# Metaproteomic Profiling of the Secretome of a Granule-forming Ca. Accumulibacter Enrichment

van Olst, Berdien; Eerden, Simon A.; Eštok, Nella A.; Roy, Samarpita; Abbas, Ben; Lin, Yuemei; van Loosdrecht, Mark C.M.; Pabst, Martin

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

WILEY-VCH

**Proteomics**
Proteomics and Systems Biology

| **RESEARCH ARTICLE** OPEN ACCESS

# Metaproteomic Profiling of the Secretome of a Granule-forming *Ca*. Accumulibacter Enrichment

Berdien van Olst [ID] | Simon A. Eerden [ID] | Nella A. Eštok | Samarpita Roy [ID] | Ben Abbas | Yuemei Lin [ID] |
Mark C. M. van Loosdrecht [ID] | Martin Pabst [ID]

Department of Biotechnology, Delft University of Technology, Delft, the Netherlands

**Correspondence**: Martin Pabst (m.pabst@tudelft.nl)

## ABSTRACT

Extracellular proteins are supposed to play crucial roles in the formation and structure of biofilms and aggregates. However, often little is known about these proteins, in particular for microbial communities. Here, we use two advanced metaproteomic approaches to study the extracellular proteome in a granular *Candidatus* Accumulibacter enrichment as a proxy for microbial communities that form solid microbial granules, such as those used in biological wastewater treatment. Limited proteolysis of whole granules and metaproteome isolation from the culture's supernatant successfully classified over 50% of the identified protein biomass to be secreted. Moreover, structural and sequence-based classification identified 387 proteins, corresponding to over 50% of the secreted protein biomass, with characteristics that could aid the formation of aggregates, including filamentous, beta-barrel containing, and cell surface proteins. While various of these aggregate-forming proteins originated from *Ca*. Accumulibacter, some proteins associated with other taxa. This suggests that not only a range of different proteins but also multiple organisms contribute to granular biofilm formation. Therefore, the obtained extracellular metaproteome data from the granular *Ca*. Accumulibacter enrichment provides a resource for exploring proteins that potentially support the formation and stability of granular biofilms, whereas the demonstrated approaches can be applied to explore biofilms of microbial communities in general.

## 1 | Introduction

Microbial communities are dynamic and complex environments where individual microbes interact and compete, which can lead to synergistic effects [1, 2]. Consequently, an improved understanding of the microbial community dynamics requires a holistic and detailed knowledge of both the cellular and community functionality. This understanding will accelerate the development of strategies to control or design microbial communities to provide solutions to societal challenges [3–6].

Among the various metaomic approaches, metaproteomics is particularly powerful for community characterisation because it provides insights into the expressed proteins of different community members, their interactions, and their individual contributions to the community [7]. Characterising the complete metaproteome of microbial communities is more challenging than characterising the proteome of single, laboratory-grown cultures. For example, protein extraction can be hindered by the complex structure of the communities or the complex environment surrounding the communities [8]. Additionally, protein interference from

**Summary**

- Biofilm-forming microbial communities are widespread and pose both challenges and opportunities in various settings in life.

- Structure-providing, extracellular proteins likely play a crucial role in the formation of the biofilm matrix, but these proteins are challenging to characterise due to the dynamic and complex nature of these communities.

- We used two advanced metaproteomic approaches to enrich for the extracellular proteins in a granule-forming *Ca*. Accumulibacter enrichment culture as a proxy for granule-forming communities present in wastewater treatment plants.

- The extracellular proteins were classified using structure and sequence-based annotation tools, which identified multiple different protein categories that potentially aid in granule formation, but also may provide structure to the biofilm matrix.

- Interestingly, although the granules were highly enriched for *Ca*. Accumulibacter, several structure-providing protein categories originated from other organisms.

- The obtained metaproteomic data contribute to the understanding of proteins and processes that are potentially involved in granule formation.

- This understanding could further help to optimise processes involving granular biofilms and identify candidates for the recovery of novel biopolymers for biotechnological applications.

peptides is complicated because the large number of community members requires a large and complex reference database [9, 10]. Furthermore, metagenomic approaches for constructing comprehensive reference sequence databases are highly dependent on the chosen pipeline, and their completeness is difficult to verify [9, 11, 12].

Characterising microbial community proteins beyond those involved in central metabolic pathways presents an additional challenge, particularly when analysing structure-providing proteins. Such proteins can have unconventional properties or modifications that challenge their detection [13] and their functions are difficult to verify in expression studies resulting in many unknown members. However, structure-providing proteins are crucial in many microbial communities. For example, the crosslinking ability of proteins and other biopolymers aids in the formation of an extracellular matrix that aggregates the community together in a biofilm [1, 14, 15]. The biofilm facilitates intercellular interaction, enhances water retention and nutrient sorption, and protects the microbial community against various environmental stresses [1, 2]. On the other hand, biofilm-forming communities can pose challenges and opportunities in among others medical and industrial settings [16, 17]. Therefore, it is important to understand the role of proteins beyond the central metabolic pathways.

The proteins in the extracellular matrix may play a crucial role in fine-tuning microbial communities in both natural and engineered ecosystems. For example, in biological wastewater treatment, water is cleaned from pollutants such as carbon, phosphorus and nitrogen by microbial communities that are separated from the cleaned water by gravitational settling. These communities first form flocs that only settle slowly but can be transformed into dense granules that settle much faster, thereby offering various operational and economic advantages for wastewater treatment plants [18–20]. Interestingly, these granules were reported to have a higher protein content in their extracellular matrix than flocs with the centre of the granules mainly consisting of proteins [21]. Furthermore, while certain exopolysaccharide components were found to be important for granule structure in various granule types [22, 23], this specific component was absent in granules enriched in *Ca*. Accumulibacter [23]. This might indicate that other components, such as proteins, play an important role in the formation of these granules. *Ca*. Accumulibacter has been identified as the dominant organism in dense granular wastewater treatment plant communities that perform enhanced biological phosphorus removal [11]. Although this organism cannot be isolated from the community, it can be enriched by selective growth conditions [24], resulting in dense granular structures [25] with a reduced complexity of the proteomic pool. Therefore, granular *Ca*. Accumulibacter enrichment cultures serve as valuable models for studying extracellular proteins and their potential roles in biofilm formation.

The potential processes involved in granule formation in Ca. Accumulibacter enrichment cultures have been studied previously. For instance, Barr and colleagues examined the metaproteomic changes occurring during the transition from microbial flocs to granules and identified proteins related to stress and to outer membrane proteins that allow scavenging for extracellular nutrients [26]. Another study measured the turnover of extracellular proteins and polysaccharides by tracking the incorporation of stable isotopes and found that the extracellular biopolymer turnover was comparable to the intracellular biopolymer turnover [27]. However, the harsh extraction procedures resulted in the release of intracellular proteins [28] and only 20% of the total protein intensity was regarded as extracellular [27]. Several other studies that investigated the extracellular matrix also observed large fractions of intracellular proteins [29–33]. Therefore, a gentler approach is needed to selectively measure the proteins in the extracellular space and exposed at the surfaces of the granules. Gentle separation of the extracellular has been previously achieved by gradual erosion of granule layers [34], but structure-forming proteins were not reported.

To more selectively measure the extracellular proteome of a granular *Ca*. Accumulibacter enrichment culture, we employed two metaproteomic approaches that target the extracellular space, with a particular focus on identifying proteins that contribute to structural features. The first approach focused on the supernatant of the enrichment to study proteins that can also enter the intergranular space. The supernatant has been previously employed to study soluble extracellular proteins, albeit requiring large sample volumes and prefractionation with SDS-PAGE before MALDI-TOF identification of protein bands [31]. The second approach employed in our study utilised limited proteolysis of whole granules, allowing for the cleavage of fragments from proteins that

are exposed to the extracellular space [35]. The metaproteomic data were analysed using a metagenomics-constructed database from the same enrichment and were classified with a range of protein function and structure prediction tools. Together, these two approaches reveal a large spectrum of proteins that are released into the extracellular space, of which some may support aggregation and the development into a granular biofilm.

## 2 | Methods

### 2.1 | Reactor Operation

The *Ca*. Accumulibacter enrichment culture was cultivated in a 1L laboratory-scale sequencing batch reactor controlled by an Applikon System 10 controller as described previously [27, 36] with some modifications. The inoculum originated from the return flow (return activated sludge) of the Harnaschpolder wastewater treatment plant in Den Hoorn, the Netherlands. The sequencing batch reactor operated in 6 h cycles. Each cycle started with a 10 min nitrogen sparging phase to remove any dissolved oxygen from the liquid, followed by a 30 min feeding phase, a 105 min anaerobic phase, and a 135 min aerobic phase. In the last minute of the aerobic phase, approximately 30 mL of sludge was discharged, followed by a 30 min settling phase and a 50 min effluent discharge phase, after which the cycle was repeated. The reactor was stirred at 600 rpm in all phases, except for the settling and the effluent discharge phases. The culture was fed with 50 mL COD medium (124 mM Sodium Acetate), 50 mL mineral medium (20 mM Ammonium Chloride, 12 mM Magnesium Sulphate, 5 mM Calcium Chloride, 13 mM Kalium Chloride, 13 mM Sodium Dihydrogen Phosphate, 0.69 mM N-Allylthiourea ($C_4H_8N_2S$), 18 mL of Trace elements [37] and 500 mL of distilled water with Bacto yeast extract (10 mg/L). The pH was set to 7.2 ($\pm$0.05), maintained by addition of 0.5 M HCl (with antifoam) and 0.5 M NaOH.

### 2.2 | Sampling

Samples were taken in the last minutes of the aerobic phase, while the reactor was still stirred. A large sample (> 30 mL) was drawn, and the granules were separated from the supernatant by settling for a minute. The granules for whole granule proteome and metagenome analysis were immediately frozen and stored at −20°C until further processing, whereas the granules for shaving were freshly processed (see below). The supernatant (1.5 mL) was centrifuged (14,000 rcf, 3 min, 4°C) and the proteins were precipitated in triplicate with trichloroacetic acid (TCA; 4:1 ratio SN:TCA) at 4°C for 30 min, pelleted by centrifugation (14,000 rcf, 15 min, 4°C), and subsequently stored at −20°C until further processing. In addition to the aerobic samples, metagenome and supernatant samples were taken in the last minutes of the anaerobic phase in the same cycle and processed analogous to the aerobic samples. The *Ca*. Accumulibacter enrichment of the culture was verified by fluorescence in situ hybridisation (FISH) as described previously [38] using the probes Acc470 (*Ca*. Accumulibacter aalborgensis), Acc471 (*Ca*. Accumulibacter affinis and *Ca*. Accumulibacter proximus) and a mix of EUB338, EUB338-II, and EUB338-III (all bacteria).

### 2.3 | Limited proteolysis of granules ('enzymatic granule shaving')

Intact granules were enzymatically digested ('shaved') by either trypsin, chymotrypsin, lysC or lysC/trypsin in triplicate. Fresh granules (30 mg) were incubated with 15 µL of 0.1 µg/µL protease in 500 µL 100 mM ammonium bicarbonate (ABC) in a ThermotopMixer with Thermotop (37°C, 300 rpm). The lysC/trypsin samples were initially digested with lysC, and trypsin (5 µL of 0.1 µg/µL) was added after 2 h to prevent digestion of lysC by trypsin. Granule damage was checked every hour by comparing to a control incubation without protease. After 4 h of incubation, the granules were collected (2 min, 5000 rcf) and the supernatant was transferred to clean LoBind Eppendorf tubes. To promote denaturation of the shaved-off proteins, 180 mg urea (corresponding to 6 M) was dissolved in the supernatant of the lysC samples. After an overnight incubation, the samples were centrifuged (2 min, 14,000 rcf) to ensure no cell debris was transferred to the downstream processing. The digested sample was cleaned by solid phase extraction (SPE) on an Oasis HLB 96-well µElution Plate (Waters). The wells were conditioned in 100% MeOH and equilibrated with pure LC-MS $H_2O$. Following equilibration, the samples were loaded onto the columns, washed twice with 5% MeOH in LC-MS $H_2O$, and eluted with 2% formic acid in 80% MeOH and subsequently 1 mM ABC in 80% MeOH. The collected flow-through was transferred to LoBind Eppendorf tubes and dried in a centrifuge concentrator at 45°C. The dried sample was dissolved in 3% acetonitrile in 0.1% formic acid in LC-MS $H_2O$ to a concentration of approximately 0.25 mg/mL as estimated from the absorbance at 280 nm using a Nanodrop.

### 2.4 | Protein extraction from whole granule samples

Granules for whole granule proteome analysis were lysed in duplicate by combining 40 mg granules, 150 mg acid-washed glass beads, 175 µL of B-PER reagent, and 175 µL of 1% NaDOC in 50 mM TEAB pH 8 followed by three rounds of bead beating for 90 s with intermediate cooling on ice for 30 s. The cell suspension was subsequently incubated for 3 min at 80°C in a ThermotopMixer with Thermotop (1000 rpm) followed by 10 min of sonication in an ultrasonic bath. The remaining cell debris was pelleted (10 min, 14,000 rcf, 4°C), the supernatant was transferred to a clean LoBind Eppendorf tube, and the proteins were precipitated with TCA (ratio 4:1) at 4°C for 30 min followed by centrifugation (14,000 rcf, 15 min, 4°C).

### 2.5 | Sample processing for whole granule and supernatant samples

Both the whole granule and the supernatant protein pellets were washed with 200 µL ice-cold acetone (3 min, 14,000 rcf) and dissolved in 6 M urea (granules: 100 µL; supernatant: 50 µL). The dissolved pellet was reduced with 10 mM dithiothreitol (DTT; granules: 30 µL; supernatant: 5.5 µL) for 60 min using a ThermotopMixer with Thermotop (37°C, 300 rpm) and alkylated with 20 mM iodoacetamide (IAA; granules: 30 µL; supernatant:

5.5 µL) for 30 min in the dark. Subsequently, the samples were diluted in 100 mM ABC (granules: 440 µL; supernatant 285 µL) digested with either 0.1 µg/µL trypsin or 0.1 µg/µL chymotrypsin (granules: 15 µL; supernatant: 10 µL) for approximately 18 h in a Thermomixer with Thermotop (37°C, 300 rpm). The digested sample was cleaned by SPE as described for granule shaving and concentrated to approximately 0.5 mg/mL and 0.25 mg/mL for granule and supernatant samples, respectively. The whole granule samples digested by trypsin were injected twice in the LC-MSMS.

## 2.6 | Metaproteomics

Metaproteome samples were loaded on an EASY nano-LC 1200 equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm × 150 mm, 2 µm and 100 Å). For the granule-based samples, the flow rate was maintained at 350 nL/min with 0.1% formic acid in water (solvent A) and a 120 min linear gradient from 5 to 65 of 80% acetonitrile and 1% formic acid in water (solvent B). The elute was sprayed using a Nanospray Flex (Proxeon) in a Thermo Q Exactive Plus mass spectrometer (Thermo Scientific, Germany) operated in data-dependent acquisition (DDA) mode. Full spectra were recorded with peptide signals between 385 and 1250 $m/z$ at a 70K resolution with an automatic gain control of 3e6 and a maximum injection time of 75 ms. The top 10 most abundant precursor ions with a charge state between 2 and 5 were fragmented (28% normalised collision energy, 2.5 $m/z$ isolation window, and 0.2 $m/z$ isolation offset) and recorded at a 17.5K resolution with an automatic gain control of 2e5 and a maximum injection time of 100 ms. The supernatant samples were analysed with a 60 min linear gradient from 2% to 40% of solvent B and full spectra were recorded between 350 and 1250 $m/z$ with a maximum injection time of 100 ms. The top 10 most abundant precursor ions with a charge state between 2 and 5 were fragmented (3.0 $m/z$ isolation window) with an automatic gain control of 5e5. The mass spectrometric raw data was analysed per approach using PEAKS Studio (Bioinformatics Solutions Inc., Canada) and searched against the metagenomic-constructed database (see below) and the cRAP contaminant database. A maximum of two missed cleavages was allowed for the whole granule and supernatant samples and four missed cleavages for the shaved samples. The digestion mode was set at specific for the whole granule and supernatant samples, and semispecific for the shaved samples. Carbamidomethylation (C) was set as fixed modification in samples that were alkylated (i.e., whole granule and supernatant samples), whereas deamidation (NQ), oxidation (M) and acetylation (Protein N-term) were set as variable modification in all samples. Peptide identifications were filtered for 1% false discovery rate, protein identifications required at least one unique peptide and the top proteins per protein group were exported. Using these approach-specific protein groups, the identified proteins were assigned to experiment-wide protein groups: proteins were assigned to the same experiment-wide protein group if they belonged to the same approach-specific protein group in at least one approach and did not belong to different approach-specific protein groups in any of the approaches. The metaproteome intensities were converted to iBAQ using the theoretical number of peptides per protease [39].

## 2.7 | DNA extraction from granules

DNA extractions were performed in duplicate based on the Phenol/Chloroform extraction as described by Trigodet et al. [40]. Briefly, 80 mg of frozen granules was homogenised using a potter device in a 2 mL Eppendorf and incubated at 37°C for 1 h in 1 mL TLB (10 mM Tris-HCL pH 8.0, 25 mM EDTA pH 8.0, 100 mM NaCl, 0.5% [w/v] SDS, 20 µg/mL RNase A). Subsequently, the samples were incubated with proteinase K (final concentration 200 µg/mL) at 50°C for 2 h with gentle inversing every 30 min. The lysate was transferred to a 2 mL phase-lock tube (Quantabio, Germantown MD, USA) containing 800 µL phenol, inverted for 10 min and centrifuged for 15 min at 17,000 rcf, which was repeated with 800 µL phenol/chloroform (PCI 25:24:1). The DNA was precipitated by adding 720 µL (0.8 vol) 100% isopropanol and centrifuging for 15 min. The resulting pellet was washed twice with 1 mL 70% ethanol and resuspended overnight in TE-buffer. The resuspended sample was cleaned with a Qiagen Genomic Tip-20 (Hilden, Germany) according to the manufacturer's protocol with isopropanol precipitation. The high molecular weight fraction was subsequently selected based on the method described by Stortchevoi et al. [41]. In short, 4000 ng DNA was incubated with 10 mM TRIS pH8, 0.14 vol Mg-PEG buffer (70 mM $MgCl_2$ in 0.7% (v/v) Polyethylene glycol 8000) and washed AMPure XP beads (Beckman Coulter) for 60 min while shaking to keep the beads in suspension. The beads were collected on a magnet and washed twice with 500 µL 80% ethanol. The DNA was released from the beads by incubating the beads in 50 µL $H_2O$ for 10 min at 37°C while shaking (500 rpm) and stored at −20°C until further processing. Part of the DNA was loaded onto an agarose gel and electrophoresis was performed to evaluate the size of the DNA and its suitability for Illumina library preparation.

## 2.8 | Metagenomics

DNA libraries were prepared using the Nextera XT DNA library preparation kit (Illumina) and sequenced on an Illumina NovaSeq6000 with a paired-end 150 bp strategy (Baseclear, Leiden, The Netherlands) for 10GB. The data quality was assessed, the metagenome was assembled using MEGAHIT (v 1.2.9) [42], and genes were predicted and annotated using Prokka (v 1.14.5). The genes predicted in the four samples were clustered to reduce sequence redundancy using CD-HIT [43] at 95% sequence identity level, resulting in a single database that was used for protein identification. Taxonomic classification of the raw reads was predicted using Kraken 2 [44].

## 2.9 | Processing of metaproteomic data

The identified proteins were annotated with information about their subcellular location, taxonomy, functionality and sequence properties. Secreted proteins were identified by predicting the presence of a signal peptide using SignalP (v 6.0) [45], whereas transmembrane domains were predicted using DeepTMHMM (v 1.0) [46]. The location of the identified proteins was further predicted by PSORTb (v 3.0.3) [47] and by DeepLocPro (v 1.0) [48], a prokaryotic extension to DeepLoc [49]. The taxonomic origin was predicted using GhostKoala [50] and further validated for

the identified aggregating proteins by DIAMOND [51]. Proteins were functionally annotated using Eggnog mapper (v 2.1.12) [51–53]. Sequence-based protein properties were calculated using the ProtParam module in the SeqUtils package of Biopython (v 1.78), whereas protein structural properties were calculated using NetSurfP (v 3.0) [54]. To classify proteins as potentially aggregating, we ranked all measured proteins by the properties describing charge, hydrophobicity (GRAVY), secondary (percentage of sequence in beta-sheet) and tertiary (percentage of cysteines) structures, aromaticity, protein length, surface accessibility and disorder. Subsequently, proteins were selected as aggregating if (i) at least one of these properties belonged to the top 10% (GRAVY, beta-sheet percentage, cysteine percentage, aromaticity, protein length), or bottom 10% (surface accessibility, disorder), (ii) the protein abundance belonged to the top 50% in at least one approach, and (iii) the protein was predicted to be secreted or extracellular. The selected proteins were clustered using a Euclidean distance and a complete clustering method. As both a negative and a positive charge might contribute to aggregating, the protein charges were converted to absolute numbers before selecting the top 10%, but the nonconverted values were used for clustering.

## 3 | Results

Here we employed two different metaproteomic approaches to identify and characterise the secretome of a *Ca*. Accumulibacter enrichment (hereafter Accumulibacter) as a proxy for aerobic granular sludge employed in wastewater treatment, with a focus on identifying potential aggregate-forming proteins that play a role in the formation and structure of the granular matrix. As Accumulibacter cannot be cultured in pure form [24], we enriched a sample obtained from a wastewater treatment plant for Accumulibacter by cyclic anaerobic feeding of acetate and aerobic growth phases in a bioreactor. Subsequently, the enrichment was analysed by metagenomics and metaproteomics, where the metagenome served as a reference for database searching experiments. Two metaproteomic approaches were employed to identify the secreted metaproteome of the enrichment cultures, namely 'enzymatic granular shaving' and 'supernatant proteomics' (Figure 1A). The first approach aimed to identify proteins inside the granule's biofilm by directly adding proteases to intact granules, thereby 'shaving' the granules of their extracellular and transmembrane proteins. This approach relies on the digestion of native proteins, which are not easily accessible for proteases. Therefore, multiple proteases were employed in parallel to enhance protein identification (Figure S2). The second approach aimed to identify proteins in the culture's supernatant, which is expected to be enriched in secreted proteins. For this supernatant approach, we compared samples harvested from the anaerobic feeding phase and the aerobic growing phase, which showed only minimal proteome differences between these two phases (Figure S3). These two approaches were compared to the proteome obtained from lysed whole granules (Figure S1).

### 3.1 | The extracellular metaproteome of a *Ca*. Accumulibacter enrichment

The three different approaches identified in total almost 4000 protein groups, of which 71% was identified in the whole granules,

52% was identified in the supernatant, 25% was identified in the shaved granules, and 11% was identified in all approaches (Figure 1B). These shared protein groups were mostly high-abundant but also included some protein groups with average abundances (Figure 1C). Only 1% of the identified protein groups were found in both the shaved granules and the supernatant, which correlated only moderately (Pearson: 0.41; Figure S4A), and both approaches shared more proteins and correlated better with the whole granule approach (Pearson: 0.58 for shaved granules and 0.57 for supernatant; Figure S4A). Each approach also identified a unique set of protein groups (Figure 1B) which were mainly average and low abundant (Figure S4B). Notably, the total protein biomass (i.e., all identified protein abundances summed) was two times lower in the shaved granule samples compared to whole granule samples (Figure 1D), which corresponds to the 2-fold lower amount of proteolytic digest analysed. Conversely, the protein biomass observed in the supernatant samples was 7-fold lower compared to the whole granules, which likely reflects both the lower injection volume and the shorter gradient used to analyse the samples, thereby providing fewer identifications.

The shaving and supernatant approaches were developed to investigate the proteins that are secreted by the microbes in the enrichment. Therefore, we examined the proteins identified in each approach for the presence of transmembrane domains and signal peptides that indicate secretion. These tools predicted that of the almost 4000 protein groups identified by all approaches, 19% had a signal peptide, 8% had a transmembrane domain, and 8% had both. These percentages were similar for the whole granule and supernatant approaches (Figure 2A), indicating that a substantial fraction of the identified protein groups can be regarded as intracellular. Conversely, 25% of the protein groups in the shaved granules were predicted to be secreted and only a limited number of transmembrane proteins was identified. The lack of transmembrane proteins in the shaved granules also partly explains the substantial number of proteins that were only identified in whole granules and the supernatant (Figure 1B) of which 30% was predicted to have a transmembrane domain. Interestingly, the small number of secreted proteins accounted for the majority of the protein biomass. Of the shaved biomass, 54% was predicted to be secreted and transmembrane proteins contributed only marginally (Figure 2B) while 66% of the supernatant biomass was predicted to be secreted, of which almost half was also predicted to have a transmembrane domain. In the whole granule samples, only 43% of the protein biomass was predicted to be secreted, of which half was also predicted to have a transmembrane domain, similar to the supernatant samples. Moreover, in the shaved granules, 50% and 75% of the secreted protein biomass was covered by 2 and 9 proteins, respectively. Conversely, in the supernatant, 6 and 26 proteins covered over 50% and 75% of the secreted protein biomass, respectively, which is similar to the whole granules where 4 and 21 proteins covered 50% and 75% of the secreted protein biomass, respectively.

The absence of a predicted secretion signal does not necessarily mean that the proteins are not secreted, as there might be alternative routes to the extracellular space. Therefore, the protein location was further examined with PSORTb and DeepLocPro (a prokaryotic extension to DeepLoc). Both these tools confirm the high number but low abundance of cytoplasmic protein groups in shaved granules and the supernatant, albeit with different
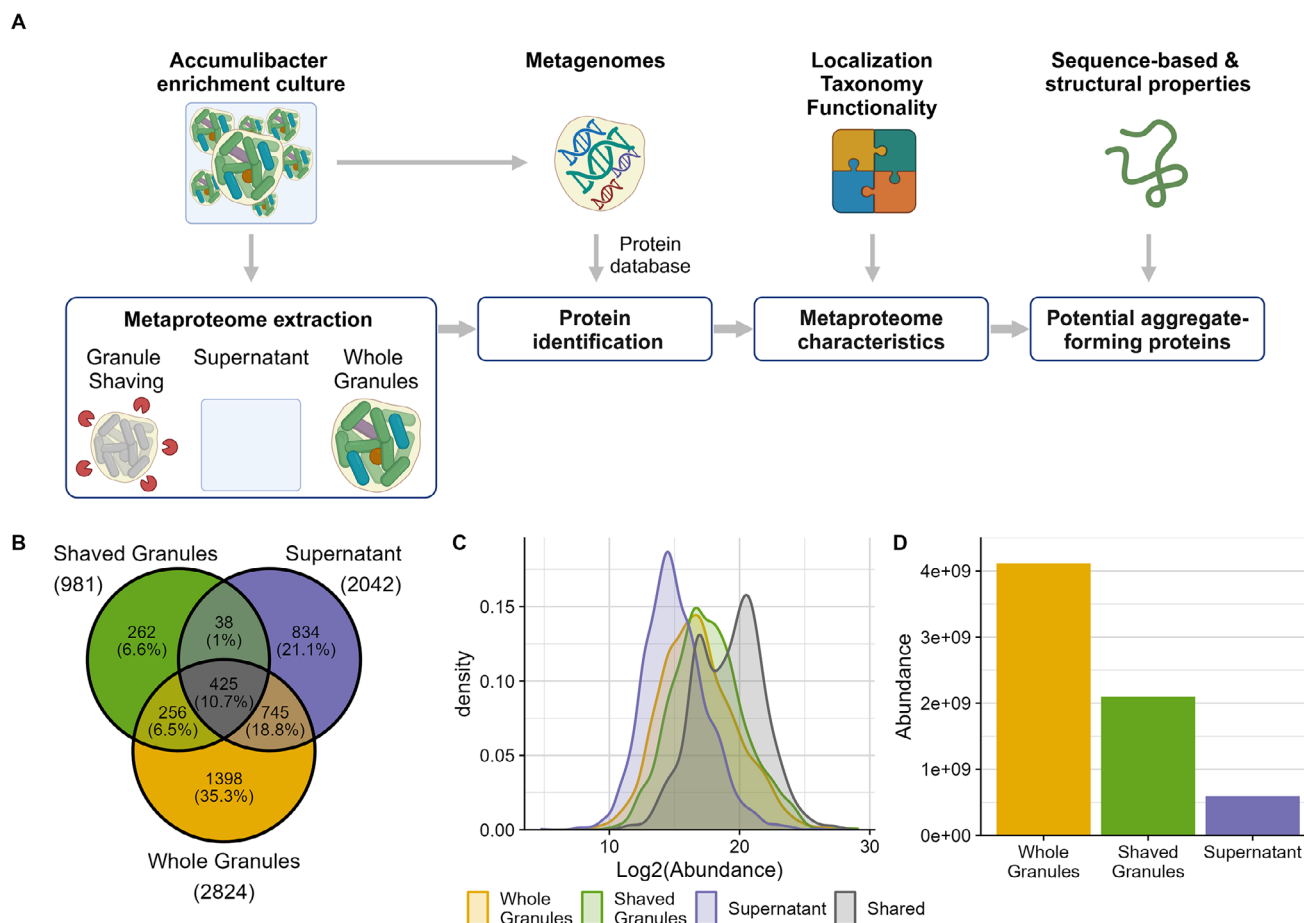
**FIGURE 1** | Workflow and global comparison of different metaproteomics approaches applied to *Ca*. Accumulibacter enrichment. (A) Workflow (B) Venn diagram displaying the identified protein groups per approach and the overlap between them. (C) Overlap in the log2 protein abundance (x-axis) for the protein groups identified by the different approaches and the protein groups shared by all approaches (colours). (D) Total protein biomass calculated from the summed protein abundances (y-axis) per approach (x-axis).

percentages (Figure 2C,D; Figure S5). Moreover, they validate the substantial noncytoplasmic fraction in whole granules. Interestingly, both tools indicate that the different approaches enrich for proteins at different locations (Figure 2C,D). Most biomass in the shaved granules is assigned to periplasmic protein groups, whereas most biomass in the supernatant is assigned to outer membrane proteins. In addition, a large fraction of the supernatant biomass is assigned to extracellular proteins when predicting with DeepLocPro, which is substantially less when predicting with PSORTb. In whole granules, the biomass mainly contains cytoplasmic proteins, followed by outer membrane and periplasmic proteins. Hence, while the shaved granule and supernatant approaches still result in a significant number of intracellular proteins, these proteins only account for a small fraction of the protein biomass confirming that both approaches enrich for secreted proteins, although from different cellular locations.

The secreted proteins are investigated in an Accumulibacter enrichment culture, as Accumulibacter is expected to be involved in dense granule formation [11, 25]. To examine the Accumulibacter enrichment, we investigated the culture's taxonomy as identified by the different approaches. The two approaches that immediately target granules displayed similar enrichment patterns across the different taxonomy levels and indicated a

high enrichment (96%) of the Pseudomonodota phylum, the Betaproteobacteria class (∼89%) and the *Ca*. Accumulibacter genus (∼78%; Figure 3). The high enrichment of Accumulibacter was supported by FISH (Figure S6), but was not supported by metagenome-based taxonomic classification (Figure S7), although the same phyla and classes were highlighted albeit at different relative abundances. Conversely, only 88% of the protein biomass in the supernatant belonged to the Pseudomonodota phylum, only 66% to the Betaproteobacteria class and only 43% to the *Ca*. Accumulibacter genus. The remaining proteins in the supernatant could be attributed to more than 100 different taxonomies (based on protein sequence alignment) but were mostly low abundant (< 2%). Nevertheless, 2% of the protein biomass could be attributed to viruses, 6% to the Bacteroidota phylum, 17% to the Gammaproteobacteria class, and 5% to the Hydrogenophaga genus, suggesting the presence of non-Accumulibacter microcolonies or microbial aggregates in the bioreactor that are not directly part of the granular community, and such cells were indeed visible in the bioreactor culture (data not shown). Hence, while the granule community is highly enriched for Accumulibacter, this was not reflected in the identified supernatant proteome. Nevertheless, the observed Accumulibacter-originating biomass in the supernatant was well enriched for secreted proteins (Figure S8).
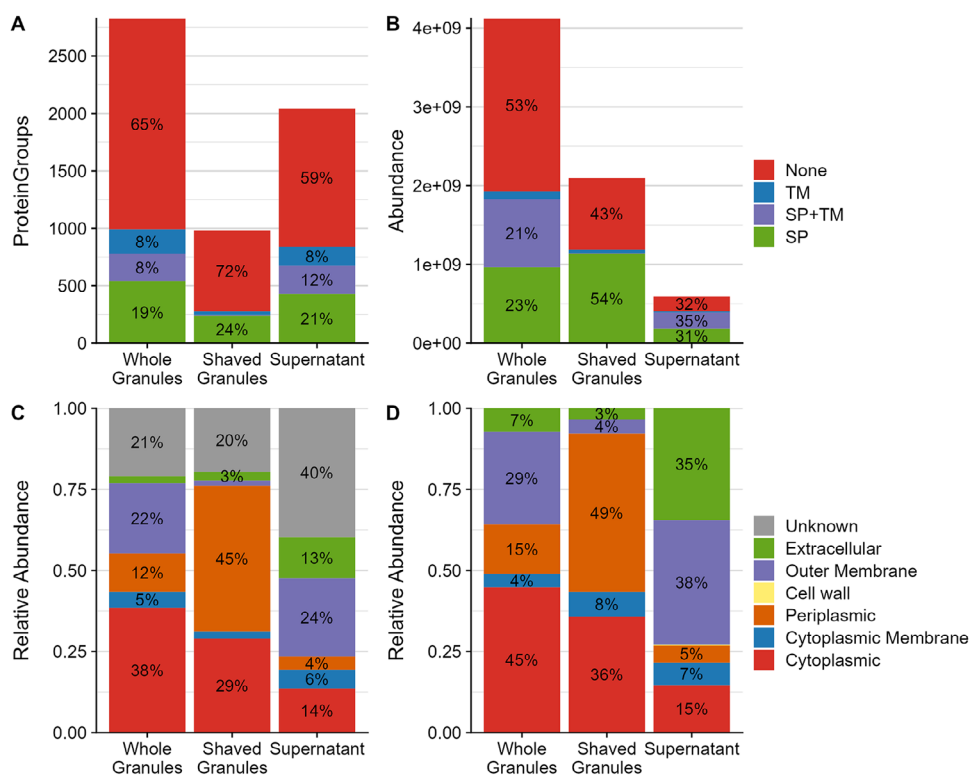
**FIGURE 2** | Enrichment of secreted and extracellular proteins by the different metaproteomics approaches. (A) The number of identified protein groups (y-axis) and (B) the summed protein abundance (y-axis) of proteins with a predicted SP (green), a predicted TM (blue), both predictions (SP+TM; purple) or no predictions (None; red) for the different approaches (x-axis). (C, D) The relative abundance of proteins (y-axis) predicted at different cellular locations (colours) as predicted by (C) PSORTb and (D) DeepLocPro for the different approaches (x-axis). TM, transmembrane domain; SP, signal peptide.

## 3.2 | Functional classification of identified proteins

To gain more insight in the protein functions, proteins were further assigned to Clusters of Orthologous Groups (COGs) [55, 56]. Of the 4000 identified protein groups, 21% could not be assigned to a COG group and 12% had an unknown function (S category). The missing functions predominantly impacted the supernatant samples as 45% of the protein biomass and 51% of the secreted and extracellular protein biomass could not be assigned (i.e., either NA or S; Figure 4). The most abundant protein group (10%) without assignment corresponded with a hypothetical protein from Accumulibacter with a secretion signal, which was also found in whole granules (2%) but not in shaved granules (Table S1). The most abundant group (3%) with unknown function corresponded to a tail sheath protein that was only found in the supernatant, which was predicted to originate from Hydrogenophaga but also shared a 100% sequence identity with a major tail sheath protein from Accumulibacter. The 'Cell wall, membrane and envelop biogenesis' category (M) was the most abundant assigned category in the whole granules and supernatant, covering about 21% of the protein biomass, which corresponded to 44% and 28% of the secreted and extracellular protein biomass, respectively (Figure 4). In contrast, only 6% of the secreted and extracellular protein biomass in the shaved granules belonged to this category, which was explained by the high number of transmembrane proteins that were only limitedly identified in the shaved granules. The main contributing

proteins in this category were various high-abundant, secreted, outer membrane porins (Table S1) that cover 19% and 17% of the protein biomass in the whole granules and supernatant, respectively, making it the most abundant proteins in these approaches. The 'Energy production and conversion' category (C) was the most abundant fraction in the shaved granules, covering 31% of the protein biomass and 46% of the secreted and extracellular protein biomass, with the latter being almost 14 and 44 times higher compared to the whole granules and supernatant, respectively. The main contributing protein in this category was a secreted cytochrome C from Accumulibacter, which was also the most abundant protein in the shaved granules covering 23% of the protein biomass. The 'Signal transduction mechanisms' category (T) also covered a large fraction of the secreted and extracellular protein biomass in both the shaved granules (24%) and whole granules (11%), whereas it only covered 2% in the supernatant. The main contributing protein in this category was a secreted, hypothetical protein with a Calcium/Chemotaxis domain. This protein was the second and third most abundant protein in the shaved and whole granules, making up 12% and 4%, respectively. Both the 'Cell motility' and 'Intracellular trafficking, secretion, vesicular transport' categories (N and U) covered ~12% of the secreted and extracellular protein biomass in the supernatant, which was over 2-fold and 9-fold larger than in the whole and shaved granules, respectively. The main contributing proteins in these categories were various fimbrial proteins that are involved in cell adhesion and assigned to both classes. Interestingly, the most abundant fimbrial protein
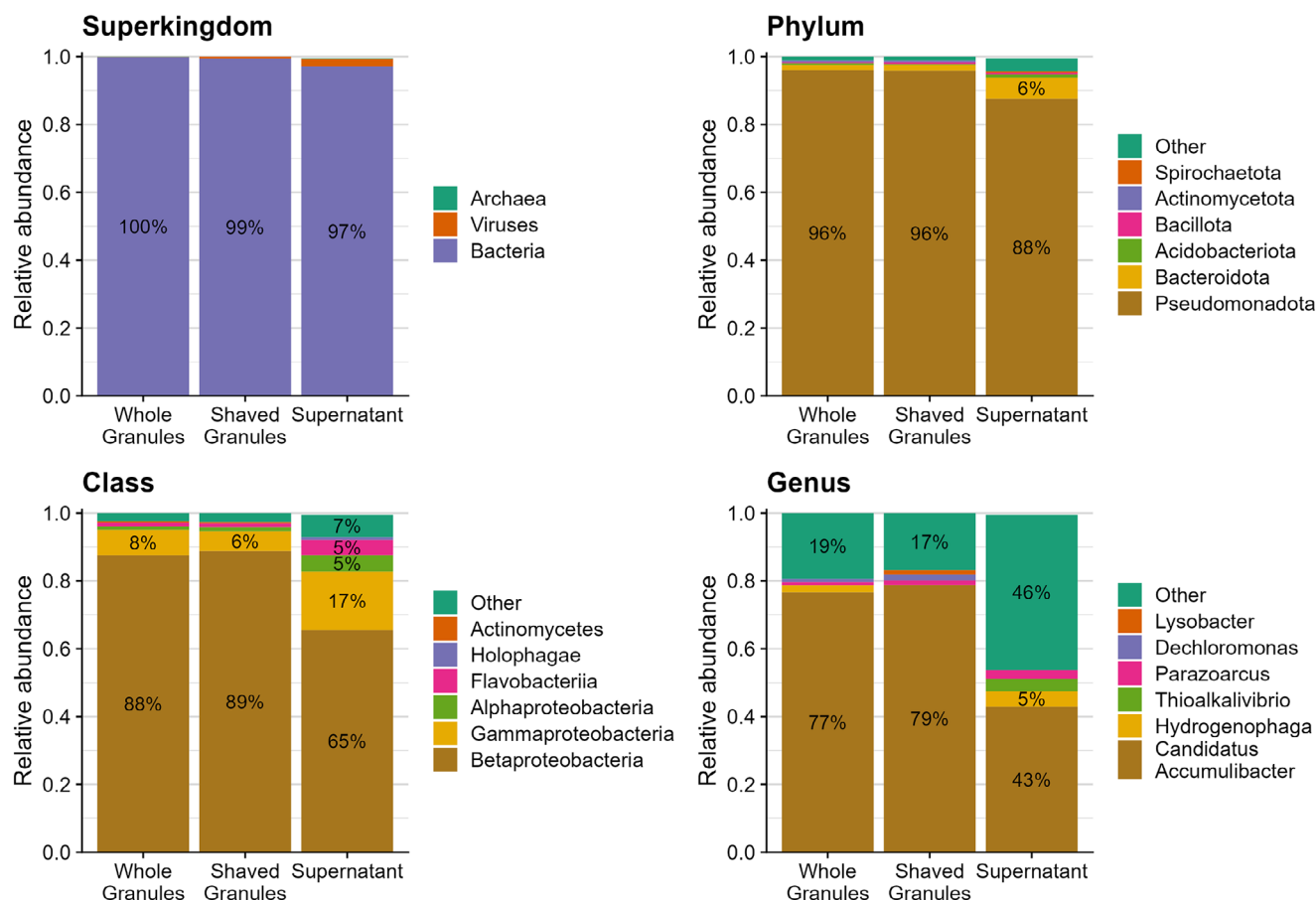
**FIGURE 3** | Taxonomic composition obtained from different metaproteomics approaches for *Ca*. Accumulibacter enrichment. Enrichment for different organisms based on relative contribution to the protein biomass (y-axis) in the different approaches (x-axis) on different taxonomy levels (panels).

was assigned to Accumulibacter, which was twice as abundant as the fimbrial proteins of two other Betaproteobacteria. Taken together, the approaches enriched for different proteins in different functional categories although there was some overlap.

### 3.3 | Large spectrum of proteins with aggregate-forming properties

The higher extracellular protein content in granules compared to floccular sludge [21] suggests that structure-providing proteins are potential contributors to biofilm formation in granules. Therefore, the secretome was additionally investigated for proteins with structure-providing functionalities. We assumed that these proteins can be recognised by intrinsic sequence-based properties that allow protein-protein interactions towards self- or subunit-aggregation or allow noncovalent interactions with other biopolymers that aid in aggregating the biofilm matrix [14]. To this end, we predicted net charge, hydrophobicity (GRAVY), aromaticity, cysteine percentage, and protein length as aggregate-forming properties that allow interactions and predicted beta-sheet percentage, relative surface accessibility, and disorder as aggregate-forming properties that are the resultant of (multiple) interactions. Proteins were considered to be potentially aggregat-

ing when at least one of the properties was in the top 10% of the measured proteins and the protein abundance was in the top 50% in at least one approach. To further narrow down our search, we only selected proteins that were predicted to be secreted or extracellular as granule formation is expected to be intentional. This resulted in a list of 387 protein groups that were clustered together into 9 main clusters (Figure 5; Table S2).

Cluster 1 and 2 mainly contained short proteins with high surface accessibility. In Cluster 1, the proteins were also hydrophobic, and the most abundant protein contained a PEP-CTERM motif and clustered with proteins with similar motifs. The proteins in Cluster 2 additionally had a high beta-sheet percentage. The most abundant protein in Cluster 2 contained a Phage_T4_gp19 domain and clustered together with proteins containing a T6SS_HCP domain, which suggests that these proteins are subunits for different tubular proteins (IPR010667; IPR008514). Similar to the first clusters, Cluster 3 contained short proteins, but they had a low beta-sheet percentage and a high cysteine percentage. The most abundant proteins in this cluster were pilus and cytochrome C proteins. The proteins in Cluster 4 were rather the opposite of proteins in Cluster 3 as they were generally longer, had a low cysteine percentage, a high beta-sheet percentage, and a high aromaticity. The proteins in this cluster were mainly beta-barrel membrane proteins, including porins and outer membrane
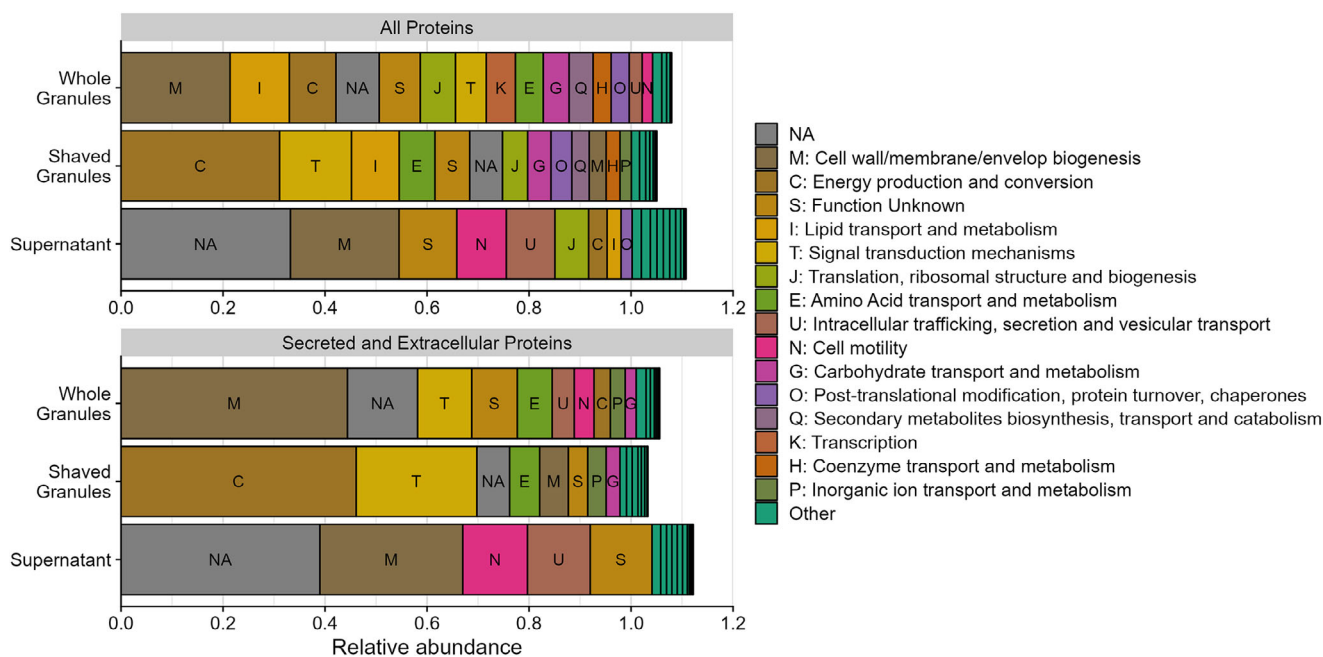
**FIGURE 4** | Relative protein biomass of COGs for whole granule, shaved granule, and supernatant metaproteomes. Distribution of the relative protein abundance (x-axis) per metaproteomic method (y-axis) over different COG categories for the total protein biomass (upper panel) and for the protein biomass predicted to be secreted or extracellularly located (lower panel). Categories with a relative abundance lower than 2% are coloured as 'Other.' Proteins annotated with more than one COG category are allocated to all those categories, resulting in a total relative abundance larger than 1. COG, clusters of orthologous groups.

proteins. Interestingly, this cluster also contained a Flavobacterium protein with a curlin repeat domain which is involved in cell adhesion (IPR009742). The proteins in Cluster 6 had similar properties as those in Cluster 4 and also contained abundant beta-barrel membrane proteins, although these proteins were longer. The properties in Cluster 7 resembled those in Cluster 6, except that the proteins in Cluster 7 had a lower beta-sheet percentage and a negative charge. The most abundant proteins in this cluster bind extracellular solutes, including starch, or have a sulfatase domain. The proteins in Cluster 5 were long and negatively charged, but relatively hydrophobic and contained a relatively large fraction of beta-sheets. The two most abundant proteins in this cluster were labelled hypothetical but showed decent sequence homology (49%–58%) against proteins from the *Paracidovorax* genus labelled as cell surface proteins. Furthermore, this cluster contained multiple proteins with a polycystic kidney disease (PKD) domain, which are mainly found in large extracellular cell surface glycoproteins (IPR000601), multiple proteins with a fibronectin domain that is among others involved in protein binding and cell adhesion (IPR003961), and multiple proteins annotated as outer cell wall protein that shared sequence homology (32%–33%) with proteins from the Bacillota phylum containing an S-layer homology domain. Remarkably, despite the high enrichment of Accumulibacter (75%–80%), most proteins in this cluster did not originate from Accumulibacter (Table S2). Clusters 8 and 9 were the smallest clusters and mainly contained proteins with a high positive or negative charge, with Cluster 8 containing proteins with a low surface accessibility without transmembrane domain and Cluster 9 containing proteins with a high surface accessibility. The most abundant proteins in these clusters did not have similar functions, indicating that these clusters were not yet well defined.

In total, our identified proteins covered between 55% and 70% of the secreted and extracellular protein biomass. Clusters 3 and 6 were the most abundant in the supernatant and whole granules, although Cluster 6 was over 2-fold more abundant than Cluster 3 in whole granules (Figure S9). Conversely, only Cluster 3 was highly abundant in shaved granules. Furthermore, Cluster 1 was more abundant in the supernatant compared to the whole and shaved granules. Taken together, we identified different clusters of proteins based on physicochemical properties, which included abundant proteins such as pili and porins.

## 4 | Discussion

We employed two advanced approaches that successfully enriched for extracellular and secreted proteins when compared to the whole granule approach. This obtained enrichment also exceeded the enrichment observed when extracting extracellular polymeric substances in alkaline conditions at high temperatures [27, 28]. Nevertheless, a substantial number of intracellular proteins was identified by both approaches that were employed in this study, although their low abundance indicates that these proteins are likely released in the extracellular space following cell lysis [57].

Interestingly, both approaches enriched for different proteins. The enzymatic granule shaving approach uniquely identified multiple cytochrome C proteins, but could not observe transmembrane and structural proteins, which are challenging to digest in their native form. We found that post-shaving digestion under denaturing conditions (6 M urea) tremendously increased the reproducibility and number of identified peptides. As most pro-
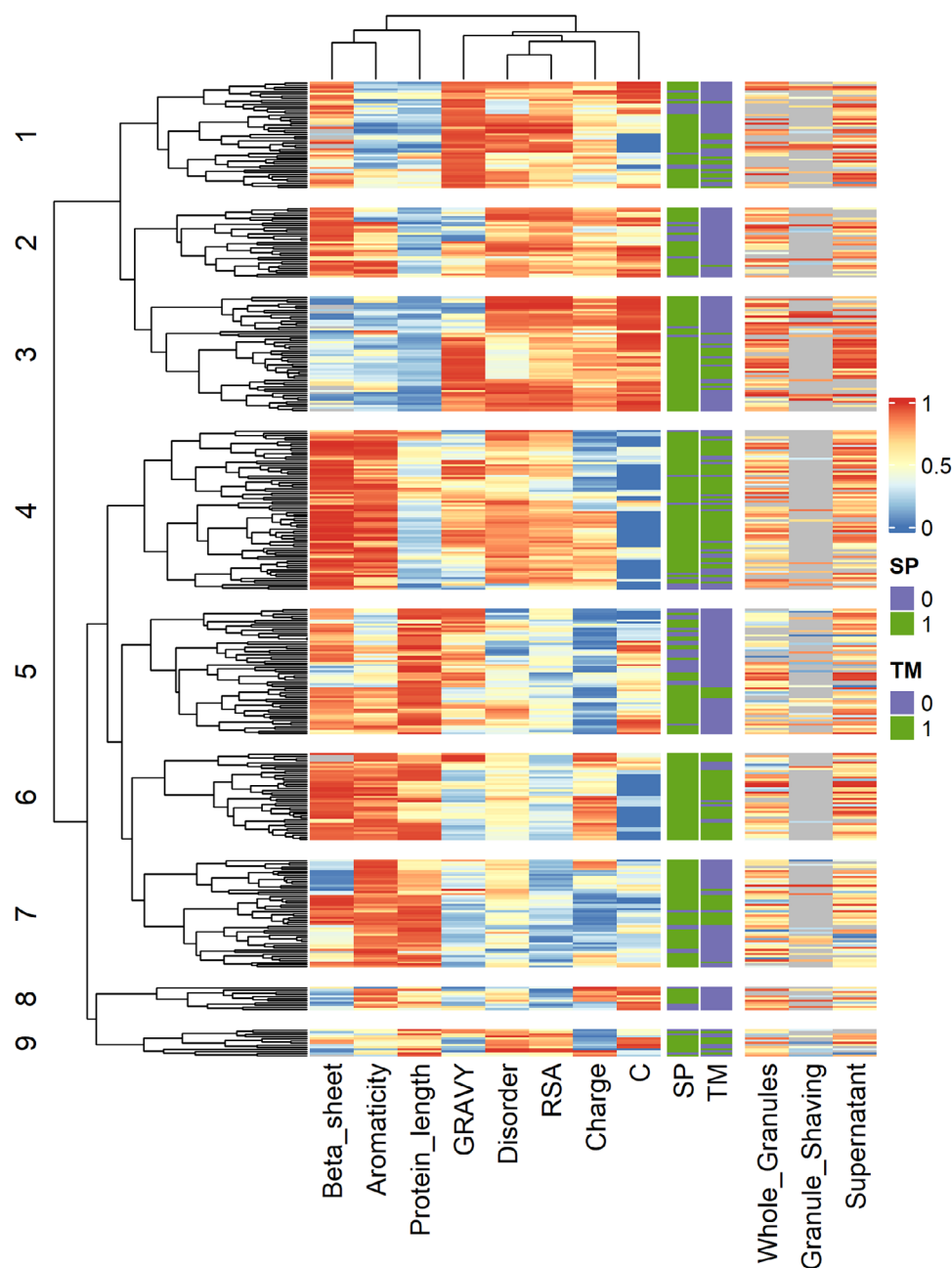
**FIGURE 5** | Clustering of proteins according to aggregating properties. Clustering of the different protein groups (y-axis) using the ranked sequence-based properties (x-axis; colours). The proteins were classified based on the following aggregate forming properties: the beta-sheet percentage, aromaticity, protein length, hydrophobicity (GRAVY), disorder, RSA, net charge, and cysteine percentage. The protein groups were annotated based on the presence of an SP, a TM and the ranked protein abundance for the different approaches. Detailed information about each cluster can be found in Table S2. RSA, relative surface accessibility; SP, signal peptide; TM, transmembrane domain.

teases are not resistant to high urea concentrations, other denaturing agents such as Rapigest might improve the identification of these proteins during the shaving treatment [58]. Nevertheless, the digestion of native proteins results in an underestimation of the extracellular protein biomass and complicates direct comparison between the shaving approach and the other approaches, both qualitatively and quantitatively. In the supernatant, the fraction of Accumulibacter proteins was smaller compared to the granule-based samples. The lower Accumulibacter enrichment could be the result of non-Accumulibacter microcolonies or microbial aggregates present in the reactor [25], even though

we aimed to remove them by centrifugation. Therefore, it might be more effective to select against these cells in the culture themselves, for example, by shorter settling or shortening the aerobic phase. Alternatively, the low Accumulibacter enrichment might originate from the taxonomic classification method. We have identified multiple highly abundant proteins that were not assigned to Accumulibacter, even though their sequence corresponded to Accumulibacter. Analogously, Accumulibacter was underrepresented in the contig-based taxonomic classification, even though FISH confirmed the high enrichment for Accumulibacter, which is in line with previous environmental

metagenomic studies [11, 59, 60]. The taxonomic classification might improve when using a less strict least common ancestor approach or when performing taxonomic classification on the contigs or even metagenome-assembled genome level [11, 12]. This would result in more clarity about the genes originating from a single organism.

Using the advanced approaches, we aimed to identify potential aggregate-forming proteins that could play a structural role in the biofilm matrix. Various biofilm components have been identified before to play a role in biofilm formation, stabilisation, and structure [1, 2, 14] and our approaches have identified various proteinaceous components. For example, we identified filamentous phage tails and other tube-forming proteins. Such proteins have been previously identified in *Pseudomonas aeruginosa* biofilms to promote organised and stable biofilms [61]. In addition, we identified pili proteins which are reported play a role in bacterial self-organisation in biofilms and stability of the biofilm [62]. In that same cluster, we found cytochrome C proteins, which may be stacked by their heme groups thereby forming pili-like structures called nanowires [63, 64]. The location of cytochrome C could be further examined by staining and visualisation. Interestingly, phage tail, pili and cytochrome C proteins have been reported to be more expressed in granular compared to floccular Accumulibacter-enriched cultures, suggesting that these filamentous structures indeed play a role in granule formation [26].

In addition to filamentous proteins, we identified multiple clusters that contained porins and other transmembrane beta-barrel proteins, which form membrane channels that allow the diffusion of small molecules. However, rising evidence indicates that these beta-barrel proteins can form functional amyloids [65–68], which play pivotal roles in biofilm formation and stabilisation [14, 69]. Our dataset might shine some light on this hypothesis as our supernatant approach is designed to enrich for nontransmembrane proteins over transmembrane proteins, suggesting that the identified beta-barrels could be functional amyloids. Interestingly, we indeed identified an amyloid curli protein among these beta-barrel proteins in the supernatant but not in whole granules. This curli protein originated from Flavobacterium [70], a genus that was 1.7-fold more present in the supernatant compared to the granules, which might explain why it was not identified in the whole granules. To further investigate the presence of amyloid beta-barrels in the supernatant, it would be interesting to stain the supernatant proteins for amyloid structures before and after digestion to correct for other amyloid structures that are resistant to denaturation and proteases and therefore not identified [13].

We also found proteins that have been previously associated with biofilm formation, although their function is less clear. For example, we identified proteins with PEP-CTERM motifs that are exclusively found in bacteria that have both an outer membrane and exopolysaccharide production genes [71]. Moreover, these proteins have been reported to play a role in floc formation in other bacteria found in wastewater treatment plants [72–74] and have been found overexpressed in granular compared to floccular Accumulibacter-enriched cultures [26]. Therefore, further unravelling the function of this motif may advance our understanding of the granule formation. In addition, we identified proteins with

polycystic kidney disease (PKD) and fibronectin 3 (fn3) domains which clustered together with proteins homologous to proteins with an S-layer domain. While all these proteins are linked to the cell surface [75], S-layer proteins have been found in other places in the biofilm, aiding in bacterial assembly and aggregation [76] and host fibronectin was reported to promote aggregation in *Staphylococcus aureus* [77, 78]. This might suggest that proteins with PKD and fn3 domains could also play a role in biofilm formation beyond cell adhesion.

For most clusters, at least one of the highly abundant proteins was attributed to Accumulibacter, confirming that Accumulibacter is a main contributor to the extracellular matrix of the granules. However, the clusters with shorter beta-barrels (Cluster 4) and cell surface proteins (Cluster 5) mainly contained proteins from other organisms. This suggests that proteins from multiple organisms may influence the granule formation and structure, where Accumulibacter mainly contributes various types of filamentous proteins and other community members contribute cell surface proteins and potentially amyloid-like proteins. Since the formation of dense granules is linked to Accumulibacter enrichment [25] and to increased filamentous protein production [26], filamentous protein production might be an important player in granule formation. The aggregating role of filamentous proteins like pili has been described in the formation of microcolonies [62], but the complex and interactive nature of biopolymers in the biofilm matrix complicates the characterisation of single protein types.

In conclusion, the two advanced metaproteome approaches successfully enriched the extracellular metaproteome from a *Ca*. Accumulibacter enrichment, which is crucial for understanding biofilm and granule formation, as well as for optimising or even controlling microbial communities in medical and industrial settings. Among many yet uncharacterised proteins we also identified several proteins with properties that were previously linked to granule or biofilm formation and biofilm structure. This data resource is a starting point to further unravel the influence of proteinaceous components on biofilm and granule formation.

### Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Data Availability Statement

The metaproteomic raw data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD057220.

## References

1. H. C. Flemming and J. Wingender, "The Biofilm Matrix," *Nature Reviews Microbiology* 8 (2010): 623–633.

2. T. Seviour, N. Derlon, M. S. Dueholm, et al., "Extracellular Polymeric Substances of Biofilms: Suffering From an Identity Crisis," *Water Research* 151 (2019): 1–7.

3. L. S. Zaramela, O. Moyne, M. Kumar, C. Zuniga, J. D. Tibocha-Bonilla, and K. Zengler, "The Sum Is Greater Than the Parts: Exploiting Microbial Communities to Achieve Complex Functions," *Current Opinion in Biotechnology* 67 (2021): 149–157.

4. M. Strous and C. Sharp, "Designer Microbiomes for Environmental, Energy and Health Biotechnology," *Current Opinion in Microbiology* 43 (2018): 117–123.

5. P. S. Li, S. Roos, H. Luo, B. Y. Ji, and J. Nielsen, "Metabolic Engineering of Human Gut Microbiome: Recent Developments and Future Perspectives," *Metabolic Engineering* 79 (2023): 1–13.

6. Y. L. Qian, F. Lan, and O. S. Venturelli, "Towards a Deeper Understanding of Microbial Communities: Integrating Experimental Data With Dynamic Models," *Current Opinion in Microbiology* 62 (2021): 84–92.

7. M. Kleiner, "Metaproteomics: Much More Than Measuring Gene Expression in Microbial Communities," *mSystems* 4 (2019): 6.

8. D. J. Nebauer, L. A. Pearson, and B. A. Neilan, "Critical Steps in an Environmental Metaproteomics Workflow," *Environmental Microbiology* 26 (2024): 24.

9. J. A. Blakeley-Ruiz and M. Kleiner, "Considerations for Constructing a Protein Sequence Database for Metaproteomics," *Computational and Structural Biotechnology Journal* 20 (2022): 937–952.

10. T. Muth, B. Y. Renard, and L. Martens, "Metaproteomic Data Analysis at a Glance: Advances in Computational Microbial Community Proteomics," *Expert Review of Proteomics* 13 (2016): 757–769.

11. H. B. C. Kleikamp, D. Grouzdev, P. Schaasberg, et al., "Metaproteomics, Metagenomics and 16S rRNA Sequencing Provide Different Perspectives on the Aerobic Granular Sludge Microbiome," *Water Research* 246 (2023): 120700.

12. H. B. C. Kleikamp, R. van der Zwaan, R. van Valderen, et al., "NovoLign: Metaproteomics by Sequence Alignment," *ISME Communications* 4 (2024): ycae121.

13. P. H. Nielsen, M. S. Dueholm, T. R. Thomsen, J. L. Nielsen, and D. Otzen, in: *Biofilm Highlights*, ed. H. C. Flemming, J. Wingender, U. Szewzyk (Springer, 2011): 41–62.

14. H. C. Flemming, E. D. van Hullebusch, T. R. Neu, et al., "The Biofilm Matrix: Multitasking in a Shared Space," *Nature Reviews Microbiology* 21 (2023): 70–86.

15. L. Hobley, C. Harkins, C. E. MacPhee, and N. R. Stanley-Wall, "Giving Structure to the Biofilm Matrix: An Overview of Individual Strategies and Emerging Common Themes," *FEMS Microbiology Review* 39 (2015): 649–669.

16. V. Vishwakarma, "Impact of Environmental Biofilms: Industrial Components and Its Remediation," *Journal of Basic Microbiology* 60 (2020): 198–206.

17. S. Sharma, J. Mohler, S. D. Mahajan, S. A. Schwartz, L. Bruggemann, and R. Aalinkeel, "Microbial Biofilm: A Review on Formation, Infection, Antibiotic Resistance, Control Measures, and Innovative Treatment," *Microorganisms* 11 (2023): 1614.

18. Y. V. Nancharaiah and G. K. K. Reddy, "Aerobic Granular Sludge Technology: Mechanisms of Granulation and Biotechnological Applications," *Bioresource Technology* 247 (2018): 1128–1143.

19. S. S. Adav, D. J. Lee, K. Y. Show, and J. H. Tay, "Aerobic Granular Sludge: Recent Advances," *Biotechnology Advances* 26 (2008): 411–423.

20. M. Pronk, M. K. de Kreuk, B. de Bruin, P. Kamminga, R. Kleerebezem, and M. C. M. van Loosdrecht, "Full Scale Performance of the Aerobic Granular Sludge Process for Sewage Treatment," *Water Research* 84 (2015): 207–217.

21. B. S. McSwain, R. L. Irvine, M. Hausner, and P. A. Wilderer, "Composition and Distribution of Extracellular Polymeric Substances in Aerobic Flocs and Granular Sludge," *Applied and environmental microbiology* 71 (2005): 1051–1057.

22. T. Seviour, Z. G. Yuan, M. C. M. van Loosdrecht, and Y. M. Lin, "Aerobic Sludge Granulation: A Tale of Two Polysaccharides?" *Water Research* 46 (2012): 4803–4813.

23. T. W. Seviour, L. K. Lambert, M. Pijuan, and Z. G. Yuan, "Selectively Inducing the Synthesis of a Key Structural Exopolysaccharide in Aerobic Granules by Enriching for *Candidatus* "*Competibacter phosphatis*"," *Applied Microbiology and Biotechnology* 92 (2011): 1297–1305.

24. H. B. Lu, A. Oehmen, B. Virdis, J. Keller, and Z. G. Yuan, "Obtaining Highly Enriched Cultures of *Candidatus Accumulibacter* Phosphates Through Alternating Carbon Sources," *Water Research* 40 (2006): 3838–3848.

25. J. J. Barr, A. E. Cook, and P. L. Bond, "Granule Formation Mechanisms Within an Aerobic Wastewater System for Phosphorus Removal," *Applied and Environmental Microbiology* 76 (2010): 7588–7597.

26. J. J. Barr, B. E. Dutilh, C. T. Skennerton, et al., "Metagenomic and Metaproteomic Analyses of *Accumulibacter phosphatis*-Enriched Floccular and Granular Biofilm," *Environmental Microbiology* 18 (2016): 273–287.

27. S. Tomás-Martínez, E. J. Zwolsman, F. Merlier, et al., "Turnover of the Extracellular Polymeric Matrix of Granules Performing Biological Phosphate Removal," *Applied Microbiology and Biotechnology* 107 (2023): 1997–2009.

28. S. Felz, S. Al-Zuhairy, O. A. Aarstad, M. C. M. van Loosdrecht, and Y. M. Lin, "Extraction of Structural Extracellular Polymeric Substances From Aerobic Granular Sludge," *Journal of Visualized Experiments* 115 (2016): 54534.

29. P. Zhang, Y. Shen, J. S. Guo, et al., "Extracellular Protein Analysis of Activated Sludge and Their Functions in Wastewater Treatment Plant by Shotgun Proteomics," *Scientific Reports* 5 (2015): 11.

30. C. Park, J. T. Novak, R. F. Helm, Y. O. Ahn, and A. Esen, "Evaluation of the Extracellular Proteins in Full-Scale Activated Sludges," *Water Research* 42 (2008): 3879–3889.

31. A. F. Silva, G. Carvalho, R. Soares, A. V. Coelho, and M. T. B. Crespo, "Step-by-Step Strategy for Protein Enrichment and Proteome Characterisation of Extracellular Polymeric Substances in Wastewater Treatment Systems," *Applied Microbiology and Biotechnology* 95 (2012): 767–776.

32. S. H. Hansen, A. Stensballe, P. H. Nielsen, and F. A. Herbst, "Metaproteomics: Evaluation of Protein Extraction From Activated Sludge," *Proteomics* 14 (2014): 2535–2539.

33. M. Albertsen, A. Stensballe, K. L. Nielsen, and P. H. Nielsen, "Digging Into the Extracellular Matrix of a Complex Microbial Community Using a Combined Metagenomic and Metaproteomic Approach," *Water Science and Technology* 67 (2013): 1650–1656.

34. A. Cydzik-Kwiatkowska, N. De Jonge, J. S. Poulsen, and J. L. Nielsen, "Unravelling Gradient Layers of Microbial Communities, Proteins, and Chemical Structure in Aerobic Granules," *Science of the Total Environment* 829 (2022): 154253.

35. A. Olaya-Abril, I. Jiménez-Munguía, L. Gómez-Gascón, and M. J. Rodríguez-Ortega, "Surfomics: Shaving Live Organisms for a Fast Proteomic Identification of Surface Proteins," *Journal of Proteomics* 97 (2014): 164–176.

36. L. G. Da Silva, K. O. Gamez, J. C. Gomes, et al., "Revealing the Metabolic Flexibility of "*Candidatus Accumulibacter phosphatis*" Through Redox Cofactor Analysis and Metabolic Network Modeling," *Applied and Environmental Microbiology* 86 (2020): 17.

37. G. J. F. Smolders, J. Van der Meij, M. C. M. Van Loosdrecht, and J. J. Heijnen, "Model of the Anaerobic Metabolism of the Biological Phosphorus Removal Process: Stoichiometry and pH Influence," *Biotechnology and Bioengineering* 43 (1994): 461–470.

38. J. P. Bassin, M. Pronk, G. Muyzer, R. Kleerebezem, M. Dezotti, and M. C. M. van Loosdrecht, "Effect of Elevated Salt Concentrations on the Aerobic Granular Sludge Process: Linking Microbial Activity With Microbial Community Structure," *Applied and environmental microbiology* 77 (2011): 7942–7953.

39. B. Schwanhäusser, D. Busse, N.a Li, et al., "Global Quantification of Mammalian Gene Expression Control," *Nature* 473 (2011): 337–342.

40. F. Trigodet, K. Lolans, E. Fogarty, et al., "High Molecular Weight DNA Extraction Strategies for Long-Read Sequencing of Complex Metagenomes," *Molecular Ecology Resources* 22 (2022): 1786–1802.

41. A. Stortchevoi and S. S. Levine, "Improved Yield of SPRI Beads–Based Size Selection in the Very High Molecular Weight Range," *Journal of Biomolecular Techniques* 35 (2024): 3fc1f5fe.d13e7666.

42. D. H. Li, C. M. Liu, R. B. Luo, K. Sadakane, and T. W. Lam, "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct *de Bruijn* Graph," *Bioinformatics* 31 (2015): 1674–1676.

43. W. Z. Li and A. Godzik, "Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences," *Bioinformatics* 22 (2006): 1658–1659.

44. D. E. Wood, J. Lu, and B. Langmead, "Improved Metagenomic Analysis With Kraken 2," *Genome Biology* 20 (2019): 13.

45. F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, et al., "SignalP 6.0 Predicts all Five Types of Signal Peptides Using Protein Language Models," *Nature Biotechnology* 40 (2022): 1023–1025.

46. J. Hallgren, K. D. Tsirigos, M. D. Pedersen, et al., "DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks," *BioRxiv* (2022): 2022-04.

47. N. Y. Yu, J. R. Wagner, M. R. Laird, et al., "PSORTb 3.0: Improved Protein Subcellular Localization Prediction With Refined Localization Subcategories and Predictive Capabilities for all Prokaryotes," *Bioinformatics* 26 (2010): 1608–1615.

48. J. Moreno, H. Nielsen, O. Winther, and F. Teufel, "Predicting the Subcellular Location of Prokaryotic Proteins With DeepLocPro," *Bioinformatics* 40 (2024): btae677.

49. V. Thumuluri, J. J. A. Armenteros, A. R. Johansen, H. Nielsen, and O. Winther, "DeepLoc 2.0: Multi-Label Subcellular Localization Prediction Using Protein Language Models," *Nucleic Acids Research* 50 (2022): W228–W234.

50. M. Kanehisa, Y. Sato, and K. Morishima, "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences," *Journal of Molecular Biology* 428 (2016): 726–731.

51. B. Buchfink, K. Reuter, and H. G. Drost, "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND," *Nature Methods* 18 (2021): 366–368.

52. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale," *Molecular Biology and Evolution* 38 (2021): 5825–5829.

53. J. Huerta-Cepas, D. Szklarczyk, D. Heller, et al., "eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses," *Nucleic Acids Research* 47 (2019): D309–D314.

54. M. H. Høie, E. N. Kiehl, B. Petersen, et al., "NetSurfP-3.0: Accurate and Fast Prediction of Protein Structural Features by Protein Language Models and Deep Learning," *Nucleic Acids Research* 50 (2022): W510–W515.

55. R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A Genomic Perspective on Protein Families," *Science* 278 (1997): 631–637.

56. M. Y. Galperin, Y. I. Wolf, K. S. Makarova, R. Vera Alvarez, D. Landsman, and E. V. Koonin, "COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens," *Nucleic Acids Research* 49 (2021): D274–D281.

57. J. Wingender, T. R. Neu, H.-C. Flemming, eds., *Microbial Extracellular Polymeric Substances: Characterization, Structure and Function* (Springer Berlin Heidelberg, 1999), 1–19.

58. A. Tanca, G. Biosa, D. Pagnozzi, M. F. Addis, and S. Uzzau, "Comparison of Detergent-Based Sample Preparation Workflows for LTQ-Orbitrap Analysis of the *Escherichia coli* Proteome," *Proteomics* 13 (2013): 2597–2607.

59. H. B. C. Kleikamp, M. Pronk, C. Tugui, et al., "Database-Independent *De Novo* Metaproteomics of Complex Microbial Communities," *Cell Systems* 12 (2021): 375–383.e5.

60. M. Stokholm-Bjerregaard, S. J. McIlroy, M. Nierychlo, S. M. Karst, M. Albertsen, and P. H. Nielsen, "A Critical Assessment of the Microorganisms Proposed to be Important to Enhanced Biological Phosphorus Removal in Full-Scale Wastewater Treatment Systems," *Frontiers in Microbiology* 8 (2017): 18.

61. P. R. Secor, J. M. Sweere, L. A. Michaels, et al., "Filamentous Bacteriophage Promote Biofilm Assembly and Function," *Cell Host & Microbe* 18 (2015): 549–559.

62. L. Craig, K. T. Forest, and B. Maier, "Type IV Pili: Dynamics, Biophysics and Functional Consequences," *Nature Reviews Microbiology* 17 (2019): 429–440.

63. Y. Gu, M. J. Guberman-Pfeffer, V. Srikanth, et al., "Structure of *Geobacter* Cytochrome OmcZ Identifies Mechanism of Nanowire Assembly and Conductivity," *Nature Microbiology* 8 (2023): 284–298.

64. F. Wang, Y. Gu, J. P. O'Brien, et al., "Structure of Microbial Nanowires Reveals Stacked Hemes That Transport Electrons Over Micrometers," *Cell* 177 (2019): 361–369.e10.

65. M. V. Belousov, A. O. Kosolapova, H. Fayoud, et al., "OmpC and OmpF Outer Membrane Proteins of *Escherichia coli* and *Salmonella Enterica* Form *Bona Fide* Amyloids," *International Journal of Molecular Sciences* 24 (2023): 15522.

66. E. J. Danoff and K. G. Fleming, "Aqueous, Unfolded OmpA Forms Amyloid-Like Fibrils Upon Self-Association," *PLoS ONE* 10 (2015): e0132301.

67. A. I. Sulatskaya, A. O. Kosolapova, A. G. Bobylev, et al., "$\beta$-Barrels and Amyloids: Structural Transitions, Biological Functions, and Pathogenesis," *International Journal of Molecular Sciences* 22 (2021): 11316.

68. E. Villain, A. A. Nikekhin, and A. V. Kajava, "Porins and Amyloids Are Coded by Similar Sequence Motifs," *Proteomics* 19 (2019): 6.

69. S. A. Levkovich, E. Gazit, and D. L. Bar-Yosef, "Two Decades of Studying Functional Amyloids in Microorganisms," *Trends in Microbiology* 29 (2021): 251–265.

70. M. S. Dueholm, M. Albertsen, D. Otzen, and P. H. Nielsen, "Curli Functional Amyloid Systems Are Phylogenetically Widespread and Display Large Diversity in Operon and Protein Structure," *PLoS ONE* 7 (2012): e51274.

71. D. H. Haft, I. T. Paulsen, N. Ward, and J. D. Selengut, "Exopolysaccharide-Associated Protein Sorting in Environmental Organisms: The PEP-CTERM/EpsH System. Application of a Novel Phylogenetic Profiling Heuristic," *BMC Biology* 4 (2006): 16.

72. N.a Gao, M. Xia, J. Dai, et al., "Both Widespread PEP-CTERM Proteins and Exopolysaccharides Are Required for Floc Formation of Zoogloea Resiniphila and Other Activated Sludge Bacteria," *Environmental Microbiology* 20 (2018): 1677–1692.

73. M. Xia, D. Yu, H. Chen, et al., "An RpoN-Dependent PEP-CTERM Gene Is Involved in Floc Formation of an *Aquincola Tertiaricarbonis* Strain," *BMC Microbiology* 23 (2023): 13.

74. W. An, F. Guo, Y. Song, et al., "Comparative Genomics Analyses on EPS Biosynthesis Genes Required for Floc Formation of *Zoogloea Resiniphila* and Other Activated Sludge Bacteria," *Water Research* 102 (2016): 494–504.

75. H. Jing, J. Takagi, J. H. Liu, et al., "Archaeal Surface Layer Proteins Contain $\beta$ Propeller, PKD, and $\beta$ Helix Domains and Are Related to Metazoan Cell Surface Proteins," *Structure (London, England)* 10 (2002): 1453–1464.

76. L. Li Wong, Y. Lu, J. C. S. Ho, et al., "Surface-Layer Protein Is a Public-Good Matrix Exopolymer for Microbial Community Organisation in Environmental Anammox Biofilms," *ISME Journal* 17 (2023): 803–812.

77. M. J. Pestrak, T. T. Gupta, D. H. Dusane, et al., "Investigation of Synovial Fluid Induced *Staphylococcus aureus* Aggregate Development and Its Impact on Surface Attachment and Biofilm Formation," *PLoS ONE* 15 (2020): 15.

78. S. Knott, D. Curry, N. Zhao, et al., "*Staphylococcus aureus* Floating Biofilm Formation and Phenotype in Synovial Fluid Depends on Albumin, Fibrinogen, and Hyaluronic Acid," *Frontiers in Microbiology* 12 (2021): 13.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.