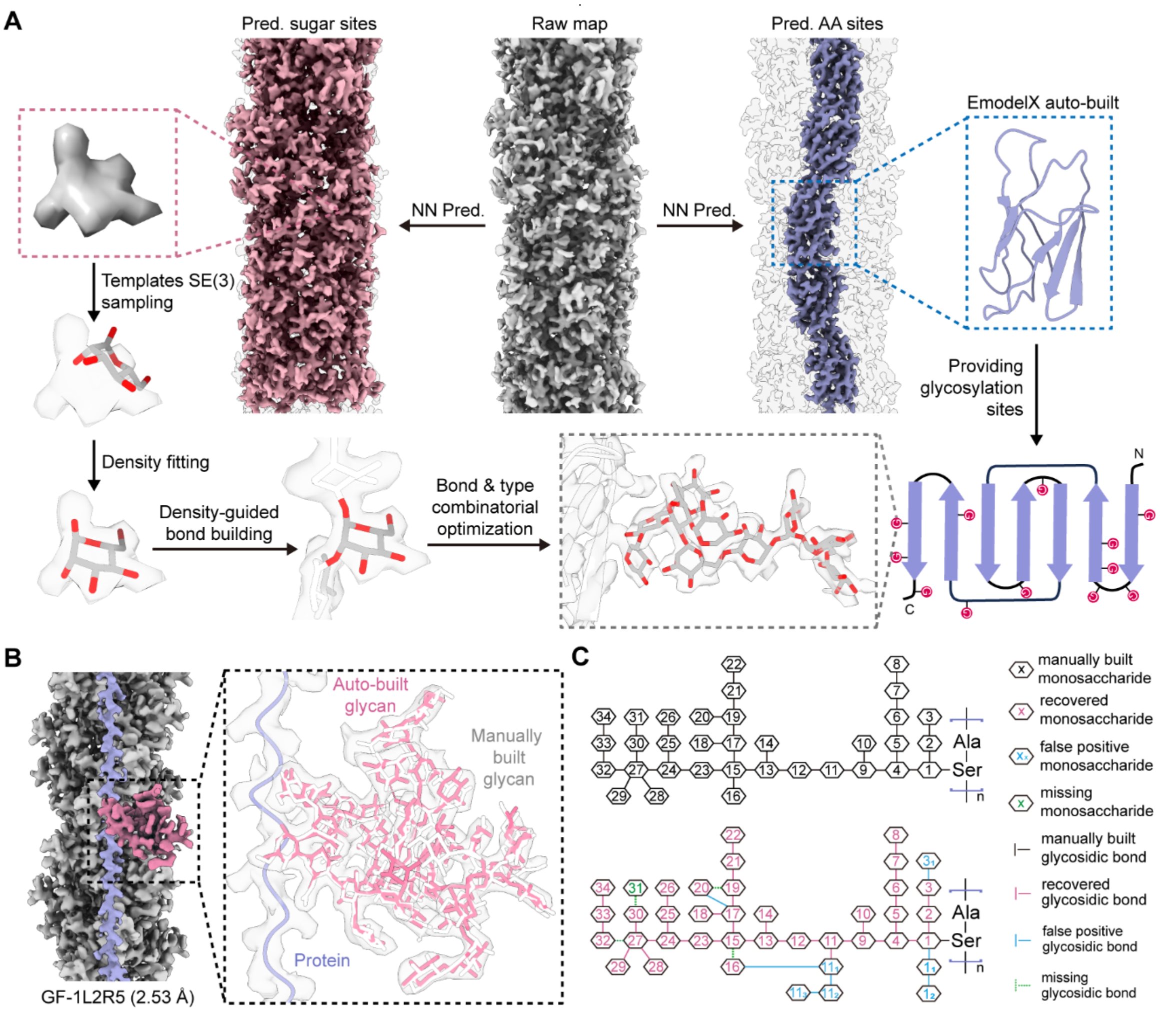


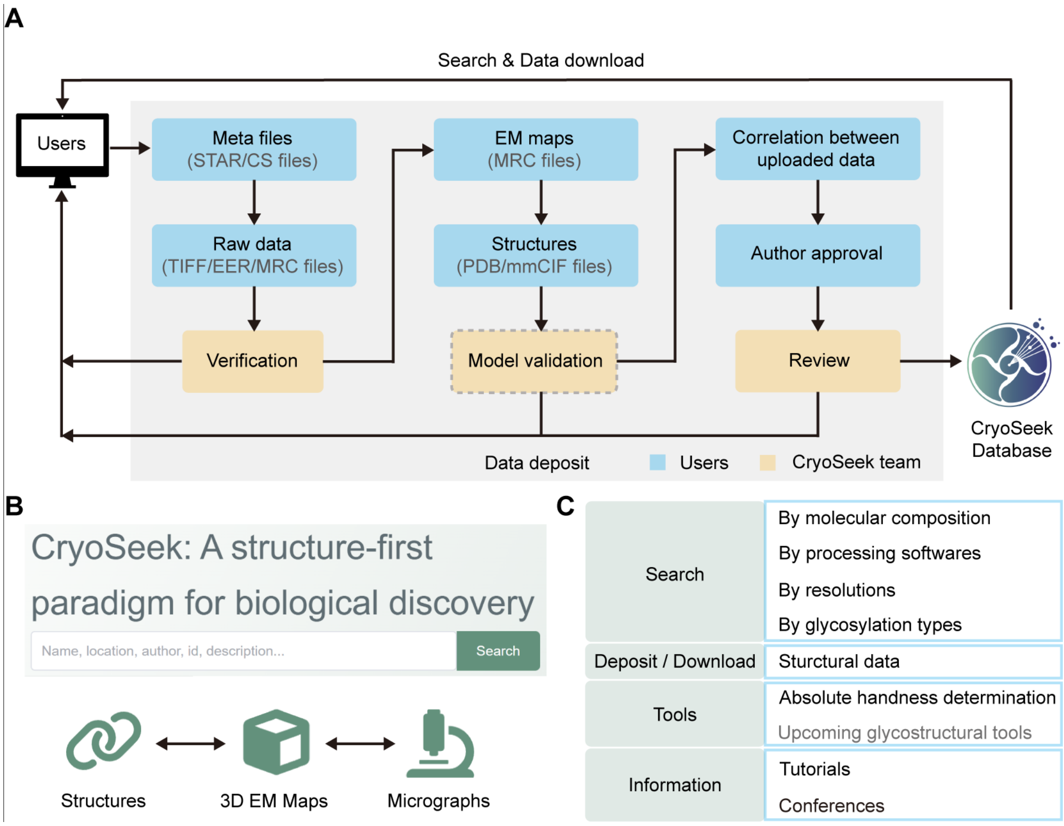
Fig. 4 | Overview of the EModelX-Glycan (EModelG) pipeline for automated Cryo-EM glycoprotein modeling. (**A**) The workflow begins with a raw cryo-EM density map, from which a neural network predicts regions corresponding to carbohydrate (left) and amino acid (right) densities. Predicted amino acid regions are used to automatically build a protein backbone model using EModelX, revealing potential glycosylation sites as anchors for subsequent glycan modeling. For each voxel predicted as carbohydrate, monosaccharide templates are sampled across SE (3) rotations and optimized through gradient-based density fitting. Density-guided glycosidic bond building and combinatorial type optimization are iteratively performed until chain growth terminates.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.

(**B**) Representative example of model building for glycans using EModelG. Shown here is the model building for GF-1L2R5 (2.53 Å). The glycan model colored white is manually built as the reference. The auto-built glycans and protein peptide are colored pink and blue, respectively. (**C**) Comparison between manually and EModelG auto-built glycan models. The automated method recovers nearly all monosaccharide units and glycosidic bonds, with only a few false-positive monosaccharides built in peripheral density regions.

.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/bkvrt910. This version posted November 20, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All Rights Reserved. No reuse allowed without permission.



Search	By molecular composition By processing softwares By resolutions By glycosylation types
Deposit / Download	Sturctural data
Tools	Absolute handness determination Upcoming glycostructural tools
Information	Tutorials Conferences

Fig. 5 | Introduction the CryoSeek database. (**A**) The workflow for registered users to deposit data into the CryoSeek database. Users upload metadata files (for movies, micrographs, and particles), raw data, EM maps, and structure files. Following validation, the correlation of these data is established. All entries are reviewed and validated by the administration team prior to public release. Model validation is designated for future work, as robust methods for glycofibrils are yet to be established. (**B-C**) Architecture and functionalities of the CryoSeek database. CryoSeek Database is an integrated resource made of structures, EM maps, and micrographs. The database enables entry search using different criteria. Authorized users can deposit to and download from the database. A suite of tools (e.g. Ahaha) is included in the server to facilitate quick analysis of bio-assemblies.