Evolutionary plasticity of bacterial surface layer protein exoskeletons

Short title: Evolutionary plasticity of biological exoskeletons

Anna Barwinska-Sendra*[1] ✉, Paula S. Salgado[1] ✉, Kacper M. Sendra*[1] ✉

[1]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.

*These authors contributed equally to this work.

✉Authors to whom correspondence should be addressed:

Kacper M. Sendra, email: kacper.sendra@newcastle.ac.uk.

Anna Barwinska-Sendra, email: anna.barwinska-sendra@newcastle.ac.uk.

Paula S. Salgado, email: paula.salgado@newcastle.ac.uk.

Abstract

All cellular life possesses environmental interfaces like cell membranes or cell walls, yet the compositional complexity of these major cell components limits structural and evolutionary studies. Prokaryotic surface layer (S-layer) exoskeletons, with their homopolymer paracrystalline architecture, offer a more tractable evolutionary model of an environmental interface. In this study, we reveal the functional, structural and evolutionary diversity of S-layers in Gram-positive *Peptostreptococcaceae*, including pathogens and cancer-promoting species. We uncover novel S-layer architectures with diverse biochemical and physiological properties, enabled by a modular design co-evolving with other cell envelope components. We elucidate the mechanisms and evolutionary pathways underpinning the emergence of novel S-layers and the diversification of existing ones. Our findings establish the S-layer as a paradigm of cellular and molecular modularity and evolutionary plasticity. The demonstrated adaptability of these biological exoskeletons enables rapid reconfiguration of bacterial cell surface architecture and physiology, facilitating immune evasion in pathogens.

Introduction

The cell wall and cell membrane usually constitute the core components of the bacterial cell envelope, the interface between a bacterial cell and its environment. The layered architecture of the cell envelope can be further augmented by the presence of an outer cell surface-layer (S-layer) exoskeleton. S-layers are found across Archaea, Gram-positive, and Gram-negative Bacteria(*1-4*), are visible using high-resolution imaging techniques, and their building blocks are some of the most abundant cellular proteins. The S-layers studied so far are generally composed of a major protein that assembles into the paracrystalline lattice, and S-layer-associated proteins that confer additional functionality(*5*), as observed in pathogens such as *Bacillus anthracis*(*6*) and *Clostridioides difficile*(*7*). Proposed physiological roles of S-layers and S-layer proteins include protection against cell wall degradation and osmotic stress(*8-11*), biofilm formation(*12-14*), adhesion(*15-17*), phage recognition(*18, 19*), and/or function as a molecular sieve(*20, 21*). Understanding the biology of S-layers is significant within biotechnology, with potential applications as scaffolds in bio-product design(*22, 23*), drug nanocarrier(*24*) or vaccine delivery system(*25*), and within healthcare, since they have been implicated in immunomodulation by microbiota important for optimal health(*26-29*) and in the infection process of different pathogens(*30-32*), and are a target for the development of novel therapeutics(*32, 33*).

Although the diversity of S-layers across different, distantly related bacterial taxa is well established(*2, 34, 35*), general evolutionary mechanisms shaping this diversity across the Tree of Life remain unexplored. S-layer proteins, including those sharing the same structural fold, often lack any identifiable sequence homology(*2*). This complicates studies of S-layer origins and evolution and provides a major challenge for the identification of S-layer proteins and prediction of conserved S-layer functions such as protection against the human immune system. Recent accomplishments in structure-based predictive algorithms provide major advancements in our capacity to predict S-layer protein candidates and S-layer architectures(*34, 35*). However, due to their extreme sequence variation and low number of experimentally verified and studied S-layers, our understanding of functional diversity, ecological significance, and evolution of this important element of prokaryotic cell biology is still limited.

The evolutionary origins of these paracrystalline layers surrounding entire prokaryotic cells, including that of many pathogens, are also not clear. It is commonly hypothesised that S-layers have originated multiple independent times(*34, 36*), mainly based on the sequence and structural diversity of S-layer proteins and lattices(*34*). However, the observed variation of surface exoskeletons could, at least partially, also be explained by an alternative hypothesis of a common origin followed by an extreme divergence and lateral transfer of S-layer proteins. In pathogenic *C. difficile*, the S-layer protein (*Cd*SlpA), encoded within

60 the genomic S-layer cassette, is shared across *C. difficile* isolates *via* homologous recombination(*37*), and its
61 highly variable sequence enables its use for typing (S-layer cassette type, SLCT) of different clinical
62 isolates(*37, 38*). A homologue of *Cd*SlpA, *Peptostreptococcus anaerobius* surface protein PCWBR2 was
63 recently implicated in promoting colorectal carcinogenesis(*39*). Despite their medical significance, our
64 understanding of the structural and functional landscape of S-layers formed by these diverse orthologues
65 from closely related species is limited and current experimental structural information is available for only
66 two *C. difficile* sequence types(*40*).
67 Here, to address these major gaps in our understanding of S-layer biology and to circumvent the
68 inherent challenges in studying these cellular structures, we focus on narrower phylogenetic scales, where
69 S-layer evolution is more tractable. We experimentally characterise novel S-layer types and present
70 evidence of independent S-layer losses and gains across a bacterial family *Peptostreptococcaceae*. We
71 reveal the role of modular evolution in shaping S-layer diversity and enabling anchoring to variable cell wall
72 architectures. We present novel structures of *C. difficile* S-layers and uncover how, at the narrowest
73 phylogenetic scales, extreme sequence divergence, recombination, and adaptive evolution drive extensive
74 structural changes of the crystalline lattice surrounding the bacterial cells, without compromising its role in
75 protection against the human immune system. Physiological tests and CRISPRi experiments demonstrate
76 that the resistance to the innate immunity effectors emerged specifically in pathogenic lineages. Our
77 results provide insights into the mechanisms of adaptation of host-exposed cell surfaces of S-layer-
78 containing pathogens, and the evolution of the environmental interfaces across the bacterial Tree of Life.

Results

The evolutionary and structural diversity of S-layers across the Tree of Life

It is often hypothesized that S-layers are common in Prokaryotes(*1-3*). To test this hypothesis, we compiled a list of previously studied examples of S-layer-positive species and S-layer proteins with strong experimental evidence identified in the available literature, including verification using high-resolution imaging and molecular techniques (Supplementary Data 1). Mapping this evidence onto the Tree of Life (Fig. 1a) revealed that, although verified S-layer-positive species are distributed across many lineages, they represent only a small subset of the analysed Bacteria and Archaea. This reflects the disparity between the number of known species of Prokaryotes and those that have been cultured(*41*), of which only a small number have been tested for the presence of S-layers. Our analysis highlights the extensive gaps in the current understanding of S-layer diversity and distribution.

To address these gaps, we investigated the distribution of homologues of the known full-length S-layer protein sequences and their constituent domains across the Tree of Life using protein sequence homology searches (Supplementary Fig. 1), hidden Markov model (HMM) profile searches (Supplementary Fig. 2), and protein similarity network analyses (Supplementary Fig. 3). The main building block of the paracrystalline S-layer array, here referred to as Surface Layer Lattice Protein (SLLP), usually displays a bipartite architecture(*2*). It consists of a cell anchoring domain ($anchor_{DOM}$), which often represents a general anchoring mechanism of cell surface proteins, and an S-layer 'crystallisation' or surface-forming domain ($surface_{DOM}$), which plays a more specific role in forming the paracrystalline lattice layer exposed to the extracellular environment (Fig. 1b and Supplementary Fig. 1a). Anchoring domains such as SLH have additional physiological roles, including outer membrane tethering in diderm Negativicutes(*42*), and in our analysis, are found across a broad taxonomic range, often in multiple proteins other than SLLPs from the same genome (Supplementary Fig. 3g). In addition, multiple different types of $anchor_{DOM}$s are often found within a single cell-anchored protein family (Supplementary Fig. 3b, d), indicating that they can be swapped during cell-anchored protein evolution. Therefore, $anchor_{DOM}$s are not specific to SLLPs and constitute swappable protein modules that enable the binding of surface proteins to alternative cell envelope architectures. This likely reflects the structural co-evolution of different cell-envelope layers, such as the cell wall and the S-layer.

In contrast to the $anchor_{DOM}$s, specific surface domain sequence profiles are less common and display narrower taxonomic distribution, suggesting lineage-specific adaptations and more specific functions in S-layer formation (Fig. 1a and Supplementary Fig. 3c). The majority of the analysed $surface_{DOM}$s, including those sharing the common immunoglobulin-like domain architecture (Fig. 1b, and Supplementary Fig. 1a), display extensive sequence diversity. The substantial discontinuity in the structural and sequence landscape of bacterial S-layers is further apparent in SLLPs sharing similar domain folds (Fig. 1b, and Supplementary Fig. 1a) but no sequence homology (Supplementary Figs. 1c, 2a, 3b) like bacterial *B. anthracis* EA1 and *Deinococcus radiodurans* HPI, or archaeal *Haloferax volcanii* Csg, which assemble into extremely diverse paracrystalline layers (Fig. 1b) and utilise different cell surface anchoring mechanisms (Fig. 1b, and Supplementary Fig. 1a).

Combined with the identified gaps in the appreciation of S-layer diversity and distribution, and the possible existence of SLLPs with distinct $surface_{DOM}$ folds, such as *Tannerella forsythia* TfsA (Supplementary Fig. 1a), our analyses are consistent with the existence of undiscovered diversity of prokaryotic exoskeletons in nature. However, the lack of sequence homology often observed between SLLPs belonging to the same structural family and the utilisation of common structural folds complicate analyses of their evolution over wide phylogenetic scales. Given the structural and sequence diversity of surface-forming domains and the multitude of anchoring mechanisms, it is likely that the various S-layers have either

emerged multiple independent times from non-SLLPs and/or are a result of extreme surface$_{DOM}$ divergence and anchor$_{DOM}$ swapping among existing S-layer proteins.

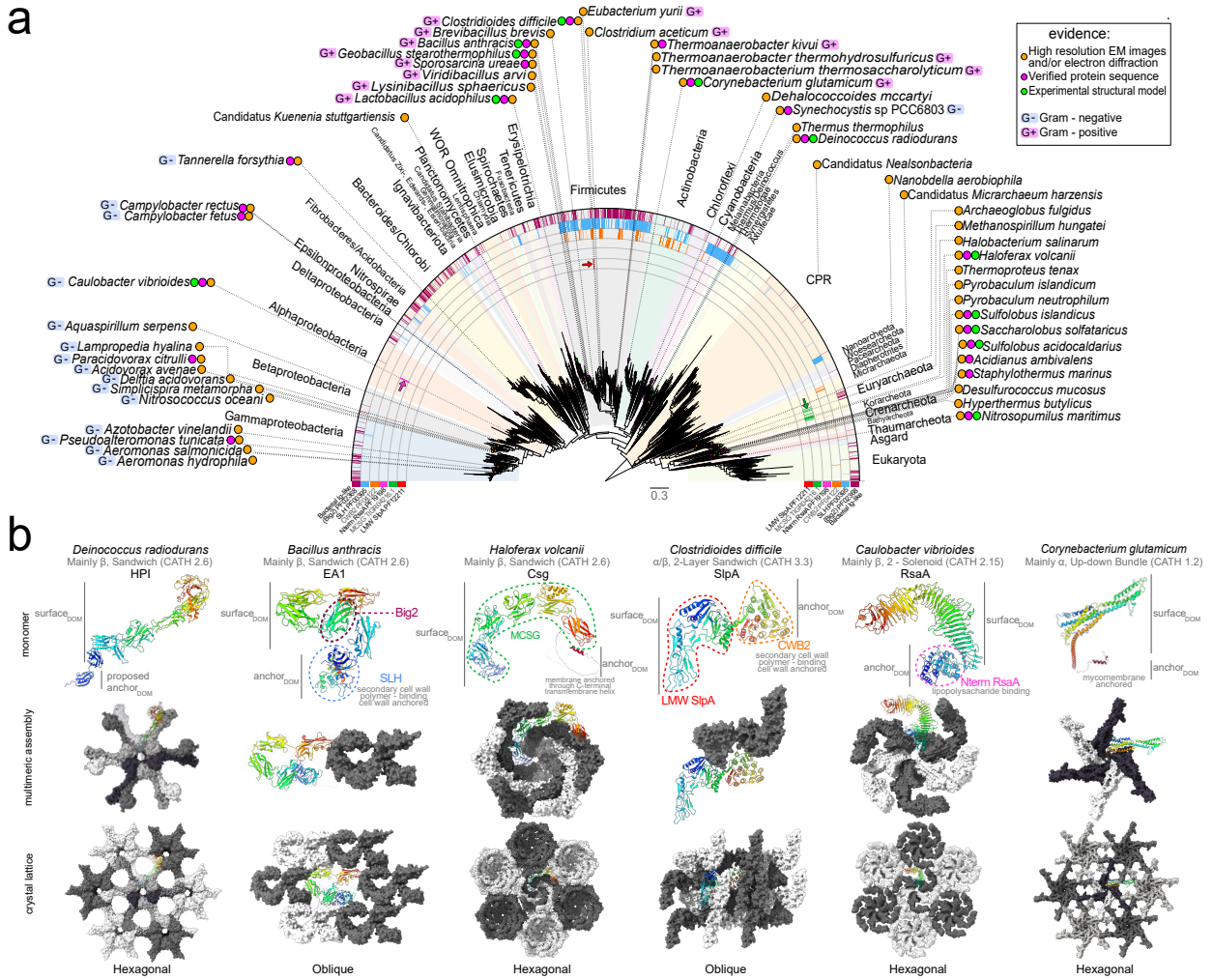

Fig. 1: Diversity of S-layer protein lattices across the Tree of Life.

a, Species names of representative 47 Bacteria and 21 Archaea with previously experimentally verified S-layers (Supplementary Data 1) were annotated onto a phylogeny of 2,705 Bacteria, 315 Archaea, and 149 Eukaryota. Coloured circles represent the available experimental evidence including verification *via* high-resolution imaging techniques (orange), structural biology (green), and protein sequence verification of SLLP (magenta). The coloured semi-circles represent hmmsearch results using HMM profiles of domains found in SLLPs with known structures presented in **b**, demonstrating the presence of both common (e.g. Big2 and SLH), and species- or lineage-specific domain profiles (coloured arrows, e.g. LMW SlpA, SLP$_L$).
b, Cartoon representations of structural models of SLLPs (monomer) and surface representations of their higher order organisation (multimeric assembly and crystal lattice) within the S-layer. The sequence regions (dashed outlines) identified with the available HMM profiles (**a** and Supplementary Fig. 2) can be found within two distinct domains, mainly involved in either cell surface anchoring (anchor$_{DOM}$) or paracrystalline S-layer lattice formation (surface$_{DOM}$). The experimental surface$_{DOM}$ structures (Supplementary Fig. 1a) belong to at least four types of structural folds identified using Foldseek (CATH: 2.6, 3.3, 2.15, 1.2). Multiple monomers (shades of grey) within the multimeric assemblies assume diverse higher-order geometries, which constitute a repeating unit within the crystal lattice with varying degrees of complexity and symmetry types (e.g. hexagonal and oblique). These diverse S-layer architectures are attached to the cell surfaces using a variety of mechanisms including membrane anchoring through transmembrane helices, and binding to different secondary cell wall polymers (SLH and CWB2) or lipopolysaccharides (Nterm RsaA).
Structures of the previously determined representative S-layer lattice proteins (8CKA, 8OPR, 7PTR, 7ACY, 6Z7P, 9GK2) were visualised using ChimeraX.

149 **Novel SLLPs form structurally diverse S-layers within a single family of Bacteria**

150 To experimentally investigate the diversity of the S-layer exoskeletons at narrower phylogenetic scales,
151 we first identified the lineages with a high degree of variation in the known SLLP domains. Our analyses
152 across the Tree of Life (Fig. 1a) revealed an interspersed distribution of different anchoring domains – SLH
153 and CWB2 – in Firmicutes. This bacterial phylum includes two important pathogens*, Bacillus anthracis* and
154 *Clostridioides difficile*, with two distinct S-layer surface domain folds (Fig. 1b). Within Firmicutes,
155 *Peptostreptococcaceae* represent a particular 'hotspot' of SLH/CWB2 anchoring domain repertoire
156 variation. Therefore, this lineage was selected for combinatorial bioinformatic, cell biology, microbiological,
157 and structural biology analyses to characterise their S-layers and SLLPs and unravel evolutionary
158 mechanisms shaping their diversity.

159 The experimental investigations (Fig. 2a and d, and Supplementary Fig. 4a) revealed four
160 *Peptostreptococcaceae* (29%, 4/14 species) with no apparent S-layer in whole cell electron micrographs, no
161 'ghosts' in S-layer electron microscopy (EM) preparations, and no abundant protein bands (Supplementary
162 Fig. 4a), typically observed for SLLPs, in surface protein preparations from S-layer-positive species. In three
163 out of four S-layer-negative species, the identified instances of apparent S-layer absence had scattered
164 distribution across the phylogeny, and no identifiable genomic SLLP homologue sequences, consistent with
165 multiple independent losses of S-layer (Fig. 2d). The existence of multiple S-layer-negative lineages of
166 Bacteria could lend some support to the hypothesis that S-layers have evolved multiple independent
167 times(*2, 34*). However, the presence of life cycle-specific or environmentally induced S-layers, or S-layer
168 loss in laboratory conditions, cannot be fully excluded. This is apparent in previously described difficulties in
169 the detection of S-layers in some bacteria(*43*) and, in our dataset, in the single case of *Romboutsia hominis*
170 RFIFI, which contains an SLLP homologue but displays an S-layer-negative morphotype (Fig. 2a and d, and
171 Supplementary Fig. 4). Therefore, currently available data suggests the possibility of independent S-layer
172 losses within *Peptostreptococcaceae*, although these represented the minority of the studied species.

173 In the majority (71%, 10/14 species) of the studied *Peptostreptococcaceae*, we found evidence of S-
174 layers and SLLPs (Fig. 2d; and Supplementary Fig. 4a). AlphaFold2 structural predictions and Foldseek
175 analyses revealed that six of the identified SLLPs (60%, 6/10 SLLPs) contained surface domains homologous
176 to the α/β-sandwich structural fold (Fig. 2b, c), previously only found in *C. difficile* SLLP (*Cd*SlpA). This type
177 of SLLP, with only two or three surface-exposed subdomains (Fig. 2c), features a distinctively shorter
178 $surface_{DOM}$ than other identified SLLP types and can potentially assemble into a more compact S-layer
179 lattice, as observed in *C. difficile*. Another distinction of this SLLP type is the higher proportion of α-helices
180 within its $surface_{DOM}$(*40*), as opposed to predominantly β-strand $surface_{DOM}$s of S-layers commonly found
181 across the Tree of Life (Fig. 1b). Four of these shorter SLLPs were found in close relatives of *C. difficile*,
182 including *Asaccharospora irregularis* and *Clostridioides mangenotii*, suggesting that this type of S-layer was
183 present in their last common ancestor. Interestingly, the SLLP from *Peptostreptococcus anaerobius*
184 contained the same fold, and our experiments identified it as PCWBR2 (Supplementary Fig. 4a), a protein
185 implicated in promoting colorectal carcinogenesis(*39*).

186 Three other identified *Peptostreptococcaceae* SLLPs (30%, 3/10) contained β-sandwich $surface_{DOM}$s,
187 with two from the most basal *Peptostreptococcaceae* outgroup lineages (Fig. 2a-d and Supplementary Fig.
188 4). Interestingly, while α/β-sandwich $surface_{DOM}$ shared enough sequence homology to group weakly within
189 a single protein similarity network, the β-sandwich fold SLLPs grouped within two distinct networks
190 (Supplementary Fig. 4b). The more broadly distributed β-sandwich S-layers were found within the outgroup
191 lineages including *Acetoanaerobium sticklandii* and *Peptoanaerobacter stomatis*. Their broader distribution
192 across the species tree suggested more ancient S-layer acquisition, contrasted by the second type of β-
193 sandwich S-layers, found only in a narrow subsample of closely related *Intestinibacter bartlettii* genome
194 assemblies (Fig. 2e). Inspection of all available *I. bartlettii* genomes revealed a recent SLLP acquisition likely

195 *via* transposition, indicated by its co-distribution with an adjacent IS256 gene (Fig. 2e) encoding a
196 transposase homologue. Furthermore, structural and sequence comparisons of the *I. bartelettii* SLLP
197 revealed that it has evolved *via* multiple tandem duplications of a single common β-sandwich
198 Immunoglobulin-like (PFAM ID, PF16403) surface subdomain (Fig. 2f). At this stage, we were able to
199 sequence-verify SLLP from the single *I. bartelettii* isolate available in culture collections. The significance of
200 subdomain number variation observed across available *I. bartlettii* genome assemblies remains unclear, as
201 SLLP-coding sequences are often incomplete and found at the end of a contig. Nevertheless, all six
202 $surface_{DOM}$ subdomains of the verified sequence shared, on average, 96 % of pairwise nucleotide sequence
203 identity, consistent with the recent emergence of this β-sandwich SLLP type.
204 The single SLLP from *Peptoclostridium acidaminophilum* contained the third distinct type of S-layer
205 surface-forming domain, β-barrel (10%, 1/10 SLLPs). Its similarity to the β-sandwich fold suggests that it
206 may have evolved either from a shared β-sandwich SLLP ancestor or originated independently from a non-
207 SLLP β-barrel-containing protein. The β-barrel SLLP contained the largest $surface_{DOM}$ among the identified
208 *Peptrostreptococcaece* SLLPs, with an extended 10-subdomain architecture, previously also found in the β-
209 sandwich S-layer of ammonia-oxidizing marine archaea(*21*). The available crystal structures or calculated
210 SymProFold S-layer lattice models indicate that the β-barrel SLLP can assemble into an hexagonal lattice, in
211 contrast to the oblique lattice determined in *C. difficile* α/β-sandwich S-layer or the predicted square
212 symmetry of the *A. sticklandii* β-sandwich S-layer model (Fig. 2c).
213 By combining bioinformatics and experimental investigations, we uncovered novel types of SLLPs and S-
214 layers, recent SLLP emergence and S-layer acquisition, and multiple independent SLLP losses in
215 *Peptostreptococcaceae*. The revealed structural diversity of *Peptostreptococcaceae* SLLPs, including
216 multiple distinct S-layer surface-forming folds assembled into multiple distinct S-layer architectures, reflects
217 what we observed in our analyses across the Tree of Life (Fig. 1). Therefore, the presence of diverse S-layer
218 architectures utilising different surface domain folds does not only represent distinct features of distantly
219 related bacterial phyla but can be generated within much shorter evolutionary time frames. The emergence
220 of a novel SLLP-type *via* tandem subdomain duplication in *I. bartelettii* provides an insight into the
221 mechanism likely involved in the origins of novel S-layer architectures.
222

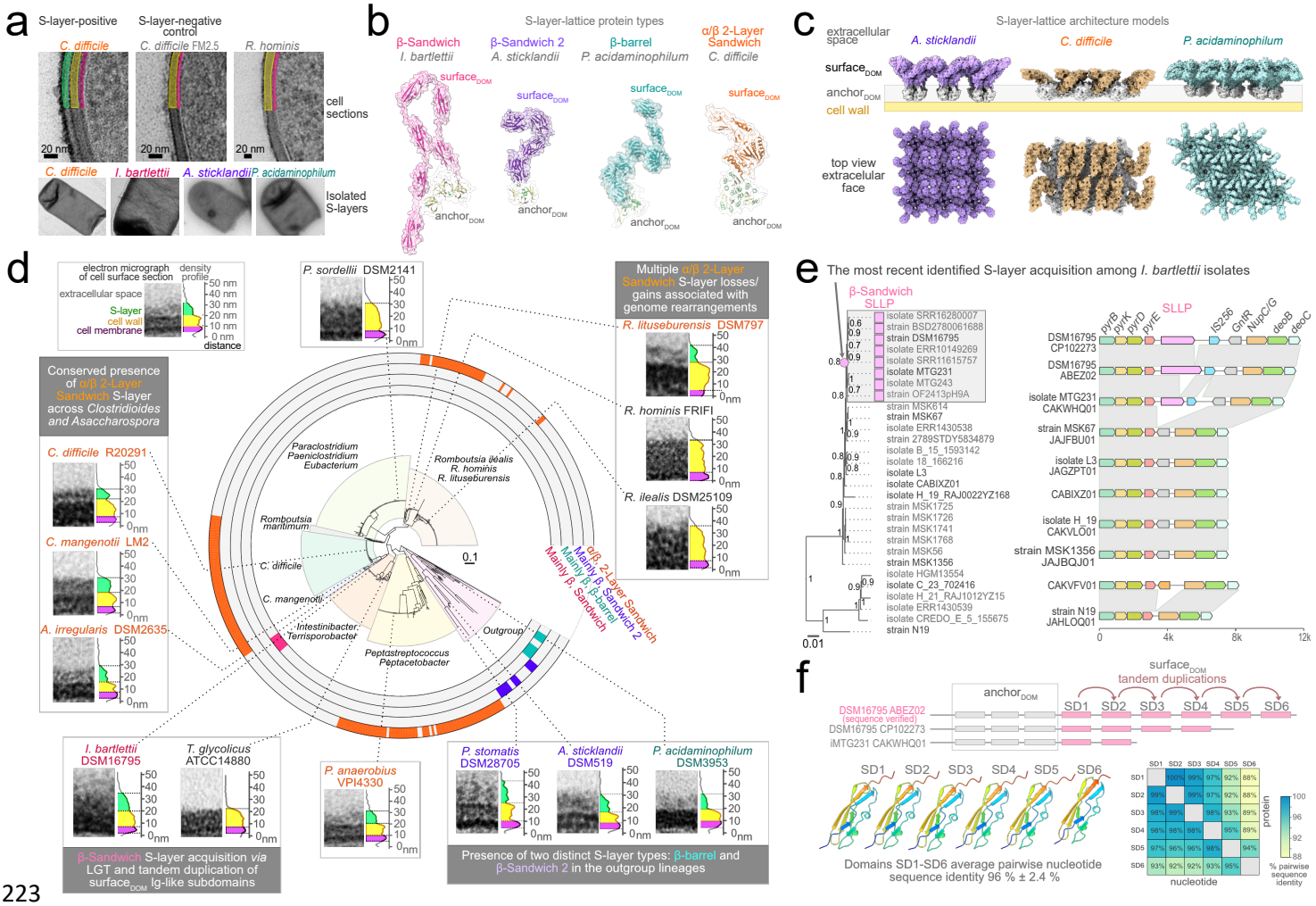

Fig. 2: Identification of new types of S-layer and mechanism of emergence of novel S-layer protein.

a, Representative electron micrographs of uranyl acetate and ruthenium red stained vegetative cells (top panel) of S-layer-positive *C. difficile,* S-layer-negative control of *C. difficile* FM2.5, and *R. hominis* with no apparent S-layer. The coloured regions correspond to S-layer (green), cell wall (yellow), and cell membrane (pink). S-layers (bottom panel) can be stripped from the vegetative cells due to the structural stability of their paracrystalline lattice.

b, Determined crystal structure (*C. difficile)* and AlphaFold2 predicted structural models of the S-layer lattice proteins (SLLPs) identified by mass spectrometry. The experimentally verified *Peptostreptococcaceae* SLLPs belonged to one of the four distinct surface domain ($surface_{DOM}$) folds.

c, Surface representations of S-layer models, including top-scoring SymProFold predictions (*A. sticklandii, P. acidaminophilum*), demonstrate that different *Peptostreptococcaceae* SLLP $surface_{DOM}$s (coloured) can assemble into distinct lattice architectures on the bacterial cell surface.

d, Species tree mapping of the *Peptostreptococcaceae* experimentally tested for the presence of S-layers and SLLPs. Density profiles next to the electron micrographs of the cell envelopes were colour-coded to indicate the location of the S-layer (green), cell wall (yellow), and cell membrane (pink). The scale bars measure distance (in nm) from the bottom of the micrograph. Species names were coloured based on the structural classification (**b**) of the verified SLLP $surface_{DOM}$ of *Peptostreptococcaceae* with apparent S-layer or those with no evidence of S-layer (black). Coloured semi-circles represent the distribution of the protein homologues of each SLLP-type identified in 453 *Peptostreptococcaceae* genomes.

e, The presence of the β-sandwich SLLP (pink rectangles) was mapped onto the species tree of *I. bartlettii* isolates (left panel). The recent acquisition of the SLLP (pink circle) is apparent in genomic region alignment (right panel), indicating ancestral insertion of the SLLP (pink) and IS256 transposase (blue) coding genes within a broadly conserved nucleotide metabolism locus.

f, Domain architecture of the *I. bartlettii* SLLPs (top panel) demonstrates a possible variation in the number of the β-sandwich $surface_{DOM}$ subdomains (SD1-6) among different isolates. All six SD1-6 (bottom panel, left) share high levels of pairwise protein and nucleotide sequence identity (bottom panel, right) consistent with SLLP origin *via* multiple tandem subdomain duplications.

248 **Homologous S-layers can impart different resistance to antimicrobial agents**

249 Although the structural and sequence diversity of S-layers at both wider and narrower taxonomic scales
250 is clear, the conservation of their physiological roles remains elusive. Furthermore, the function of S-layers
251 in many organisms has not been experimentally investigated(*2*). S-layers of bacteria such as that of gut
252 pathogen *C. difficile* protect cells from natural antimicrobial agents produced by the host immune system
253 such as lysozyme or human cathelicidin peptide LL-37(*10*). However, it is not known if this is also true for
254 other S-layers we identified across *Peptostreptococcaceae*.

255 Our initial experimental investigations demonstrated that *Peptostreptococcaceae* S-layers display
256 variable susceptibility to extraction using different chemical methods (Fig. 3a), indicating that at least some
257 of the properties of the outermost cell envelope layer of these bacteria are not conserved. Surprisingly, this
258 variation was not only observed for the S-layers made of different types of SLLPs but also across
259 homologous S-layers formed by the α/β-sandwich SLLPs, including that of *C. difficile* (Fig. 3a). Protein
260 similarity network analyses (Fig. 3b) and manual inspection of multiple sequence alignments revealed
261 extensive sequence divergence within this SLLP-fold. This is particularly apparent in the case of S-layer
262 $surface_{DOM}$s from different isolates of *C. mangenotii* (LM2 and DMS29133), which are not directly
263 connected with each other or with sequences from their closest relatives (*C. difficile* and *A. irregularis*)
264 within the network (Fig. 3b).

265 Structural comparisons of the experimentally verified α/β-sandwich SLLPs (Fig. 3c) and homologous S-
266 layer-associated proteins Cwp2 and Cwp8, revealed that, despite their extensive sequence variation, the
267 overall fold architecture remains conserved. The general α/β-sandwich SLLP $surface_{DOM}$ consists of three
268 subdomains, equivalent to *C. difficile* SlpA D1, D2 and ID(*40*) and S-layer associated Cwp8 D1, D2, D3(*44*),
269 which are connected via flexible linkers, providing further structural variability as rotation of the relative
270 orientation of the subdomains leads to varying surfaces (Fig. 3b,c). The AlphaFold3(*45*) predictions for the
271 experimentally verified SLLPs fold with three α/β-sandwich subdomains follows the topology seen in
272 *Cdl*SlpA, where D1 leads to the loop-rich D2. Insertion of a β-strand into D1 then links to the third
273 subdomain (ID), which connects to the CWB2 $anchor_{DOM}$. The most structurally conserved subdomain in
274 $surface_{DOM}$ is D1, with an average RMSD of 1.7 Å across all SLLPs, while the least conserved is D2 (average
275 RMSD 3.7 Å). The most distinct modification to the general α/β-sandwich architecture was observed in *A.*
276 *irregularis* DSM2635 where an additional, elongated α-helical region (~80 Å long) has been acquired and
277 inserted within D2 (Fig. 3c). These additional helices likely correspond to the protrusions extending out of
278 the lattice towards the extracellular environment observed in the cryo-electron micrographs of the *A.*
279 *irregularis* S-layer preparations (Fig. 3b). Importantly, acquisition of novel types of $surface_{DOM}$ subdomains
280 provides another mechanism of S-layer lattice evolution, alongside variation in the number of subdomains
281 as identified in *I. bartlettii* (Fig. 2f).

282 Next, we tested if S-layer dependent resistance to innate immune effectors observed in *C. difficile* is
283 conserved in other *Peptostreptococcaceae* containing α/β-sandwich S-layers. Physiological growth tests in
284 the presence of lysozyme and LL-37 revealed variable resistance to these natural antimicrobials (Fig. 3d).
285 Consistent with previous investigation(*10*), S-layer-positive *C. difficile* R20291, but not S-layer-negative *C.*
286 *difficile* FM2.5, displayed resistance against LL-37 and lysozyme. Interestingly, similar resistance was also
287 observed in cancer-promoting pathogen(*39*) *P. anaerobius* but not in any of the other α/β-sandwich S-layer
288 *Peptostreptococcaceae*, including closest *C. difficile* relatives *A. irregularis* and *C. mangenotii* (Fig. 3d). The
289 presented data is consistent with resistance to these antimicrobials representing specific and potentially
290 independent adaptations of *C. difficile* and *P. anaerobius* cell envelopes to the gut environment.

291 To further test the variable resistance of α/β-sandwich S-layers to antimicrobial agents, and to
292 experimentally verify that the identified SLLPs are essential for S-layer formation, we performed CRISPRi
293 gene knock-down experiments in *C. difficile* (Fig. 3e) and *R. lituseburensis* (Fig. 3f), two more distantly

294 related *Peptostreptococcaceae* species but with SLLPs sharing a close structural relationship, according to
295 Dali analyses (Fig. 3c). Consistent with the experiments in the wild type-strains (Fig. 3d), knock-down of the
296 SLLP gene abolishes *C. difficile* resistance to lysozyme and LL-37 (Fig. 3e), as opposed to *R. lituseburensis*,
297 where the presence of SLLP provided no apparent protection against either (Fig. 3f).
298 Previous experiments have also demonstrated that the presence of S-layer plays a role in *C. difficile*
299 sporulation(*10*). In our experiments, CRISPRi S-layer knock-down strains of *C. difficile* and *R. lituseburensis*
300 exhibited sporulation deficiencies in the tested conditions (Figure 3e, f, Supplementary Table 1), suggesting
301 that S-layer integrity plays a conserved role necessary for sporulation of these two α/β-sandwich S-layer
302 containing *Peptostreptococcaceae*. In addition, both mutants displayed a propensity to aggregation in liquid
303 culture, highlighting the conserved role of their S-layers in its prevention.
304 In summary, despite their structural conservation, homologous α/β-sandwich S-layers from closely
305 related species display a high degree of biochemical and sequence diversity and differ in their contribution
306 to the protection of the cells they surround against natural antimicrobials lysozyme and LL-37. Despite
307 these differences, they also play conserved roles in cellular processes such as preventing cell aggregation
308 and sporulation, important in bacterial survival and infection. Therefore, not all phenotypic effects
309 associated with the presence of homologous S-layers are conserved. Certain functional features could
310 reflect specific ecological niches, such as the gut environment for *C. difficile* and *P. anaerobius*, while more
311 general life cycle adaptations like sporulation and cell-cell interactions might be under broader selective
312 pressure.

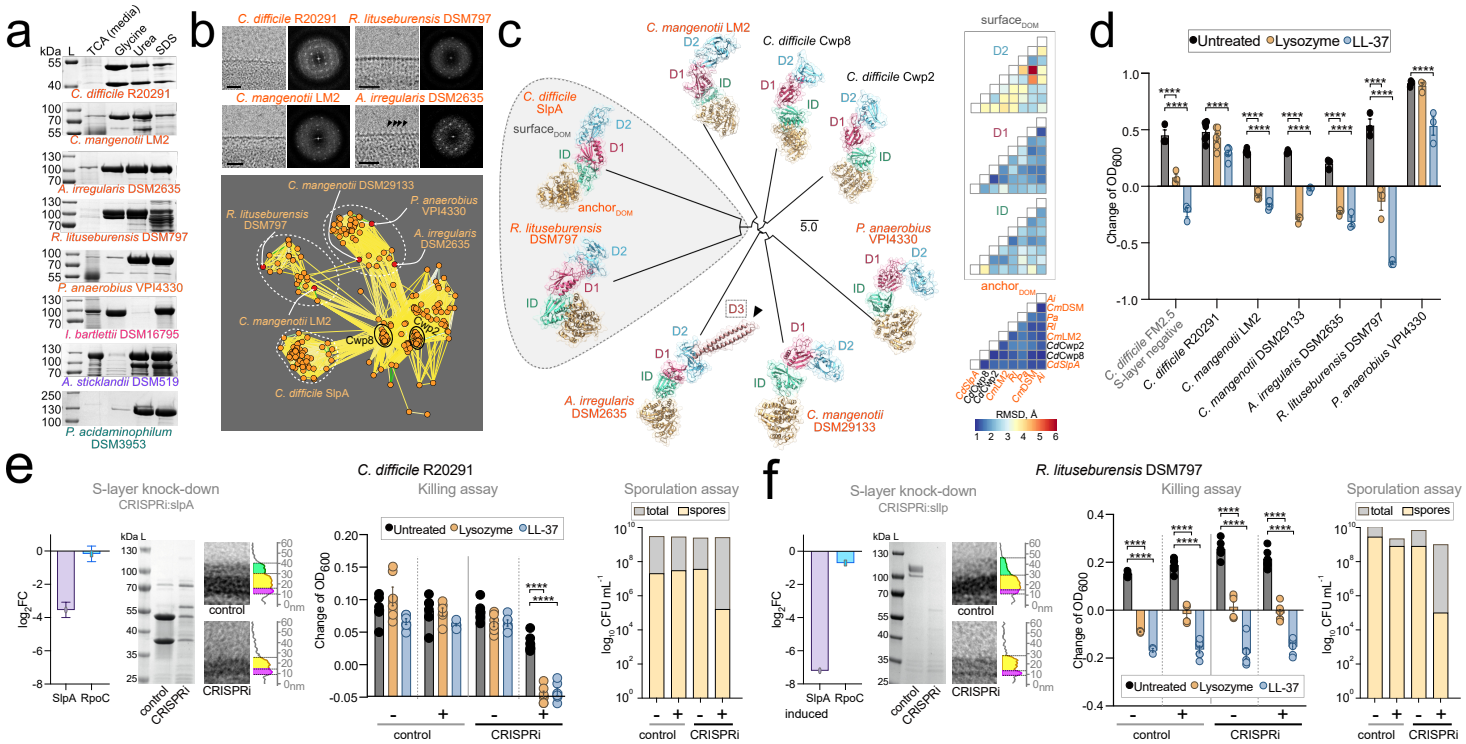

Fig. 3: Variable biochemical and physiological properties within a single structural type of S-layer.

a, SDS-polyacrylamide gels of *Peptostreptococcaceae* low pH glycine, urea, and SDS S-layer preparations. The α/β-sandwich S-layers (orange) displayed contrasting sensitivity to extraction in SDS and glycine. β-sandwich S-layers (pink or purple) varied when extracted in glycine and urea. Trichloroacetic acid (TCA) precipitation determined the SLLP presence in the medium. Molecular weight ladder (L).

b, Cryo-EM micrographs and Fast Fourier Transform (FFT) diffraction patterns of the α/β-sandwich S-layers (top panel). Additional protrusions were apparent at the extracellular face of *A. irregularis* S-layer (black arrowheads).

Protein sequence similarity network (bottom panel) of the α/β-sandwich SLLP surface$_{DOM}$s homologues (orange circles) including S-layer-associated Cwp2 and Cwp8 from the available *Peptostreptococcaceae* genomes. Highly divergent sequences of the verified S-layer surface$_{DOM}$s (red circles) formed three distinct groups (dashed outlines).

c, The Alphafold3 structural models of identified *Peptostreptococcaceae* α/β-sandwich SLLPs and Cwp2/8 were mapped onto their Dali surface$_{DOM}$s structural comparison dendrogram. Two additional α-helices of *A. irregularis* (dashed rectangle), likely correspond to the S-layer extracellular protrusions (**b**). Heatmap matrices for RMSD values across subdomains D1, D2, ID and anchor$_{DOM}$ show the highest structural divergence within the D2 (average RMSD 3.7 Å). Conversely, the anchor$_{DOM}$ CWB2 fold is highly conserved (average RMSD 1.2 Å).

d, Broth-based killing assay in the presence of immune effectors (Lysozyme, orange; LL-37, blue) relative to the non-treated control (black). Resistance to growth impairment and cell death was observed only for two pathogenic S-layer-positive species (*C. difficile* and *P. anaerobius*).

e, f, *C. difficile* CRISPRi SlpA knock-down (**e**) resulted in the sensitivity to both antimicrobial agents in broth-based killing assay (as in **d**), as observed in the S-layer-negative strain FM2.5 (**d**), consistent with S-layer protecting against the host's innate immunity. The S-layer knock-down in *R. lituseburensis* (**f**) did not change its natural sensitivity (**d**) to lysozyme and LL-37, yet it impacted sporulation in plate-based assay for both species.

The gene knock-down was induced with xylose (+) relative to negative plasmid controls (control). CRISPRi SLLP gene knock-downs were validated with RT-qPCR (left panel), SDS-polyacrylamide gels (middle panel) and electron microscopy (right panel). All experiments were performed in biological triplicate, with a standard error of the mean and statistical significance (****, P<0.0001) shown.

341 **S-layer anchoring domains belong to a broader system of swappable cell attachment modules**

342 The observed diversity of *Peptostreptococcaceae* S-layer surface-forming domains results in multiple S-
343 layer architectures (Fig. 2b-d) with distinct biochemical properties (Fig. 3a) and variable physiological
344 functions (Fig. 3d-f). In addition to surface$_{DOM}$s, SLLPs also contain various anchoring domains that attach
345 the S-layer to the cell. Our analysis across the Tree of Life demonstrated that anchor$_{DOM}$s are not unique to
346 SLLPs, can be found in multiple other proteins encoded in bacterial genomes, and that alternative
347 anchor$_{DOM}$s can be swapped during cell-anchored protein evolution (Supplementary Fig. 3). In pathogenic
348 Firmicutes, S-layers utilise at least two different cell-wall anchoring domains, CWB2 in *C. difficile* and SLH in
349 *B. anthracis*, binding to different secondary cell wall polymers (SCWPs)(*6, 46, 47*). The capacity to utilise
350 swappable cell wall anchors could enable attachment of S-layers to different cell wall components, enabling
351 lateral transfer of S-layers between bacteria utilising different cell wall architectures and facilitating co-
352 evolution of different cell envelope layers. Therefore, we investigated if the *Peptostreptococcaceae* S-layers
353 utilise variable cell anchoring mechanisms and if their anchor$_{DOM}$s evolve as part of the wider cell surface
354 proteome.

355 The four types of SLLPs we identified in *Peptostreptococcaceae* (Fig. 2b,c) utilise four distinct cell-
356 anchoring domains (Fig. 4a and Supplementary Fig. 5a,b). All the SLLPs with α/β-sandwich surface$_{DOM}$s,
357 including *C. difficile* SlpA, contained CWB2 anchor$_{DOM}$, while the single β-barrel SLLP contained the SLH
358 anchor$_{DOM}$ (Fig. 3a), also found in β-sandwich surface$_{DOM}$-containing S-layers of *Bacillaceae,* including
359 pathogenic *B. anthracis* (Fig. 1b). However, *Peptostreptococcaceae* S-layers with β-sandwich surface$_{DOM}$s
360 contained one of two distinct anchors: SH3 or Cu-oxidase-like N-terminal domain(*48-50*) (Cu-OX, Fig. 3a).
361 This demonstrated that, within the single bacterial family of *Peptostreptococcaceae,* a variety of anchor$_{DOM}$s
362 are utilised by different S-layer lattice-forming surface$_{DOM}$s, reminiscent of the patterns observed across the
363 Tree of Life (Fig. 1 and Supplementary Fig. 1a).

364 The utilisation of the four distinct cell wall anchoring domains in *Peptostreptococcaceae* S-layers,
365 including those sharing similar structural folds (Fig. 4a), suggested that these domains may represent
366 swappable modules. Our analyses revealed multiple independent swaps, gains, and/or losses of the entire
367 SLH/CWB2 cell wall anchoring systems within *Peptostreptococcaceae* (Fig. 4c), consistent with the patterns
368 we observed within the broader sampling of Firmicutes (Supplementary Fig. 5). These changes usually
369 affect multiple proteins, as SLH and CWB2 anchor$_{DOM}$ are found on average in 12.5 and 9.8 proteins per
370 anchor$_{DOM}$-containing Firmicute genome, respectively (Supplementary Fig. 5b). Based on protein similarity
371 network analyses, the identified anchoring domain swaps were often observed with retention of the
372 attached functional domain homology (Fig. 4e,f, and Supplementary Fig. 3a-d), suggesting that anchoring
373 system swap does not necessarily reflect changes in the surface proteome functionality. Although at least
374 some of the inferred swaps involved rearrangement of entire genomic regions encoding multiple surface
375 proteins (Supplementary Fig. 6), individual surface proteins are often scattered across the chromosome
376 (Supplementary Fig. 5c). Phylogenetic, structural, and sequence homology analyses are consistent with
377 multiple independent anchor replacements among closely related proteins (Supplementary Fig. 6)
378 indicating that, in addition to the swaps of gene cassettes encoding multiple surface proteins, surface
379 proteomes evolve by replacing anchor$_{DOM}$s in individual proteins. Multiple anchor$_{DOM}$ swaps were also found
380 in species with no evidence of S-layer (Supplementary Fig. 6). Therefore, anchoring domain variation is not
381 specific to SLLPs and S-layers, but it rather reflects a more general trend affecting entire cell wall-anchored
382 protein repertoires.

383 Neither CWB2 nor SLH systems were found in any of the analysed *Terrisporobacter/Intestinibacter*
384 genomes. Instead, they contained an expanded repertoire of SH3 domain-containing proteins compared to
385 other *Peptostreptococcaceae* (Fig. 4c), and other Firmicutes (Supplementary Fig. 5a), suggesting CWB2/SLH-
386 loss and SH3-expansion represents their lineage-specific adaptation. Of the two species we studied

387 experimentally, *Terrisporobacter glycolicus* and *Intestinibacter bartlettii,* only *I. bartlettii* exhibited evidence
388 of S-layer presence (Fig. 2d, and Supplementary Fig. 4) and its SLLP contained the SH3 anchoring domain
389 (Fig. 4a). Although the presence of SH3 domains is conserved across all *Terrisporobacter/Intestinibacter*, the
390 SH3-anchored β-sandwich SLLP homologues were specific to a distinct group of *I. bartlettii* isolates (Fig. 2e)
391 suggesting that this novel S-layer has evolved subsequently to the major changes to the cell envelope
392 architecture in the *Terrisporobacter/Intestinibacter* ancestor.
393 For the anchoring domains to perform their function in the attachment of S-layers and other surface
394 proteins to the cell, specific secondary cell wall polymers (SCWP) need to be synthesised and incorporated
395 into the cell wall(*51*). In protein similarity network analyses of all proteins encoded in the 117 *Romboutsia*-
396 and *Paeniclostridium*-genomes which have undergone multiple anchoring system swaps/gains/losses, we
397 identified only a single protein without anchor enriched in CWB2-positive or SLH-positive species (Fig. 4d).
398 The CWB2 system-associated protein was a homologue of a polysaccharide synthase (CD2777) involved in
399 the synthesis of SCWP for CWB2 anchoring in *C. difficile*(*47*). The SLH system-associated protein was a
400 homologue of TagA, involved in SLH-specific SCWP biosynthesis in Firmicutes, including *B. anthracis*(*52*) and
401 *Paenibacillus alvei*(*51*). Therefore, enzymes required for the biosynthesis of SCWP anchors co-distribute
402 with their cognate anchoring domains. The presence of these secondary cell wall polymer biosynthetic
403 enzymes was also detected in organisms where both SLH and CWB2 proteins were encoded in the same
404 genome (Fig. 4d). However, in those cases, the inferred SCWP type reflected that of the more abundant
405 anchoring domain, while the second domain was found in only a few proteins (Fig. 4d), suggesting residual
406 presence from incomplete anchoring domain swap, an indication that this can be a stepwise process.
407 For S-layers to assemble on the cell surface, in addition to the presence of specific cell surface anchors, a
408 protein secretion system is needed to transport SLLP monomers from the cytosol across the cell
409 membrane. An SLLP-specific SecA2 secretion system is utilised in the assembly of α/β-sandwich *C. difficile*
410 S-layer(*53*). However, our analyses revealed that neither of the *Peptostreptococcaceae* genomes encoding
411 SH3- or CuOX-anchored β-sandwich SLLPs encoded any identifiable orthologues of SecA2. This suggests
412 that, in *Peptostreptococcaceae*, SecA2 is associated predominantly with α/β-sandwich CWB2-anchored S-
413 layers and some homologues of β-barrel SLH-anchored SLLPs*,* and that β-sandwich SH3- and CuOX-
414 anchored SLLPs can utilise a more broadly distributed canonical SecA or a not yet identified alternative
415 secretion system. Therefore, S-layers within a single bacterial family are not only assembled from different
416 building blocks but also utilise different mechanisms for their assembly and the evolution of S-layer needs
417 to be coordinated with associated cell wall modification and protein secretion machineries.
418 Together, our analyses revealed that S-layer anchor$_{DOM}$s are part of broader cell wall anchoring systems
419 shaped by modular evolution. Entire systems and anchor$_{DOM}$ of individual proteins were swapped, acquired,
420 or lost on multiple independent occasions, resulting in a mosaic of different anchors and functional
421 domains often found in the genomes of closely related species. Homologues of enzymes known to be
422 involved in the biosynthesis of CWB2- and SLH-specific secondary cell wall polymer anchors are commonly
423 swapped alongside the SLH/CWB2 anchoring systems. This indicates that, at least in
424 *Peptostreptococcaceae*, evolution of S-layer anchoring mechanisms is coordinated with that of cell wall
425 chemistry. The capacity for utilisation of pre-existing swappable anchor$_{DOM}$ modules can also provide an
426 explanation for the diversity of anchors observed in S-layers sampled from across the Tree of Life, including
427 those utilising the same surface$_{DOM}$-fold (Fig. 1b).

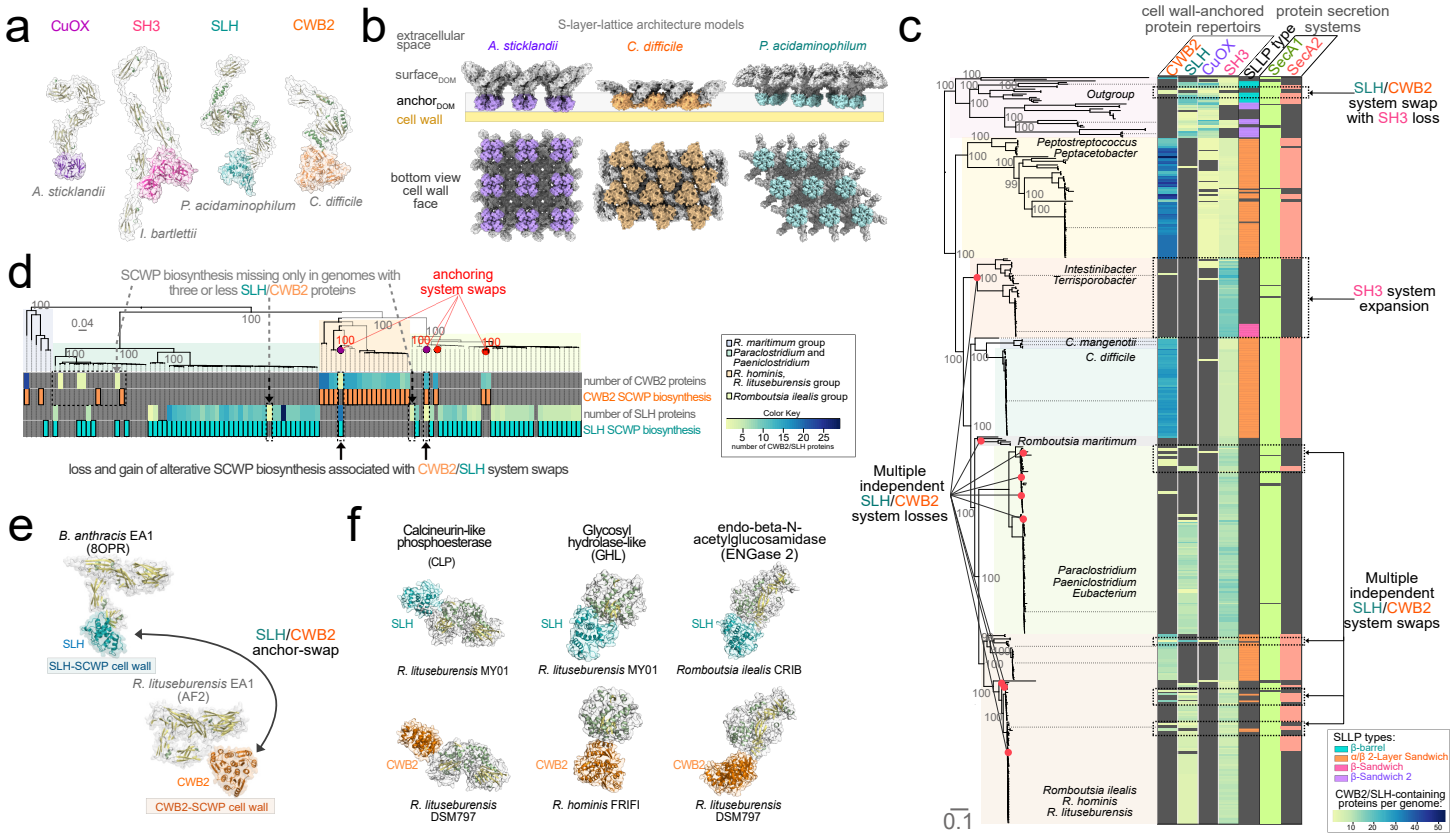

Fig. 4: Modular architecture and evolution of the S-layer and other cell envelope systems.

a, The four experimentally identified *Peptostreptococcaceae* S-layer surface forming fold-types (Fig. 2b) are anchored to the cell using four distinct anchoring domains (orange, CWB2; cyan, SLH; magenta, CuOX; pink, SH3).
b, Surface representations of S-layer lattice models (as in Fig. 2c) with anchor$_{DOM}$s coloured as in **a**.
c, Species tree mapping of the total number (yellow-blue heatmap) of cell wall anchored proteins identified in genomes of 453 *Peptostreptococcaceae*. SH3 anchoring is more broadly conserved, while relatively less common CWB2 and SLH systems are interchangeable and were lost (red circles) or swapped (dashed rectangles) multiple independent times. The more general SecA1 (green) secretion system is present in all lineages, whereas SecA2 (red) with an SLLP-specific role in in *C. difficile*, is found primarily with CWB2-anchored S-layers (SLLP type).
d, Numbers of CWB2 and SLH anchor-containing proteins (yellow-blue heatmaps), and the presence of the secondary cell wall polymer (SCWP) biosynthesis enzymes specific to SLH (cyan) or CWB2 (orange) were mapped onto the *Romboutsia/Paraclostridium/Paeniclostridium* species tree. Enzymes involved in the biosynthesis of SLH- and CWB2-specific SCWPs co-distribute with their respective anchor$_{DOM}$s and are missing only in the genomes containing a low number of anchored proteins (dashed arrows). Multiple independent SLH/CWB2 system swaps (red arrows) and possible complete (red circles) and incomplete (purple circles) system losses (dashed arrows) are apparent.
e, AlphaFold2 structure predictions of the *R. lituseburensis* homologue of *B. anthracis* SLH-anchored S-layer protein EA1 contains a CWB2 anchor$_{DOM}$ consistent with anchoring domain swap and lateral gene transfer between *Peptostreptococcaceae* and *Bacillaceae*.
f, AlphaFold2 structure prediction of individual non-SLLP cell wall-anchored proteins such calcineurin-like phosphoesterase (CLP) and peptidoglycan hydrolases (endo-β-*N*-acetylglucosamidase, ENGase; and glycosyl hydrolases, GHL), with swapped SLH/CWB2 anchor$_{DOM}$s.

451 **Extreme divergence, LGT, and recombination shape diversity of homologous S-layers**

452 The observed diversity of S-layer cell anchoring mechanisms and variation in the number and type of
453 individual subdomains within a single surface$_{DOM}$ structural fold can be explained by modular evolution.
454 However, at this point, it is unclear what mechanisms shape the structural and functional diversity of
455 homologous S-layers utilising the same anchor$_{DOM}$- and surface$_{DOM}$-type. The extremely high level of
456 sequence diversity of homologous S-layer surface-forming domains such as those of α/β-sandwich fold-
457 type (Fig. 3b), complicates evolutionary analyses across the broader taxonomic scales. To circumvent this
458 difficulty, and to better understand the mechanisms of the SLLP-surface$_{DOM}$ evolution, we leveraged the
459 established model of *C. difficile,* an important human pathogen where the S-layer provides a selective
460 advantage within the host during infection(*30*).
461 Our bioinformatic analyses of available *C. difficile* SLLP (*Cd*SlpA) sequences (Fig. 5a) using Duplication-
462 Transfer-Loss reconciliation identified signatures of at least 15 unambiguous recent LGT events (Fig. 5b),
463 consistent with an active process, likely involved in immune evasion(*54, 55*). Of all 18,035 non-atypical *C.*
464 *difficile* genome assemblies available in NCBI, 99.7% contained *Cd*SlpA coding genes. This suggests selection
465 pressure for retention of a single copy of an *slpA* gene, in contrast to *Bacillus anthracis* with its two
466 SLLPs(*6*), for example. Recombination analyses provided strong evidence for at least a single recombination
467 event between parental SlpA$_{SLCT4}$ and SlpA$_{SLCT6}$, resulting in a recombinant divergent SlpA$_{SLCT4}$ (Fig. 5c). Thus,
468 recombination not only plays a part in exchanging different SLCT cassettes between different genome types
469 but can also contribute to generating the diversity observed within a single species, such as among *Cd*SlpAs.
470 To investigate the level of SlpA divergence relative to other protein families encoded within *C. difficile*
471 genomes, we identified 2,950 orthogroups containing a single copy orthologue in at least 90% (SCO90) of
472 265 non-redundant representative *C. difficile* genome assemblies. Within this dataset, SlpA was the least
473 conserved of all identified SCO90s (48.1% average sequence identity, Fig. 5d). Thus, *C. difficile* SlpA
474 represents the most rapidly evolving of all identifiable single-copy orthologue proteins found across all
475 analysed *C. difficile* genomes. It is therefore likely that the high level of SlpA divergence reflects an
476 adaptation of *C. difficile* to selective pressures inside the mammalian gut(*30*). This hypothesis is further
477 supported by the observation that another major cell surface antigen, S-layer-associated protein Cwp66,
478 previously implicated in adhesion(*56, 57*) and survival in the human gut(*57*), is the second least conserved
479 SCO90 (64.3 % average sequence identity, Fig. 5d). The fourth least conserved SCO90 is a homologue of the
480 cell wall teichoic acid glycosylation protein GtrA(*7*). *Listeria monocytogenes* homologue (GtcA) plays a role
481 in cell wall glycosylation(*58, 59*), likely as a flippase(*60*), and the *B. anthracis* homologue is involved in the
482 biosynthesis of SCWPs essential for S-layer surface attachment(*59*). The sixth SCO90 protein is SecA2, the
483 secondary secretion system ATPase required for specific translocation of *Cd*SlpA(*53*), which co-distributes
484 with CWB2-anchored SLLPs in *Peptostreptococcaceae* (Fig. 4c).
485 It is not clear if both LGT and high levels of sequence divergence can also contribute to the diversity of
486 SLLPs in other bacterial lineages. Our bioinformatic analysis of *Bacillus* genomes revealed patterns
487 consistent with both mechanisms in closely related *B. cereus/anthracis* isolates (Supplementary Fig. 7a-d),
488 suggesting that these mechanisms are not limited to *C. difficile* SlpA and may be more common in other S-
489 layer containing species. Interestingly, SAP/EA1 homologues were identified within the *B. cereus/anthracis*-
490 group and not in their closest relatives from SAP/EA1-negative *B. cereus*-lineage (Supplementary Fig. 7d),
491 suggesting differences in their cell envelope architectures. It is important to note that extreme levels of
492 sequence divergence such as those observed in *Cd*SlpA (Fig. 5d) could mask more ancient LGTs, leading to
493 underestimation of S-layer protein exchange between more distantly related taxa. In our analysis, we
494 identified a single clear example of a β-sandwich EA1 surface-forming domain homologue being shared *via*
495 LGT between *Bacillus* and *Peptostreptococcaceae* (Supplementary Fig. 7f,g). As *Peptostreptococcaceae* EA1
496 contained CWB2 and not SLH anchoring domains, it provided the first clear example of anchoring domain

497 swap in SLLP homologues. We were able to experimentally verify the presence of an EA1 homologue in *C.*
498 *mangenotii* LM2 surface protein preparations, where it constituted a minor S-layer component, alongside
499 its dominant CWB2-α/β-sandwich SLLP (Supplementary Fig. 7h,i).
500 Together, our data demonstrate that the observed diversity of S-layers utilising the same anchor$_{DOM}$-
501 and surface$_{DOM}$-types can be generated through extreme sequence divergence, homologous recombination,
502 and LGT between close and more distant relatives. Our analyses also revealed patterns of accelerated
503 evolution in S-layer-associated proteins and proteins involved in S-layer assembly.
504

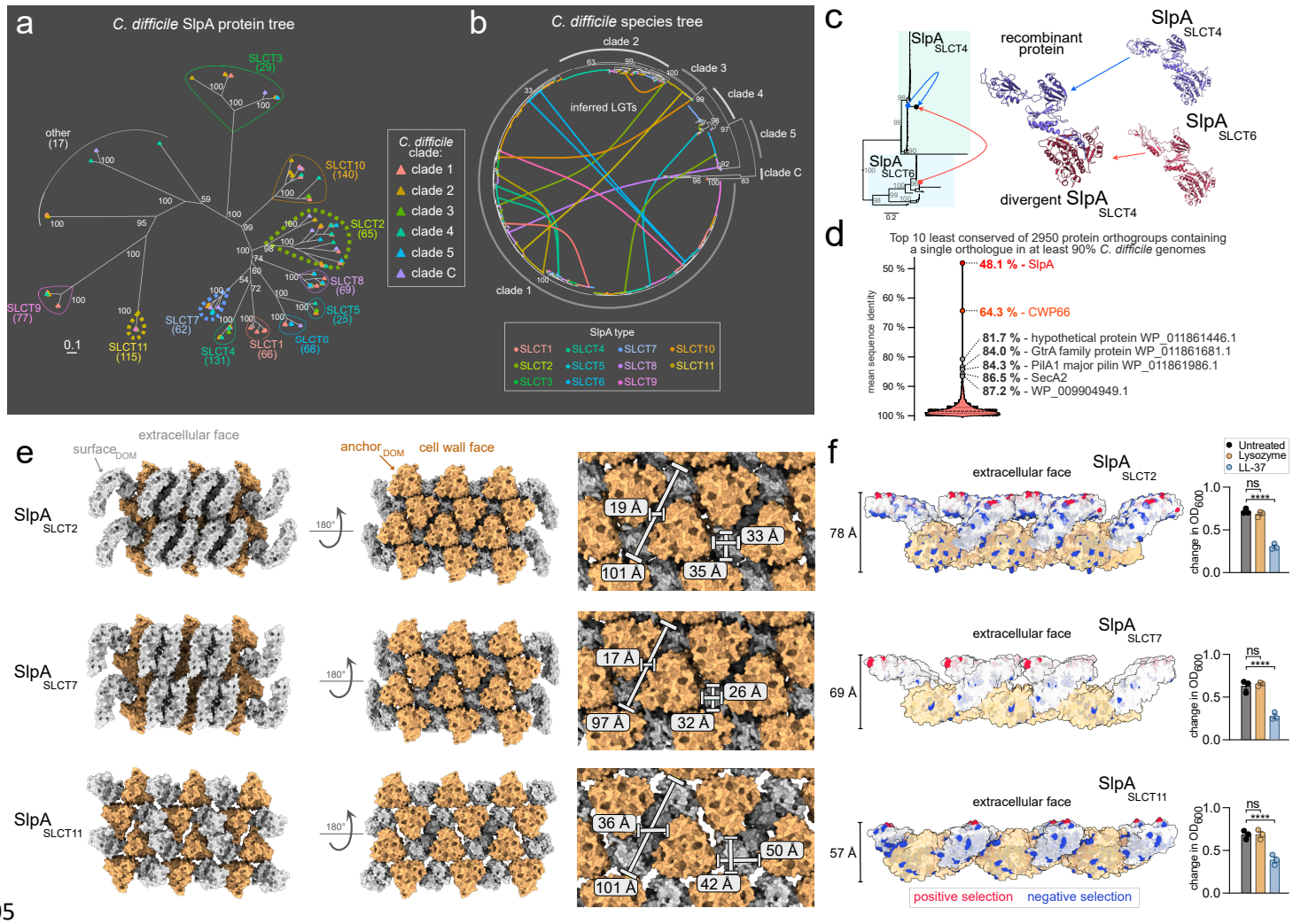

Fig. 5: Evolutionary mechanisms shaping the diversity of S-layers across clinical isolates of a pathogen.

a, Phylogeny of SlpA SLLP colour-coded based on the clade-membership of its source *C. difficile* genome assembly (tree tip colours). Diverse SlpA sequences belong to multiple distinct S-layer cassette types ($SlpA_{SLCT1-11}$).

b, The 14 unambiguous LGT events (coloured lines), identified using Duplication-Transfer-Loss reconciliation analysis, were mapped onto the core genome species tree of 866 *C. difficile* isolates including all six clades (clades 1-5 and C) and coloured based on the SlpA-type membership.

c, An unambiguous recombination event (red and blue arrows) between SLCT4 and SLCT6 identified in recombination detection program RDP4 analyses was mapped onto a $SlpA_{SLCT4}$ and $SlpA_{SLCT6}$ protein phylogeny (left) and onto AlphaFold2 structure models (right) of the two inferred parent proteins (blue or red) and the recombinant sequence (red/blue).

d, Violin plot represents the distribution of mean pairwise sequence identity within 2,950 orthogroups containing a single copy orthologue (SCO) in at least 90 % (240 of 267) analysed representative non-redundant *C. difficile* genome assemblies. Scale bars represent the number of substitutions per site.

e, Representation of the SLLP-formed lattice derived from crystal packing of the three determined SlpA structures. The three SlpAs arrange in different lattice topologies, represented here at different tilt angles: extracellular (left) and cell wall-facing orientations (middle). Distances measured between neighbouring $anchor_{DOM}$s (right, white lines) show that $anchor_{DOM}$s are packed more closely in $SlpA_{SLCT2}$ and $SlpA_{SLCT7}$ lattices and are more separated in the $SlpA_{SLCT11}$ lattice.

f, Mapping of negatively (blue) and positively (red) selected residues onto a side view of the S-layer lattice shown in surface representation for each $SlpA_{SLCT}$ ($surface_{DOM}$ – white, $anchor_{DOM}$ – orange, left). The three lattices differ in thickness reflecting the differences in structural organisation of the SlpA monomers. Despite their extreme sequence divergence (**d**) and structural differences (**e**), all analysed S-layers display conserved patterns of differentially distributed sites under adaptive evolution. While negatively selected residues are found in $surface_{DOM}$ and $anchor_{DOM}$ and on all SlpA surfaces, the sites under positive selection are exclusively found on the environment-exposed extracellular surface of the $surface_{DOM}$. Broth-based killing assay (right) of different $SlpA_{SLCT}$-carrying strains in the presence of lysozyme (orange) and LL-37 (blue) relative to the non-treated control (black) confirms conserved resistance to these innate immune effectors.

531 **Extensive S-layer rearrangements and positive selection across clinical isolates of a pathogen**

532 Over the broad phylogenetic scales, S-layer building blocks from the same structural family can
533 assemble into diverse cell surface lattice architectures (Fig. 1b). To better understand the structural effects
534 of the extreme SLLP sequence variation at the narrowest phylogenetic scales (Fig. 5d), we determined the
535 crystal structure of *C. difficile* $SlpA_{SLCT11}$ and $SlpA_{SLCT2}$, representing distinct phylogenetic SlpA-types from
536 multiple clinical isolates of a single species (Fig. 5a), and compared them to that of the previously studied
537 $SlpA_{SLCT7}$(*40*).
538 All tested SlpA homologues are proteolytically processed following secretion, resulting in two separate
539 peptides: low molecular weight surface layer protein (LMW SLP or SLP_L), containing the S-layer surface-
540 forming domain, and high molecular weight (HMW SLP or SLP_H), containing the CWB2 anchoring domain.
541 Regardless of SLCT membership, the overall topology of *Cd*SlpA structures (Supplementary Fig. 8a) is
542 maintained and consists of a triangular prism arrangement of the CWB2 anchoring domain, an interacting
543 domain (ID) linking $anchor_{DOM}$ and $surface_{DOM}$, followed by one or two domains (D1-only, or D1/D2) of
544 $surface_{DOM}$, protruding the plane of the array(*40*). The structure of the $anchor_{DOM}$ is highly conserved among
545 all determined *Cd*SlpA structures, with core root mean square deviation (RMSD) of ~1 Å (Supplementary
546 Fig. 8a). This is consistent with the higher degree of $anchor_{DOM}$ sequence conservation relative to that of
547 $surface_{DOM}$ in SLLPs from other organisms(*2*) and apparent in the analyses across the Tree of Life presented
548 here (Supplementary Fig. 1-3). Within the more conserved *C. difficile* CWB2 $anchor_{DOM}$, the highest degree
549 of structure variability was observed in loops projecting from the $anchor_{DOM}$ plane towards the extracellular
550 face of the S-layer in $SlpA_{SLCT11}$ or the $surface_{DOM}$/$anchor_{DOM}$ interspace in $SlpA_{SLCT2}$ and $SlpA_{SLCT7}$
551 (Supplementary Fig. 8a). Consistent with the more general pattern observed across the Bacterial Tree of
552 Life (Supplementary Fig. 1 and 2), the most prominent differences across the SlpA structures were observed
553 within their environment exposed S-layer surface-forming domain (Fig. 5e-f, Supplementary Fig. 8). Unlike
554 the $SlpA_{SLCT7}$ and $SlpA_{SLCT2}$, which share a 2-domain (D1/D2) arrangement, $SlpA_{SLCT11}$ lacks the entire D2
555 domain, which accounts for over 20% of total solvent accessible area of the S-layer unit, resulting in a
556 smoother (Fig. 5e) and thinner array (Fig. 5f). The stark contrast between the $surface_{DOM}$s of $SlpA_{SLCT2}$,
557 $SlpA_{SLCT7}$, and $SlpA_{SLCT11}$ provided evidence of how the entire fold that determines S-layer architecture can
558 be reshaped within the relatively short phylogenetic timescales of emergence of different isolates of the
559 same species. The *Cd*SlpA $surface_{DOM}$ amino acid sequence regions align poorly with each other, further
560 demonstrating their extreme sequence divergence (Fig. 3b and 5d) and providing a likely explanation for
561 difficulties in identifying homologous SLLPs from more distantly related species. Interestingly, $surface_{DOM}$-
562 $anchor_{DOM}$ interactions form the majority of the lattice interfaces in all studied *Cd*SlpAs (Supplementary Fig.
563 9), suggesting that the main role of the $surface_{DOM}$ may be as a crystallisation enhancer module which
564 forms an S-layer lattice only in conjunction with the more common CWB2 anchoring domain. This mode of
565 S-layer lattice assembly, involving both the anchoring and surface domains, could explain the shorter
566 $surface_{DOM}$ in *Cd*SlpA relative to those identified in other *Peptostreptococcaceae* S-layer-types (Fig. 2c) and
567 possibly the lack of any identifiable $anchor_{DOM}$ swaps involving α/β-sandwich $surface_{DOM}$.
568 Despite the observed differences in their lattice architecture, all the studied *C. difficile* isolates utilising
569 different S-layer variants displayed conserved resistance to natural antimicrobial agents LL-37 and lysozyme
570 produced by the human immune system (Fig. 5f). This demonstrates that considerable structural changes
571 of lattice can evolve without compromising the physiological functions of the S-layer. It also raises
572 important questions about the reasons behind the apparent sensitivity to the antimicrobials observed in
573 other *Peptostreptococcaceae* utilising homologous S-layers (Fig 3d-f).
574 Finally, it remained unclear if the observed structural variation in *Cd*SlpAs is mainly shaped by genetic
575 drift and recombination or if any of the sites are under positive selection. To test this, we analysed non-
576 redundant nucleotide sequences encoding SlpA identified in all 18,035 analysed *C. difficile* genomes. All

577 *Cd*SlpA types displayed patterns of positive selection and all identified sites under positive selection were
578 within the surface$_{DOM}$ and were surface exposed (Fig. 5f and Supplementary Fig. 10). This pattern was
579 contrasted by the negatively selected residues found within surface$_{DOM}$ and anchor$_{DOM}$, and at all lattice-
580 forming faces of the S-layer array (Fig. 5f). This is consistent with previous analyses indicating that surface
581 antigens of pathogenic bacteria and viruses, and especially their exposed epitopes, are often under positive
582 selection(*61, 62*).
583 The positive selection patterns detected only at the S-layer face exposed to the extracellular
584 environment are consistent with its proposed role in phage and immune evasion. Extreme sequence and
585 structural variation can facilitate the pathogen's arms race with human adaptive immunity without
586 compromising its conserved role in protection against lysozyme and LL-37, two components of the innate
587 immune response. This model is consistent with the recently described capacity to generate monoclonal
588 antibodies selective towards different SlpA variants(*63*).
589 In summary, the evolution of a single S-layer-forming protein can have a dramatic impact on the
590 architecture of the bacterial cell surface at the narrowest phylogenetic scales which, in pathogens such as
591 *C. difficile*, contributes to the protection against the human immune system.
592
593
594 **Discussion**

595 As opposed to the archetypical elements of bacterial cell envelope architecture - lipid cell membrane
596 and peptidoglycan cell wall - the main building block of a specific S-layer lattice is commonly encoded by a
597 single gene and therefore can provide a substantial level of evolutionary plasticity to the microbial cell
598 surface. This plasticity is of particular significance in bacterial pathogens such as *B. anthracis* or *C. difficile*,
599 where the S-layer plays a role in disease(*10, 16, 30, 32*).
600 Our bioinformatic analyses demonstrate that the observed diversity of *C. difficile* S-layers is shaped by
601 lateral gene transfer, homologous recombination, adaptive evolution, and extreme sequence divergence.
602 Similar mechanisms can also generate the diversity of SLLP folds, and S-layer crystal lattice architectures
603 across the bacterial Tree of Life (Fig. 1 and references(*34, 35*)), as we observed in wider sampling of
604 *Peptostreptococcaceae* and *Bacillus*. Comparative analyses of the two crystal structures of *Cd*SlpA
605 homologues experimentally determined in this study show that extreme surface$_{DOM}$ sequence divergence
606 can result in significant changes to the S-layer architecture including lattice thickness, anchor$_{DOM}$ packing,
607 and extensive rearrangements of its environment-exposed face. The observed architectural variation of *C.*
608 *difficile* S-layers does not perturb its physiological functions in spore formation and resistance against
609 effectors of the human innate immunity such as lysozyme and LL-37.
610 In the case of pathogenic bacteria, the extensive plasticity of their environment exposed surfaces may
611 reflect an arms race with external threats to the pathogen's survival such as bacteriophages(*18, 19*) and
612 human adaptive immunity(*64*). Since the S-layer is an assembly of multiple copies of *Cd*SlpA, a single
613 mutation of an environment-exposed positively selected residue can correspond to changes in up to an
614 estimated 0.1 - 1% of the entire solvent accessible outer surface of *C. difficile* SlpA crystal lattice
615 (Supplementary Fig. 9d). The observed structural plasticity, particularly at the more externally exposed
616 domains (e.g. sites under positive selection), shows how a single mutation can affect a significant
617 proportion of the host-pathogen interface. This includes altering potential surface epitopes, and therefore
618 contribute to immune evasion, as *Cd*SlpA has been directly implicated in interactions with the host and
619 eliciting immune response(*64*). Understanding the evolutionary mechanisms shaping *Cd*SlpA sequence
620 diversity and the resulting structural plasticity of the S-layer lattice is important if we are to therapeutically
621 target the S-layer(*10, 19, 33*). To avoid mutation and/or recombination-driven resistance mechanisms
622 arising against any drugs designed to disrupt the S-layer, we need to consider the short-time scale

623 evolution processes identified here. Furthermore, the structural landscape of S-layers needs to be
624 considered as recent immunological studies identified that high-affinity anti-LMW monoclonal antibodies
625 binding distinct epitopes of a single SlpA have different effects on *C. difficile*'s growth, toxin secretion, and
626 biofilm formation(*33*).
627 The high level of S-layer diversity is not limited to clinical isolates of *C. difficile* and was observed in
628 different strains of *Bacillus stearothermophilus*(*65*) and on a broader taxonomic scale in divergent SLLP
629 sequences found across multiple species of *Bacillus* and *Lactobacillus*(*66*). At an even broader phylogenetic
630 scale, among fourteen tested representatives of a single family of Bacteria, *Peptostreptococcaceae*, we
631 detected and experimentally verified nine new SLLPs with one novel $surface_{DOM}$ fold-type (β-barrel, CATH
632 2.4) and two anchoring domains (Cu-OX and SH3), not observed in any of the previously characterised
633 SLLPs. Our results demonstrate that the diversity of S-layers in *Peptostreptococaceae* is shaped by modular
634 evolution(*67, 68*). This is apparent when the mosaic of different genes, protein folds, secretion systems,
635 and surface anchoring systems, consisting of different anchoring domains and different chemical pathways
636 generating different secondary cell wall polymer modules, are mapped onto their species tree (Fig. 6a).
637 However, this modularity is not restricted to a single mode of evolution but rather encompasses multiple
638 'complexity' levels. At the genome level, entire multigene S-layer cassettes can be swapped *via* LGT and
639 recombination, and alternative multicomponent anchoring systems (e.g. CWB2 and SLH), including the
640 surface-anchor biosynthesis machinery and its cognate anchoring domains of multiple proteins, can be
641 replaced. At the individual protein level, alternative anchoring domain types can be swapped or gained,
642 enabling otherwise homologous proteins to utilise one of the multiple cell-anchoring systems that coexist
643 within a single bacterial genome (e.g. choline-binding and CWB2). Similar patterns were observed at the
644 subdomain level in lattice-forming $surface_{DOM}$s, as demonstrated by variation in the number of β-sandwich
645 (sub)domains within Ig-like arrays(*69*), the most common fold of known SLLP architectures, acquisition of a
646 novel α-helical surface subdomain in α/β-sandwich S-layer of *A. irregularis*, or as the capacity to lose the
647 entire D2 subdomain in the α/β-sandwich fold of *C. difficile* $SlpA_{SLCT11}$. In addition, the reorganisation of
648 subdomains may also play roles in the incorporation of divergent S-layer proteins into an existing lattice
649 such as that proposed for the minor S-layer component SlpX in *Lactobacillus(70)* or anchor-swapped EA1
650 homologue in *C. mangenotii.* At the cell biology level, the entire S-layer can be considered as a cell
651 envelope module which can be swapped, remodelled, or lost. This multilevel modular architecture provides
652 a flexible framework for various lineage- and species-specific cell surface adaptations (Fig. 6b). In
653 *Romboutsia*, CWB2 and SLH secondary cell wall polymers were swapped, gained, or lost multiple
654 independent times, often as part of major genome rearrangements leading to multiple inferred S-layer
655 losses. In *Clostridioides/Asaccharospora*, a single core CWB2-anchored α/β-sandwich S-layer scaffold is
656 retained, and the variation of the cell envelope is mostly due to the modification to the pre-existing lattice
657 or addition of S-layer associated proteins. In *Terrisporobacter/Intestinibacter*, major ancestral changes to
658 the cell wall chemistry have likely caused a loss or swap of the ancestral S-layer and driven the emergence
659 of a novel SH3-anchored β-sandwich SLLP.

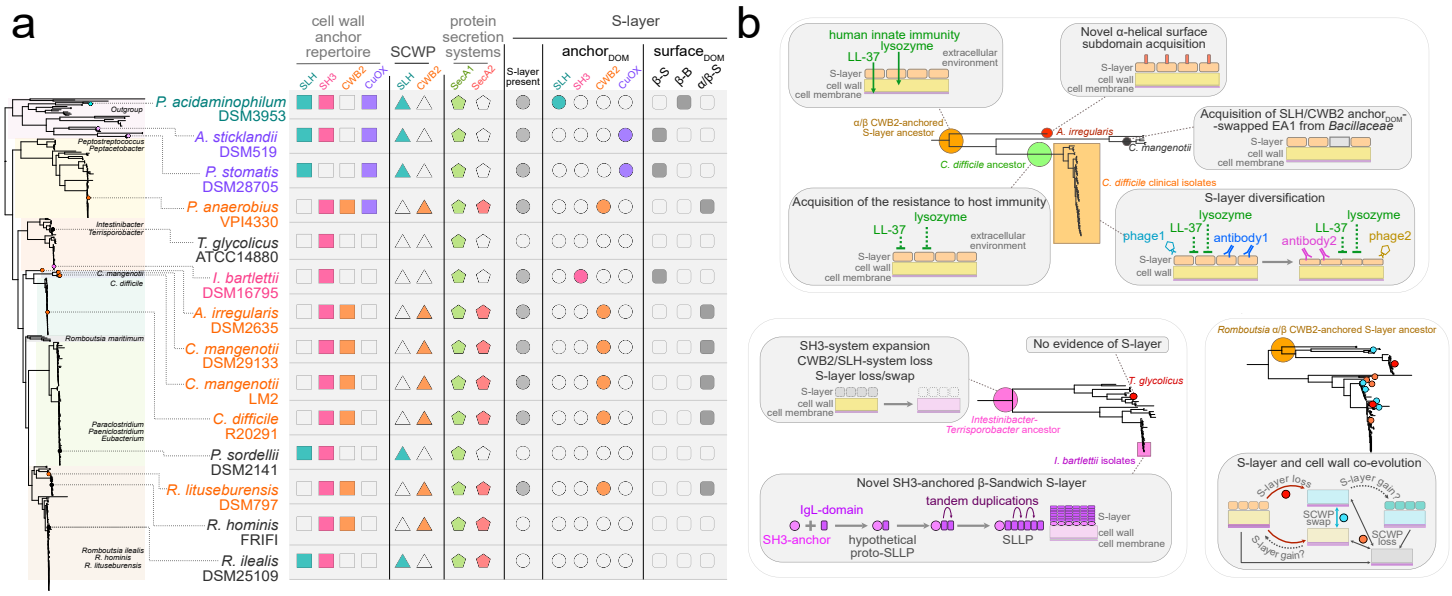

Fig. 6: Multiple evolutionary pathways to S-layer exoskeleton gain, loss and diversification in a single family of Bacteria.

a, An array representing presence (solid shapes) or absence (empty shapes) of cell wall anchoring domains found in at least a single protein per genome (rectangles), enzymes involved in biosynthesis of SLH- or CWB2-specific SCWP cell anchors (triangle), SecA protein secretion systems, including SecA2 involved in S-layer assembly in *C. difficile* (pentagons), experimentally verified presence of S-layer (grey circles), and anchor$_{DOM}$-types (coloured circles) and surface$_{DOM}$ folds (rounded rectangles) of experimentally identified SLLPs mapped onto a species tree of *Peptostreptococcaceae*. The scattered distribution of a mosaic of different cell envelope components including systems involved in S-layer assembly and anchoring to the cell is consistent with their diversity being shaped by modular evolution.

b, Three distinct models of S-layer evolution based on the results presented in this study demonstrate that the observed diversity of S-layers is generated by multiple mechanisms underlying the extreme evolutionary plasticity of this major component of bacterial cell architecture: (i) structural and functional diversification of the conserved α/β-sandwich CWB2-anchored S-layers in *Clostridioides-Asaccharospora* lineage (top panel); (ii) recent strain-specific emergence of a novel SH3-anchored β-sandwich S-layer protein in *Intestinibacter bartlettii* (bottom left panel) and (iii) evidence of a specific loss of S-layer within *Romboutsia*, a lineage displaying a high frequency of cell envelope remodelling (bottom right panel).

676 Modular domain organisation was recently observed in the structure of the proposed S-layer-
677 associated protein, extracellular Fe(III) reductase from the Gram-positive bacterium *Carboxydothermus*
678 *ferrireducens*, containing a domain homologous to *B. anthracis* SLH(*71*). This suggests that modular
679 evolution can play a role in additional functions provided by S-layer-associated proteins, in this case in the
680 Fe(III) respiratory process, as well as in shaping S-layer architecture. Furthermore, modular architecture
681 was observed in many cell wall binding proteins(*72*), including lytic enzymes containing choline-binding
682 modules(*73*), consistent with our detection of cell wall homeostasis proteins being subject to frequent cell
683 wall binding domain swaps. Therefore, the observed evolutionary mechanisms shaping the S-layer likely
684 represent a more general phenomenon of bacterial cell envelope modularity that also includes cell wall and
685 cell membranes. Although the gene clusters encoding proteins involved in peptidoglycan biosynthesis are
686 highly conserved across Bacteria, rare insertions of outlier genes were observed in Firmicutes and
687 Candidate Phyla Radiation (CPR)(*74*) and our results demonstrate that secondary cell wall polymer
688 biosynthesis machinery was changed on multiple independent occasions, at least in Firmicutes. It was also
689 hypothesized that another cell envelope module, the outer membrane, was lost in Gram-negative (diderm)
690 bacteria multiple times, resulting in the independent emergence of Gram-positive (monoderm) lineages(*42,*
691 *75, 76*). However, as opposed to more complex cell membranes and cell walls, the main S-layer building
692 block SLLP is encoded by a single gene and therefore it can provide an extreme level of evolutionary
693 plasticity to microbial cell surfaces.
694 The diversification of S-layer cell surfaces can result not only in diverse architectures, symmetry types,
695 and biochemical properties but also in variable physiological functions. Probably the most striking example
696 is that of *Peptostreptococcaceae* S-layer utilising the same α/β-sandwich surface$_{DOM}$-fold and the same
697 CWB2 anchor$_{DOM}$ where only human gut pathogens *P. anaerobius* and *C. difficile* isolates (with their diverse
698 S-layer variants) are resistant to natural antimicrobials lysozyme and LL-37. In contrast, resistance to these
699 innate immunity effectors was not observed in α/β-sandwich surface$_{DOM}$-containing species unlikely to
700 cause human disease including *A. irregularis* DSM2635 and *R. lituseburensis* DSM797 (isolated from mud),
701 and *C. mangenotii* DSM29133 (isolated from intestine of a healthy mouse(*77*)). These results indicate that
702 the presence of divergent versions of the same type of S-layer can contribute to the protection against the
703 innate immunity effectors in pathogens but not in non-pathogenic species of the studied
704 *Peptostreptococcaceae*. Therefore, S-layer evolution can play a role in adaptation to the pathogenic
705 lifestyle. Furthermore, recent data suggests that distinct S-layers of *Lactobacillus* also play an important
706 role in probiotic bacteria associated with optimal health, where it can act as a mediator of the host
707 inflammatory response and prevent activation of pro-inflammatory signalling(*26, 28, 29*). It is currently
708 unclear if these properties are conserved across diverse *Lactobacillus* S-layers(*66, 70*).
709 Our data also suggests that the adoption of certain S-layer architectures can enforce some structural
710 and evolutionary constraints. While surface domains of some SLLPs appear to be found mostly with a
711 certain type of anchoring domain, others have undergone multiple independent anchor$_{DOM}$ swaps. For
712 example, in our analyses, all the α/β-sandwich SLLPs, including *C. difficile* SlpA, contained only CWB2
713 anchoring domains. This could be due to structural or functional constraints such as the importance of the
714 surface$_{DOM}$-anchor$_{DOM}$ interactions within their S-layer lattices, or it may merely represent a limitation of
715 currently available sampling. One potential explanation for these observations is that the emergence of S-
716 layers where surface$_{DOM}$-anchor$_{DOM}$ interactions form most of the lattice interfaces results in increased
717 structural dependence between the specific surface$_{DOM}$ and anchor$_{DOM}$, leading to the entrenchment of this
718 architecture. Such evolutionary ratchet mechanism(*78, 79*) could lead to a decreased capacity to accept
719 alternative anchor$_{DOM}$s. No such restrictions were observed in the more common β-sandwich surface$_{DOM}$
720 homologues, which can function with multiple different anchor$_{DOM}$s and can be transferred *via* LGT
721 between more distantly related species.

722 Elucidating the origins and evolution of S-layers is complicated by the major gaps in our knowledge,
723 including (1) insufficient number of experimentally verified S-layer-positive and -negative species from
724 across the Tree of Life, (2) limited understanding of the conserved and lineage-specific S-layer functions
725 outside of a few model organisms, (3) unclear capacity for the *de novo* origins of S-layers from non-S-layer
726 proteins, (4) limited understanding of the physiological selection pressures and mechanisms underlying
727 possible S-layer gains and losses. The limited experimental evidence for S-layer loss includes the emergence
728 of *Cd*slpA mutants in response to the experimental exposure to specific antimicrobials (diffocins)(*10*), the
729 absence of S-layer in some strains of *Corynebacterium glutamicum(36)*, and interspersed phylogenetic
730 distribution of S-layer positive *Peptostreptococcaceae* and their relatives with no evidence of S-layer
731 presented in this study. Interestingly, pathogenic *Corynebacterium* species lack identifiable homologous of
732 canonical S-layer protein PS2 found in the non-pathogenic representatives of the genus(*80*) and the
733 presence of the S-layer in the pathogenic species awaits verification. If absent, this would suggest that, at
734 least in some lineages, loss of the S-layer cell envelope module could also play a role in the emergence and
735 evolution of pathogens. More indirect evidence of the possibility of S-layer loss comes from the absence of
736 S-layers in model bacteria Gram-negative *Escherichia coli* and Gram-positive *Bacillus subtilis*(*2*). It is also
737 unclear if any of the extant Eukaryotes contain cellular outer shell homologous to the prokaryotic S-layer.
738 Although the presence of S-layers *sensu stricto* to our knowledge has not been demonstrated in
739 Eukaryotes, paracrystalline glycoprotein cell wall layers can be found in some species such as green algae *C.*
740 *reinhardtii*(*81, 82*). The available evidence suggests that the S-layer is present in most *Archaea*(*2*), raising
741 questions about whether it was also present in the Last Common Eukaryotic ancestor and subsequently lost
742 or whether its loss was an important step in early Eukaryogenesis. Interestingly, the latest
743 computational(*34*) and experimental(*83*) approaches did not identify S-layers and SLLP candidates in
744 Asgardarchaeota, suggesting that the closest archaeal relatives of eukaryotes can be S-layer negative.
745 However, this may not be a conserved feature of all Asgardarchaeota(*84*) and further experimental studies
746 are required to robustly verify this hypothesis.
747 To our knowledge, there are no previously verified examples of S-layer gain by an S-layer-negative
748 ancestor in the natural environment. However, further studies in *Peptostreptococcaceae,* including
749 *Romboutsia,* or different species of *Corynebacterium(36, 80)* could provide a much-needed model for
750 studying possible S-layer acquisitions. The presence of multiple bacterial species with extremely diverse S-
751 layers and species with no evidence of S-layer scattered across *Peptostreptococcaceae* phylogeny
752 presented in this study could lend credibility to the hypothesis that S-layers have evolved multiple
753 independent times(*2*). Perhaps the most striking example is that of *I. bartlettii* where a lineage-specific SLLP
754 has emerged within a subset of isolates of a single species, likely providing an excellent model for studying
755 S-layer acquisitions. It remains to be tested if and how a non-S-layer protein can evolve *de novo* to form a
756 paracrystalline cellular exoskeleton in nature. One captivating example of a possible evolutionary
757 intermediate in the S-layer evolution is that of a toga, a proteinaceous surface layer with membrane
758 patches, found in *Thermotoga maritima*(*85*). The evolutionary model of S-layer origin from an outer
759 membrane or a toga-like cell envelope layer could be a reverse of that proposed for outer membrane
760 biogenesis from the S-layer(*86*), i.e. a decrease in the proportion of lipids and an increase in the proportion
761 of proteins capable of self-assembly within the outer cell envelope layer. This hypothesis, however, does
762 not account for the possibility of the independent S-layer emergence in monoderm Bacteria, where, in the
763 absence of the outer lipid-containing layer, *de novo* S-layer would have to evolve *via* a different
764 mechanism. An alternative mechanism can be postulated based on the inferred origin of the S-layer in *I.*
765 *bartlettii*. Its SLLP, containing six nearly identical β-sandwich Immunoglobulin-like surface$_{DOM}$ subdomains,
766 has emerged *via* multiple tandem duplications of a single ancestral subdomain of a type commonly found
767 across bacteria. If this mechanism indeed plays a role in *de novo* evolution of S-layers, it could provide one
768 explanation for the common use of β-sandwich fold in S-layers observed across the Tree of Life.

769 Our findings demonstrate that bacteria can rapidly gain, lose, replace, or modify the S-layer cell surface
770 exoskeleton through the utilisation of a multilevel modular architecture. Thanks to this scalable framework,
771 commonly composed of a single repeating SLLP building block, a single genetic change can significantly alter
772 the entire exoskeletal lattice. The segmented design of the SLLP adds another layer of flexibility, allowing it
773 to attach to various cell surface configurations through swappable cell anchoring domains. The ability to
774 remove, add, or remodel surface domain segments promotes rapid adaptation of the S-layer layout. This
775 versatile framework underlies the diversity of S-layer architectures observed across the prokaryotic Tree of
776 Life and offers opportunities for bioprospecting new natural lattice designs for nanotechnology, protein
777 engineering, and bio-product development. Our study uncovered the molecular, structural, and
778 evolutionary processes behind multiple independent origins, losses, and the remarkable structural and
779 functional divergence of this major component of bacterial cell biology. This extraordinary evolutionary
780 plasticity underscores the frontline role of these biological exoskeletons in adapting to extracellular
781 environments. In the case of pathogenic *Clostridioides difficile*, the S-layer has played a pivotal role in the
782 emergence of ancestral resistance to innate immunity effectors and its evolvability contributes to the
783 ongoing thug-of-war with adaptive immunity at the host-pathogen interface.
784

785

786 **Materials and Methods**

787

788 **Inference of the Tree of Life**

789 The Tree of Life was inferred as described previously(*87*), using the same dataset expanded with an
790 additional 127 genomes of organisms with available evidence of S-layers and/or SLLPs, resulting in a final
791 dataset of 2,705 Bacteria, 315 Archaea, and 149 Eukaryota genomes. Shortly, seven conserved single-copy
792 marker proteins were identified using hmmsearch (-E 1e-5) of HMMER3.3(*88*). The identified orthologue
793 sequences were aligned with MAFFT(*89*) trimmed with trimAL(*90*) (gappyout mode), concatenated with
794 Biopython module Nexus(*91*), and used for the inference of phylogenies in IQ-TREE(*92*) under LG + G4
795 model with 1,000 ultrafast bootstraps(*93*).
796

797 ***Peptostreptococcaceae, Bacillaceae* and Firmicutes species trees**

798 A set of 400 of all available NCBI(*94*) *Peptostreptococcaceae* genome assemblies, excluding
799 overrepresented *C. difficile* genomes (18,035 assemblies analysed separately in this study), was
800 downloaded and expanded by a minimal representative set of 53 *C. difficile* genomes, to include all SLCT-
801 types and clades identified in our analyses, resulting in the final set of 453 *Peptostreptococcaceae* genome
802 assemblies (Supplementary Data 2). To identify 111 broadly conserved *Peptostreptococcaceae* single-copy
803 orthologues (Supplementary Data 3), protein sequences of 609 single-copy orthologues identified in our
804 analyses of *C. difficile* genomes were used as queries in phmmer search (-E 1e-10) of HMMER3.3(*88*)
805 against a local database of 1,264,675 protein sequences encoded in 453 analysed *Peptostreptococcaceae*
806 genomes.
807 Due to a large number of available Firmicute genome assemblies (470,472), a custom set of 828
808 genomes (Supplementary Data 2), including representatives of all major lineages, was selected for further
809 analyses. All 2,247,509 protein sequences encoded in the 828 Firmicutes genomes were analysed using
810 EGN(*95*) protein similarity network analysis, and the largest network (1,442,176 proteins) was subdivided
811 based on multilevel community detection algorithm in R. Of all identified protein networks, 84 represented

812 single copy orthologues (Supplementary Data 3) broadly conserved across the analysed Firmicutes
813 genomes.
814 In a dataset of all available 7,377 GeneBank[8] *Bacillus* genome assemblies, the 84 Firmicutes single
815 copy orthologues were identified (Supplementary Data 3) using phmmer search (-E 1e-10). As many *Bacillus*
816 species are overrepresented, including multiple isolates of pathogens from *B. cereus/B. anthracis* lineage, a
817 set of non-redundant assemblies was identified by pruning the redundant tree leaves (edge length = 0)
818 from a species tree of 7,377 *Bacillus* using Treemmer(*96*) to 99.999999% of the original tree length (-RTL
819 0.99999999). The final set of 4,774 non-redundant *Bacillus* assemblies (Supplementary Data 2) encoding at
820 least 75 of 84 orthologues was used for species tree inference.
821 For each dataset, protein sequences of the identified single copy orthologues were aligned with
822 MAFFT(*89*) trimmed with trimAL(*90*) (gappyout mode), concatenated with Biopython module Nexus(*91*),
823 and used for the inference of *Peptostreptococcaceae* (Supplementary Data 4), Firmicutes (Supplementary
824 Data 5), and *Bacillus* (Supplementary Data 6) phylogenies in FastTree(*97*) under -lg -gamma model.
825
826 **Identification of the SLLP homologues and common domain sequence removal**

827 To investigate the presence of known protein domains and sequence motifs in available
828 experimentally verified SLLPs (Supplementary Data 1), the SLLP sequences were scanned against local
829 copies of HMM profile databases SUPFAM(*98, 99*), Pfam(*100*), CATH(*101*), NCBI HMM(*102*) including
830 TIGRFAM(*103*), and PROSITE(*104*) using hmmscan (-E 0.001) of HMMER3.3(*88*). To search for proteins and
831 genomes containing the domains and sequence motifs identified in SLLPs, their HMM profiles were used in
832 hmmsearch, HMMER3.3(*88*) against the Tree of Life dataset of 10,667,779 protein sequences encoded in
833 the 3169 analysed genomes. To identify the SLLP homologues, their full-length sequences (Supplementary
834 Data 1) and the surface$_{DOM}$-only sequences (following anchor$_{DOM}$ sequence removal, Supplementary Data 1)
835 were used as queries in BlastP(*105*) searches (E-value threshold = 1e^{-5}) against the Tree of Life dataset. Data
836 presented in this paper indicate that cell surface anchoring domains can display broader phylogenetic
837 distribution (Fig. 1a and Supplementary Fig. 1c and 2), be present in multiple otherwise unrelated proteins
838 encoded within the same genome (Supplementary Fig. 3g), and share higher sequence and structural
839 conservation than the domains responsible for the function of the anchored protein (e.g. surface$_{DOM}$ of
840 SLLPs, Fig. 3c and Supplementary Fig. 8a). This in turn can result in masking of their distinct evolutionary
841 histories evident as clustering into a single protein similarity network of functionally and structurally
842 unrelated functional domains sharing the same anchoring domain, and cause network 'hairballs' due to
843 frequent anchoring domain swapping (Supplementary Fig. 3). To account for that, in addition to the full-
844 length sequence and structural analyses, we analysed anchorless functional domain-only sequences
845 separately (Supplementary Data 8). The sequence regions corresponding to the common anchor$_{DOM}$s were
846 identified with InterProScan 5(*106*) and removed from full-length protein sequences using Bio.Seq module
847 of the Biopython package(*91*). Adjacent domain repeats (e.g. three subsequent CWB2 or SLH domains)
848 were treated as a single region (the start and end of the removed region corresponded to the start of the
849 first repeat and the end of the last repeat) and were removed from the full-length sequences including any
850 in-between linker amino acid residues. Correct sequence removal was verified using InterProScan 5(*106*),
851 manual sequence inspection, and AlphaFold2(*107*) structure predictions. For clarity, throughout the text,
852 we collectively refer to the sequences left following the anchoring domain removal as the 'functional
853 domain' sequence.
854 *Bacillus* SAP/EA1 SLLP homologues (Supplementary Data 7) were identified with the protein
855 similarity network analyses of all *Bacillus* SLH/CWB2/SH3-containing proteins following the anchor$_{DOM}$
856 sequence removal and verified with hmmsearch using custom HMM profiles generated with the identified
857 SAP, EA1, and SAP2 surface$_{DOM}$-only sequences.

858

859 **Protein similarity network analyses**

860 Unless stated otherwise, all protein similarity network analyses were performed using the same
861 protocol. Protein networks were generated with BLAST(*105*), and quick edge file optimisation implemented
862 in EGN(*95*) using the following thresholds: BLAST (e = $1e^{-5}$), simple link parameter (E-value threshold = $1e^{-5}$,
863 hit identity threshold = 20%, minimum hit coverage = 20%, without the best reciprocal hit condition
864 enforcement), and with the quick edge file creation (E-value threshold = $1e^{-5}$, hit identity threshold = 20%,
865 minimum hit coverage = 20%, minimal hit length = 30 amino acids). The identified networks were analysed
866 using igraph(*108*) and dnet(*109*) packages in R(*110*) and visualised in Cytoscape(*111*). Where indicated,
867 communities of the densely connected subgraphs were isolated using the multilevel and walktrap
868 community detection algorithms implemented in the igraph package(*109*).

869

870 **Analysis of the cell anchoring domain-containing protein repertoires**

871 To investigate the evolution of cell surface anchored protein repertoires, we used two separate
872 datasets of organisms with evidence of frequent anchoring domain system swaps in our across the Tree of
873 Life analyses: a broader taxonomic set of 828 Firmicutes genomes encoding 2,247,509 protein sequences,
874 and narrower taxonomic set 453 *Peptostreptococcaceae* genomes encoding 1,264,675 protein sequences.
875 Proteins containing CWB2 (PF04122.15), SLH (PF00395.23), and SH3 (PF08239.14) anchoring domains were
876 identified using hmmsearch (-E $1e^{-1}$), HMMER3.3(*88*). To extract functional domain sequences
877 (Supplementary Data 8), sequence regions corresponding to the anchoring domains (SLH, IPR001119;
878 CWB2, IPR007253; SH3, IPR003646) were removed from the full-length sequences using Bio.Seq module of
879 the Biopython package(*91*). The functional domain homology was investigated with EGN network analyses,
880 annotated with the output of InterProScan 5, mapped onto species trees in R, and further verified with
881 manual inspection of MAFFT(*89*) protein sequence alignments and structural prediction in AlphaFold2(*107*)
882 for selected representatives.
883 To identify any additional homologues containing anchoring domains other than SLH, CWB2, or SH3
884 in *Peptostreptococcaceae*, including *Romboutsia*, $anchor_{DOM}$-removed functional domain-only sequences
885 from each functional domain homology network were aligned in MAFFT(*89*) and resulting alignments were
886 used to generate custom functional domain HMM profiles. These profiles were used to search the original
887 datasets with hmmsearch (-E $1e^{-5}$). The sequences identified with the functional domain HMM profiles
888 (Supplementary Data 9 and 10) were analysed with EGN network analysis following $anchor_{DOM}$ -removal
889 (Supplementary Data 10), annotated with InterProScan 5[19] to verify the presence of conserved functional
890 domains and any additional $anchor_{DOM}$s. The identified domains were mapped onto the networks in
891 Cytoscape(*111*) and verified with manual inspection of MAFFT(*89*) protein sequence alignment and
892 structural prediction in AlphaFold2(*107*) for selected representatives.

893

894 **Genomic region synteny**

895 Initial genome alignments of *Romboutsia* and *Intestinibacter* genome assemblies downloaded from
896 NCBI were performed with Mauve(*112*) to identify regions syntenic to those encoding SLLPs that we
897 experimentally verified in S-layer-positive species. Final pairwise alignments for the identified genomic
898 regions were generated using minimap2(*113*) with parameters (-X -N 50 -p 0.1 -c) and visualised in R using
899 gggenomes(*114*) package.

900

901 **Identification of non-$anchor_{DOM}$ proteins co-distributing with anchoring systems in *Romboutsia***

902 All protein sequences encoded in the available *Romboutsia* and *Paeniclostridium* genome
903 assemblies (118), excluding those containing CWB2 (PF04122.15) and SLH (PF00395.23) anchoring domains
904 identified using hmmsearch (-E $1e^{-5}$), were analysed using protein similarity network analysis. Analysed
905 genomes were classified into two categories: CWB2-containing (23 assemblies encoding more than two
906 CWB2 proteins) and SLH-containing (65 assemblies encoding more than two SLH proteins). Non- $anchor_{DOM}$
907 proteins co-distributing with their respective $anchor_{DOM}$ systems (Supplementary Data 10) were identified
908 as protein networks containing sequences from more than 90% of the respective $anchor_{DOM}$ -containing
909 species (>21 CWB2 or >58 SLH) and no sequences from species containing the other $anchor_{DOM}$ system.
910
911 ***Clostridioides difficile* genome assemblies used in bioinformatic analyses**

912 Separate classification systems of *Clostridioides difficile* isolates are available(*37, 115*), including
913 those based on core genome phylogenies of universally conserved gene sequences most likely inherited
914 vertically, and classification based on the genomic cassettes shared between isolates *via* lateral gene
915 transfer. To account for these distinct classifications representing different evolutionary processes shaping
916 *C. difficile* isolate diversity, the following process was used to select representative genome samples. An
917 initial set of *C. difficile* genome assemblies was selected based on previous analyses(*37, 115-119*) so that all
918 major types, variants, and clades were included. This initial dataset was next expanded by including
919 assemblies encoding divergent SlpA sequences which were not present in the original dataset and
920 identified with BlastP and tBlastN(*105*) against the NCBI nr(*120*) and NCBI genome assemblies(*121*)
921 databases, resulting in the 141 genomes dataset (Supplementary Data 2). Next, the 141 genomes dataset
922 was further expanded by including a broad sample of genome assemblies from large-scale clinical isolate
923 sequencing project(*122*) downloaded from EnteroBase(*123, 124*), selected based on examination of their
924 clade- and SLCT-type phylogenies, resulting in the 866 genomes dataset (Supplementary Data 2). Within the
925 866 genomes dataset, no SlpA protein sequence was identified in a single assembly of *C. difficile*
926 NT949(*125*), which in the core genome species tree grouped (100% bootstrap support) with SlpA-positive *C.*
927 *difficile* CLO_CA2914AA_AS within clade1. However, tBlastN with SlpA from its closest relative revealed the
928 presence of a frame-shifted SlpA sequence (NCBI locus tag BM535_18170) annotated as a pseudogene. This
929 sequence is at the beginning of contig_808 (genebank MPDZ01000808.1), suggesting a possible incomplete
930 sequence or assembly rather than SlpA sequence loss. The only genome assembly that encoded 2 SlpA
931 sequences was *C. difficile* strain lsh09 (GCF_900688555.1), one of these sequences
932 (SAMEA1402364_01087) - as in other *C. difficile* genomes - is located in a 515,328 bp long contig
933 (CAADAX010000002.1) next to *secA2*, and the second partial sequence constituted of a short 412 bp
934 (CAADAX010000091) contig, suggesting it may be a contaminant. Finally, we verified that our dataset was
935 representative of known *C. difficile* diversity by analysing the complete set of 18,035 *C. diffcile*
936 (Supplementary Data 2) genome assemblies deposited in NCBI GenBank, where we identified 18,238 gene
937 sequences encoding SlpA homologues, including 845 non-redundant *slpA* gene sequences (Supplementary
938 Data 7) belonging to the $SlpA_{SLCT}$ groups (Supplementary Data 7) identified in our 866 genomes dataset. For
939 genome assemblies without available predicted gene sequences, encoded genes were predicted using
940 Prodigal(*126*). The SlpA homologues were identified using $surface_{DOM}$-only query sequences of all SLLP
941 types in tBlastN (-evalue 1e-5) searches. The dataset of all available NCBI *C. difficile* genome assemblies was
942 downloaded from the NCBI GenBank database[8,40] on 8th November 2023.
943
944 ***Clostridioides difficile* core genome species tree**

945 To identify a set of one-to-one universal orthologues for *C. difficile* core genome species tree
946 inference, an initial dataset of 866 genome assemblies sampled across all six clades and containing all SLCT
947 representatives was analysed using sequence similarity network analyses, resulting in 4,089 networks.

948 Network decomposition following walktrap analyses resulted in a dataset of networks including 605
949 networks of one-to-one orthologues with average sequence identities ranging between 85.4 % and 100 %.
950 Of these, the final set of 147 one-to-one orthologues (Supplementary Data 3) with sequence identity level
951 below 98% was used for core genome species tree inference. The identified sequences were aligned with
952 MAFFT(*89*), trimmed with trimAL (gappyout mode)(*90*) and resulting trimmed alignments were
953 concatenated with module Nexus of the Biopython package(*91*) to a final alignment of 866 concatenated
954 sequences of 37,580 amino acid sites, which was used to generate the *C. difficile* core genome species tree
955 (Supplementary Data 11) with IQ-TREE(*92*) under LG + G4 model with 1,000 ultrafast bootstraps(*93*).
956
957 **Identification *of Clostridioides difficile* orthogroups**

958 *Clostridioides difficile* orthogroups were identified using OrthoFinder(*127, 128*) with
959 DIAMOND(*129*) sequence homology search option. To remove redundancy and reduce computational
960 complexity for downstream analyses, the complete dataset of 866 genome assemblies was reduced to a
961 minimal representative subsample of 267 assemblies based on careful examination of universally
962 conserved orthologues and SlpA alignments and the core genome species tree and SlpA protein tree.
963 OrthoFinder assigned 1,004,302 genes (99.8% of the total) to 7,762 orthogroups (Supplementary Data 12).
964 Fifty percent of all genes were in orthogroups with 266 or more genes (G50 was 266) and were contained in
965 the largest 1,782 orthogroups (O50 was 1,782). There were 862 orthogroups with sequences from all
966 genome assemblies present and 609 of these consisted entirely of single-copy genes closely matching the
967 605 one-to-one orthologues identified in the full set of 866 genomes using protein similarity network
968 analysis. To account for possible incomplete genome assemblies and sequencing errors such as those
969 observed for slpA in *C. difficile* NT949 and strain lsh09 or rare cases of strain-specific gene loss and gain, we
970 expanded the set of the 862 one-to-one orthologues to include orthogroups containing zero or more than
971 one sequence in no more than 10% (27/267) genomes. This resulted in the final set of 2950 orthogroups
972 containing a single copy orthologue in at least 90% (240/267) *C. difficile* genomes (SCO90, Supplementary
973 Data 12). Mean sequence identities (Supplementary Data 12) were calculated as an average of trimAl(*90*)
974 pairwise identity comparisons (-sident) on MAFFT(*89*) alignments of all sequences within each orthogroup.
975
976 **Lateral gene transfer detection**

977 Lateral gene transfers were detected using the Duplication-Transfer-Loss reconciliation method
978 implemented in Ranger-DTL 2.0(*130*). *Ranger-DTL* analysis was performed using an optimally rooted SlpA
979 protein tree and a manually rooted species tree using default *Ranger-DTL* parameters in 100 replicates with
980 different random seeds. The results from the replicates were aggregated using *AggregateRanger* with the
981 default parameters. A single optimally rooted SlpA protein tree was identified using the *OptRoot* program
982 with default parameters. The position of the *C. difficile* core genome species tree root was selected based
983 on the published analyses(*131*), and our phylogenetic analyses of the available *Peptostreptococcaecae*
984 genomes. A high level of sequence divergence in the Low Molecular Weight domain (LMW, which includes
985 the $surface_{DOM}$) results in high levels of uncertainty in its amino acid sequence alignments and low support
986 values in inferred protein trees. To account for this uncertainty, we manually inspected species and protein
987 trees and selected only unambiguous LGT candidates from distantly related species (from distinct species
988 tree groupings), which grouped together in protein trees with high support**.**
989
990 **Detection of the amino acid sequence sites under adaptative evolution**

991 Initially, amino acid sequences encoded by the 972 non-redundant nucleotide *slpA* genes identified
992 across all 18,035 NCBI genomes, were aligned with MAFFT(*89*) and used to construct phylogeny with IQ-

993 TREE(*92*) under best fitting (according to AIC and BIC)(*132*) WAG+F+I+G4 model with 1,000 ultrafast
994 bootstraps(*93*). Due to the high level of sequence divergence, the accuracy of SlpA alignments within highly
995 variable regions (mainly within LMW) was low. To account for that, we separately analysed SlpAs belonging
996 to each SLCT category, where higher levels of sequence similarity enabled the generation of more reliable
997 alignments. In addition, this enabled the use of each SLCT as a replicate in testing trends conserved across
998 all SLCTs. SLCT types were extracted based on the SlpA protein tree topology. Amino acid sequences for
999 each SLCT type were aligned separately in MAFFT(*89*), and were used to generate corresponding nucleotide
1000 alignment with a local version of PAL2NAL_v14(*133*), with the nogap option to trimm alignment regions
1001 with gaps. ReadAl(*90*) (-phylip3.2 option) was used to convert fasta files to phylip format. The resulting
1002 trimmed nucleotide alignments were translated to amino acid alignments with EMBOSS transeq
1003 6.6.0.0(*134*). The trimmed amino acid alignments were used to reconstruct phylogenies for each SLCT with
1004 IQ-TREE(*92*) with the WAG+F+I+G4 model. Bootstrap values and branch lengths were removed for
1005 subsequent analyses in R using ape library(*135*). SLCT5 was excluded from the analysis as only five non-
1006 redundant sequences were available.
1007 The SLCT nucleotide alignments and corresponding protein phylogenies were used in CODEML analyses of
1008 PAML package(*136*) under M0 (one-ratio), and M1a (NearlyNeutral) and M2a (PositiveSelection) site
1009 models with heterogenous ω across the sites using following parameters: runmode = 0 (unrooted user
1010 tree), seqtype = 1 (codons), CodonFreq = 7 (FmutSel substitution model(*137*)), clock = 0 (no molecular clock
1011 assumption), model = 0 (sites models with homogenous ω across the branches). The different models were
1012 compared with the likelihood ratio test as described in the published protocol(*138*). Positively selected
1013 residues were identified in Bayes Empirical Bayes (BEB) analysis(*139*) with probability P > 95%.
1014 Recombination can complicate the interpretation of the dN/dS analyses, and although we
1015 identified only a single unambiguous example of recombination within our dataset, we used GARD to
1016 further account for possible recombination events(*140*). The relative simplicity of the GARD approach had
1017 the added benefit of accounting for the apparent heterogeneity in modes of evolution observed between
1018 the cell wall anchoring domains (e.g. CWB2) and the functional domains (e.g. $surface_{DOM}$) i.e. often the
1019 inferred breakpoints corresponded to the split between the two domains. Output files of the SLCT
1020 nucleotide alignment processed with GARD were used in all subsequent analyses (BUSTED, FUBAR, and
1021 MEME). FUBAR(*141*) was used to infer the positively (posterior probabilities > 90%) and negatively
1022 (posterior probabilities > 98.5%) selected sites. More inclusive thresholds were used for positive selection
1023 as it was verified with two additional methods (MEME and PAML), while more stringent thresholds were
1024 used for the negative selection identified using only a single method. BUSTED(*142*) was used to test for
1025 gene-wide episodic positive selection in at least one site on at least one branch while accounting for across
1026 the sites synonymous substitution rate variation(*143*). MEME was used to identify evidence of specific sites
1027 undergoing episodic diversifying selection (p-value < 0.1)(*144*). Finally, sites identified using all three
1028 methods (PAML, FUBAR, MEME) were mapped onto our experimentally determined *Cd*SlpA crystal
1029 structures and AlphaFold2(*145*) models of the remaining *Cd*SlpA homologues. Only sites identified with at
1030 least two of the three methods (PAML, FUBAR, MEME) were considered to display strong evidence of
1031 positive selection (Supplementary Data 13).
1032
1033 **Detection of recombination between *Clostridioides difficile* SlpAs**

1034 Evidence of recombination across the SlpAs sampled from all available *C. difficille* genomes was
1035 detected using RDP4.101 and RDP5 Beta 5.53(*146*), with no differences in the results obtained with each
1036 version. The 972 non-redundant nucleotide *slpA* genes identified across all 18,035 NCBI genomes and their
1037 corresponding MAFFT(*89*) amino acid alignment were used to generate nucleotide alignment with a local
1038 version of PAL2NAL_v14(*133*). The resulting nucleotide alignment was used as an input in a full exploratory

1039 recombination scan using RDP(*147*), GENECONV(*148*), Bootscan(*149*), Maxchi(*150*), Chimaera(*151*),
1040 SiSscan(*152*), and 3Seq(*153*) methods with default parameters. Of all 40 candidate recombination events
1041 detected using at least one of the seven methods, 9 candidate events were statistically significant in all
1042 seven methods, of which both inferred parental sequences were 'known' in four events (i.e. their
1043 sequences were present in the tested dataset). Of the four strongly supported recombination events with
1044 both parent sequences available, only a single unambiguous recombination event was identified following
1045 the manual verification within SlpA protein phylogeny and nucleotide sequence alignments of the parent
1046 and recombinant sequences . Recombinant sequence: GenBank=CP068558.1, locus_tag=JMW52_06215,
1047 protein_id=QQY54376.1; Major parent: GenBank=PRZC01000013.1, locus_tag=C4007_03100,
1048 protein_id=MDB2714594.1;
1049 Minor parent: GenBank=ABLTPH020000028.1, locus_tag=REJ96_016885, protein_id=MDS6367270.1.
1050 Significance values from RDP4 analysis: RDP (2.6 E-58), GENECONV (5.9 E-54), Bootscan (1.5 E-60), Maxchi
1051 (4.3 E-24), Chimaera (4.1 E-23), SiSscan (4.9 E-35), 3Seq (5.6 E-11).
1052
1053 **Other bioinformatics methods**

1054 Unless stated otherwise, protein phylogenies were inferred using sequences aligned with
1055 MAFFT(*89*) in IQ-TREE(*92*) with 1,000 ultrafast bootstraps(*93*) under best fitting model selected with
1056 ModelFinder(*132*) according to Bayesian information criterion implemented in IQ-TREE. Multiple sequence
1057 alignments were inspected using Jalview(*154*). Phylogenetic trees were inspected in Archaeopteryx(*155*).
1058 Other general bioinformatics packages used for data analyses and visualisation included: ggplot2(*156*),
1059 dplyr(*157*), ggmsa(*158*), ggtree(*159*), ape(*160*), treeio(*161*), ggtreeExtra(*162*), ggnewscale(*163*).
1060
1061 **General microbiology methods**

1062 *Peptostreptococaccceae* strains were selected based on phylogenetic analyses and obtained from
1063 reference culture collection (DSMZ, German Collection of Microorganisms and Cell Cultures GmbH) or
1064 individual laboratory sources. All bacterial strains and specific culture conditions used in this study are
1065 described in Supplementary Table 2. Bacteria were cultured in DG250 or A35 workstations (Don Whitley
1066 Scientific) in anaerobic conditions (10% H_2, 10% CO_2, 80% N_2).
1067
1068 **Bacterial growth**

1069 To investigate the presence of S-layer and expression of S-layer proteins in different culture
1070 conditions, all strains were cultured in strain-specific media following DSMZ strain collection protocols
1071 (Supplementary Table 2) and three media commonly used for the model *Peptostreptococcacea* bacterium
1072 *C. difficile*: TY (3% tryptose, 2% yeast extract), BHIS (brain heart infusion, 0.5% yeast extract, 0.1% L-
1073 cysteine), and chemically defined minimal medium (CDMM(*164*)). Bacterial growth was tested at 34 °C and
1074 37 °C for all media and strains, and additionally at 28 °C for *C. mangenotii* LM2. All cultures were started by
1075 inoculating fresh media with overnight pre-cultures, in Hungate-type tubes to an OD_{600nm} of 0.08 measured
1076 in WPA CO8000 Cell Density Meter. To test the presence of S-layer and SLLPs at different growth phases,
1077 samples were taken at early-, middle-, and late-exponential, and early-, and late-stationary phases time
1078 points and OD_{600nm} values indicated in the results. Supplementation of growth media with $CaCl_2$ to a final
1079 concentration of 500 μM was carried out to assess the impact of calcium on S-layer proteins in culture.
1080
1081 **CRISPR interference for targeted knockdown of *sllp* gene expression**
1082 *E. coli-C. difficile* shuttle vector for CRISPR interference, pIA33(*165*), kindly provided by Prof. Craig
1083 Ellermeier, was used to target expression of genes encoding identified SLLP in *R. lituseburensis* and *C.*

1084 *difficile*. The sgRNA was analysed using CASPER(*166*) (https://github.com/TrinhLab/CASPERapp), generated
1085 by amplification using oligonucleotide pairs: pIA33_ SLLP_2_Rlitu_F and pIA33_Insert_R; pIA33_SlpA1_F
1086 and pIA33_Insert_R listed in Supplementary Table 3 using pIA33 template, to substitute sgRNA-rfp, and
1087 introduced into MscI+PstI-digested pIA33 backbone by isothermal assembly(*167*). Negative-control plasmid
1088 (pIA34(*165*)) containing non-targeting sgRNA sequence that does not recognize any sequence in the
1089 genome, was used to control for non-specific plasmid effect. CRISPRi::*sll*p plasmids were introduced into
1090 *Peptostreptococacceae* host by conjugation(*168*). Transformants were selected on TY thiamphenicol 10 µg
1091 mL^{-1} agar, supplemented with cycloserine (250 µg mL^{-1}).
1092 Induction of dCas9-opt activity was mediated by supplementation of growth media with xylose (1%
1093 w/v) and the interference was observed for *C. difficile slpA* (as shown before(*165*)) and for *R. lituseburensis*
1094 *sllp*, as demonstrated by SDS-PAGE of surface layer extracts, RT-qPCR, and electron microscopy.
1095 Viability of CRISPRi::*sllp* knockdown strains was assessed by spot assay, where a 5 µl aliquots of
1096 serial 10-fold dilutions of xylose induced and non-induced overnight cultures, were spotted onto TY-agar.
1097 Following 24-hour incubation, CFU mL^{-1} were enumerated.
1098 A RT-qPCR was performed to confirm gene silencing, using oligonucleotide pairs:
1099 qCdR20291_SlpA_F and qCdR20291_SlpA_R; qCdR20291_RpoC_F and qCdR20291_RpoC_R;
1100 qCdR20291_GyrA_F and qCdR20291_GyrA_R; qRlitu_SLLP_F and qRlitu_SLLP_R; qRlitu_RpoC_F and
1101 qRlitu_RpoC_R; qRlitu_GyrA_F and qRlitu_GyrA_R, designed with PrimerQuest (Integrated DNA
1102 Technologies) listed in Supplementary Table 3. RNA extraction was carried out with Monarch Total RNA
1103 Miniprep kit (NEB) from 8 h cultures. DNAse treatment was carried out with TURBO DNA-free™ Kit
1104 (Invitrogen). cDNA preparation was performed using the SuperScript III VILO cDNA synthesis kit. RT-qPCR
1105 was performed using Luna Universal qPCR master Mix (NEB) in LightCycler96 (Roche). Samples of two
1106 biological replicates were analysed in 3 technical replicates. Results were calculated by the comparative
1107 cycle threshold method(*169*) and normalized to the gyrA transcript. Relative expression of *sllp* and *rpoC*
1108 control was presented as a log2 fold change, relative to non-induced CRISPRi::*sllp* expression level.
1109
1110 **Broth-based killing assay**
1111 Human innate immune effector (lysozyme; 500 µg mL^{-1}, or LL-37; 5 µg mL^{-1}) was added to mid-
1112 exponential cultures (OD_{600nm} ~0.5) and its impact on the growth of strains expressing intact SLLP or of
1113 CRISPRi *sllp* knock-down was assessed by monitoring OD_{600nm}. Two biological and three technical replicates
1114 were analysed in microplate reader (Stratus, Cerillo). OD change was estimated from a difference between
1115 OD_{600nm} at endpoint (8 h timepoint for *C. difficile* and 5 h for other *Peptostreptococacceae*) and OD_{600nm}
1116 measured upon supplementation with the additive.
1117
1118 **Sporulation assay**
1119 CRISPRi::*sllp* knock-down strains of *C. difficile* and *R. lituseburensis* presented sporulation defects
1120 compared to negative control when assessed by electron microscopy of whole cells. To estimate
1121 sporulation efficiency, the knock-down strains were cultured overnight, in biological duplicate. Aliquots of
1122 50 µl were spotted onto selective TY-agar supplemented or non-supplemented with xylose and cultured for
1123 24 h. Growth was collected from the plate and resuspended in 500 µl PBS. Cell suspension was split and
1124 one aliquot of 250 µl was heat-treated at 70 °C for 20 min. Both non-treated and heat-treated aliquots
1125 were serially 10-fold diluted in PBS, and 5 µl aliquots of dilutions were spotted onto TY, 0.1% taurocholate,
1126 0.1% cysteine agar plates. Following overnight incubation, CFU mL^{-1} was enumerated. The heat-resistant
1127 CFUs are represented as a percentage of total viable (non-treated) CFUs.
1128
1129 **Transmission electron microscopy (TEM) of bacterial cell sections**

1130 Stationary phase bacterial cultures were harvested by centrifugation at 2,000 x*g* for 10 min at room
1131 temperature. Cell pellets were collected in 1 ml of fixative (2% glutaraldehyde in 0.1 M sodium cacodylate
1132 buffer, pH 7.4) and stored at 4 °C until further processing, performed at Newcastle University Electron
1133 Microscopy Research Services (EMRS). Samples were enrobed in low melting point agarose and post-fixed
1134 in 1% osmium tetroxide, dehydrated in gradient acetone, and embedded in epoxy resin.
1135 Cell surface glycopolymers were visualised using a published protocol(*47*). Briefly, bacterial colonies
1136 were collected from 24 h agar plate cultures and pre-fixed with 2% glutaraldehyde and 10 mM L-lysine in
1137 CB-RR (0.1 M sodium cacodylate buffer, pH 7.4, 0.1% ruthenium red) for 10 min at room temperature.
1138 Subsequently, pellets were fixed in 2% glutaraldehyde in CB-RR for 2 h at room temperature, washed with
1139 CB-RR, and post-fixed in 1% osmium tetroxide in CB-RR. Samples were then enrobed in agarose,
1140 dehydrated in a gradient of ethanol, and embedded in the polyhydroxy aromatic acrylic resin LR white.
1141 Ultrathin sections of around 70 nm were deposited on copper grids and stained with uranyl acetate
1142 and lead citrate. Grids were then visualised using a Hitachi HT7800 120 kV transmission electron
1143 microscope (EMRS, Newcastle University).
1144 Line profile analysis of cell envelope across cell sections was performed with Fiji(*170*).
1145
1146 **Electron microscopy of isolated S-layer lattice fragments**

1147 S-layer preparation for electron microscopy was performed as previously described(*40*) with minor
1148 modifications. Mid-exponential cultures were harvested by centrifugation at 2,000 x*g* for 15 min at room
1149 temperature. Cell pellets were washed twice in ice-cold deionised water (0.25x of culture volume) and
1150 disrupted by pulse-sonication for 2 min on ice. Unbroken cells were removed by centrifugation at 800 x*g* for
1151 10 minutes at 4 °C and the supernatant was further centrifuged at 3,000 x*g* for 10 minutes. Harvested
1152 pellets containing S-layer fragments were washed with ice-cold 1 M NaCl (to an effective OD_{600nm} of 100, of
1153 initial culture) and resuspended in cold 2% Triton X-100 (to an effective OD_{600nm} of 200, of initial culture).
1154 Preparations were loaded on glow-discharged, carbon-coated 300 mesh copper EM grids (Agar Scientific),
1155 stained with 2% uranyl acetate, and assessed using a Hitachi HT7800 transmission electron microscope
1156 operating at 120 kV (EMRS, Newcastle University).
1157 For electron cryo-microscopy, S-layer fragments were diluted 10-fold in cold deionised water and
1158 added onto plasma-treated Quantifoil Cu R2/2, 200 mesh grids. The grids were blotted for 5 seconds at 4
1159 °C, 90% relative humidity, vitrified in liquid ethane, using Vitrobot (FEI), and stored in liquid nitrogen.
1160 Cryo-EM grid preparation and imaging were performed at the York Structural Biology Laboratory
1161 (YSBL) on a Glacios electron microscope equipped with a Falcon 4 direct electron detector (Thermo Fisher
1162 Scientific). Micrographs were taken at 200 kW with a magnification of 120,000 and -2µm defocus.
1163 Images were viewed and Fourier transforms were generated using 3dMOD(*171*) and Fiji(*170*).

1164
1165 **Identification of S-layer lattice proteins**

1166 As the main building blocks of known bacterial S-layers, SLLPs typically constitute one of the most
1167 abundant cellular polypeptides and the dominant protein in S-layers, aiding their identification from
1168 dominant bands in gel electrophoresis of S-layer preparations.
1169 To identify the *Peptostreptococcaceae* SLLPs, bacteria with evidence of S-layer presence in electron
1170 micrographs of S-layer fragments and whole cell sections were cultured to stationary phase and harvested
1171 by centrifugation at 4,000 x*g* for 10 min at room temperature. Additional experiments with sampling from
1172 across the growth cycle were performed for *C. mangenotii* LM2, which contains both CWB2-SLLP and
1173 CWB2-EA1, and for the strains where no evidence of S-layer was found in the electron micrographs.
1174 Harvested wet cell pellets were washed with a HEPES-NaCl buffer (50 mM HEPES, 100 mM NaCl, pH 7.5),

1175 followed by a critical step of the careful removal of all HEPES-NaCl liquid to ensure consistent extraction
1176 conditions in the subsequent steps. Non-covalently bound proteins, including S-layer proteins, were
1177 extracted from the cell surface using three different chemical methods. Preparations using 0.2 M glycine pH
1178 2.2 were carried out as previously described(*40*). Two additional methods of extraction: 6 M urea, 50 mM
1179 Tris pH 8.0, 1 mM phenyl-methyl-sulphonide-fluoride (PMSF); and 1% sodium dodecyl sulphate (SDS), 50
1180 mM Tris–HCl pH 8.0, 1 mM PMSF were used following the published protocols(*172, 173*). For each method,
1181 the washed cell pellets were resuspended in extraction solution to an effective OD_{600nm} of 50 and incubated
1182 at room temperature for 30 min, followed by centrifugation at 21,000 x*g* for 10 min. A 2 M Tris base was
1183 used to neutralise the pH of low-pH glycine preparations. Extracts were then mixed with a 0.25x volume of
1184 sample buffer (8% SDS, 400 mM DTT, 40% glycerol, 200 mM Tris-HCl pH 6.8, 0.4% bromophenol blue) and
1185 separated via 6-20% gradient SDS-polyacrylamide gel electrophoresis (SDS-PAGE).
1186 Recovered spent media were used to assess secreted protein content and cell autolysis, by
1187 precipitation with 10% (v/v) of trichloroacetic acid (TCA), for 30 min on ice. Following centrifugation at
1188 21,000 x*g* for 10 min at 4 °C, the precipitate was washed 3 times with ice-cold 90% acetone. The resulting
1189 air-dried pellet was resuspended in PBS to an effective OD_{600nm} of 50, mixed with a 0.25x volume of sample
1190 buffer (8% SDS, 400 mM dithiothreitol (DTT), 40% glycerol, 200 mM Tris-HCl pH 6.8, 0.4% bromophenol
1191 blue) and separated via 6-20% gradient SDS-PAGE.
1192 To identify proteins in S-layer preparations, trypsin digests of the respective bands excised from
1193 SDS-polyacrylamide gels were analysed by LC-MS/MS in positive ion mode on a Finnigan LTQ ion trap mass
1194 spectrometer (Thermo Fisher Scientific, Bremen, Germany). Protein identification was performed using the
1195 Mascot MS/MS search engine program (Matrix Science Ltd, London) against a custom protein sequence
1196 database (*Peptostreptococcaceae*, containing 179,609 unique sequences created on 26/01/2023 from
1197 UniProt).
1198
1199 **X-ray crystallography of SlpA**

1200 TY broth cultures (16 h) of *C. difficile* strains were centrifuged at room temperature at 5,000 x*g* and
1201 the resulting pellets were washed with 0.1x culture volume of 50 mM M HEPES pH 7.4, 150 mM NaCl
1202 buffer. S-layer extraction was performed by resuspending the washed pellet in 0.01x culture volume of 0.2
1203 M glycine-HCl pH 2.2, as described above. S-layer extract was filtered and resolved onto a Superdex 200
1204 26/600 column using an ÄKTA Pure FPLC system (Cytiva) in 50 mM Tris-HCl pH 7.5, 150 mM NaCl buffer.
1205 Purified proteins were concentrated to 10 mg mL^{-1} and subjected to crystallization with the sitting drop
1206 vapour-diffusion method at 20 °C with a range of commercial screens.
1207 Although $SlpA_{SLCT11}$ was purified from Ox247 (wild type strain(*37*)) and Ox247Δorf2 (mutant
1208 defective in glycosylation of SlpA(*174*), denoted SLCT11*), only the non-glycosylated protein resulted in
1209 successful crystallisation in 4 mM oxometalates, 50 mM Tris (base); 50 mM bicine 8.5, 20% v/v glycerol;
1210 10% w/v PEG 4000 (Morpheus Fusion, Molecular Dimensions). $SlpA_{SLCT2}$ purified from Opt2472 strain(*122*)
1211 produced diffraction quality crystals in 1.6 M Sodium citrate tribasic (Wizard Classic 3, Molecular
1212 Dimensions).
1213 Diffraction data were collected on I24 beamline (λ = 0.71 Å) at the Diamond Light Source
1214 Synchrotron (Didcot, UK; mx24948-136) at 100 K. The data was processed with xia2(*175*) ($SlpA_{SLCT11*}$) or
1215 DIALS(*176*) ($SlpA_{SLCT2}$), scaled with Aimless(*177*) within CCP4 Cloud(*178*). Initial phases were obtained by
1216 molecular replacement in Phaser(*179*) with models of the CWB2 anchoring domain using the $SlpA_{CD630}$ (PDB
1217 ID: 7ACY) template for $SlpA_{SLCT11*}$ or the AlphaFold2(*145*) predicted model for $SlpA_{SLCT2}$. The generated
1218 solution model was then subjected to automatic model building with ModelCraft(*180*), followed by
1219 iterative cycles of manual building with Coot(*181*) and refinement in Refmac5(*182*).

Final models were obtained after iterative cycles of manual model building with Coot and refinement in Phenix_refine(*183*) and PDBREDO(*184*). Validation of final models was performed using Coot and Phenix internal tools, as well as MolProbity(*185*) web server. Data collection and refinement statistics are summarized in Supplementary Table 4.

Other structural analyses

Structural prediction for proteins of interest was performed with ColabFold v1.5.5: AlphaFold2 using MMseqs2(*107*). Comparative structural searches were carried out with Foldseek(*186*) and Dali(*187*) servers. Structural analyses including crystal contacts, distances, angles and dimensions were performed with PDBSum(*188*), PISA(*189*) and with plugins implemented in PyMOL Molecular Graphics System (Schrödinger, LLC) or UCSF ChimeraX(*190*). Graphical representations of S-layer proteins and crystal latices were generated using UCSF ChimeraX. SymProFold pipeline was used to predict S-layer lattice assemblies(*35*) following the standard protocol (https://github.com/symprofold/SymProFold_Tutorial_Data). AlphaFold3 server, using the PDB2025 (up to 3 February 2025) template option, was used to generate structural predictions for full-length and surfaceDOMs models for experimentally identified SLLPs across *Peptotreptococcaceae.* Individual subdomains – $anchor_{DOM}$, $surface_{DOM}$, D1, D2 and ID - were manually extracted from the ful-length predictions and aligned using the LSQ function in COOT. Calculated RMSD for pairwise superimpositions in Coot were represented as a heatmap using Rstudio.

1240 **References**

1241

1242 1. P. Messner, U. B. Sleytr, Crystalline bacterial cell-surface layers. *Adv Microb Physiol* **33**, 213-275
1243 (1992).
1244 2. T. A. M. Bharat, A. von Kugelgen, V. Alva, Molecular Logic of Prokaryotic Surface Layer Structures.
1245 *Trends Microbiol* **29**, 405-415 (2021).
1246 3. D. Pum, J. L. Toca-Herrera, U. B. Sleytr, S-layer protein self-assembly. *Int J Mol Sci* **14**, 2484-2501
1247 (2013).
1248 4. P. Messner, C. Schäffer, E.-M. Egelseer, U. B. Sleytr, in *Prokaryotic Cell Wall Compounds: Structure*
1249 *and Biochemistry,* H. König, H. Claus, A. Varma, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg,
1250 2010), pp. 53-109.
1251 5. R. P. Fagan, N. F. Fairweather, Biogenesis and functions of bacterial S-layers. *Nat Rev Microbiol* **12**,
1252 211-222 (2014).
1253 6. D. Missiakas, O. Schneewind, Assembly and Function of the Bacillus anthracis S-Layer. *Annu Rev*
1254 *Microbiol* **71**, 79-98 (2017).
1255 7. W. J. Bradshaw, A. K. Roberts, C. C. Shone, K. R. Acharya, The structure of the S-layer of Clostridium
1256 difficile. *J Cell Commun Signal* **12**, 319-331 (2018).
1257 8. C. Trautner, W. F. Vermaas, The sll1951 gene encodes the surface layer protein of Synechocystis sp.
1258 strain PCC 6803. *J Bacteriol* **195**, 5370-5380 (2013).
1259 9. I. Lasa, J. R. Caston, L. A. Fernandez-Herrero, M. A. de Pedro, J. Berenguer, Insertional mutagenesis
1260 in the extreme thermophilic eubacteria Thermus thermophilus HB8. *Mol Microbiol* **6**, 1555-1564
1261 (1992).
1262 10. J. A. Kirk *et al.*, New class of precision antimicrobials redefines role of Clostridium difficile S-layer in
1263 virulence and viability. *Sci Transl Med* **9**, (2017).
1264 11. A. Fioravanti, M. Mathelie-Guinlet, Y. F. Dufrene, H. Remaut, The Bacillus anthracis S-layer is an
1265 exoskeleton-like structure that imparts mechanical and osmotic stabilization to the cell wall. *PNAS*
1266 *Nexus* **1**, pgac121 (2022).
1267 12. S. Ali *et al.*, Slr4, a newly identified S-layer protein from marine Gammaproteobacteria, is a major
1268 biofilm matrix component. *Mol Microbiol* **114**, 979-990 (2020).
1269 13. L. L. Wong *et al.*, Surface-layer protein is a public-good matrix exopolymer for microbial community
1270 organisation in environmental anammox biofilms. *ISME J* **17**, 803-812 (2023).
1271 14. N. Shimotahira *et al.*, The surface layer of Tannerella forsythia contributes to serum resistance and
1272 oral bacterial coaggregation. *Infect Immun* **81**, 1198-1206 (2013).
1273 15. J. W. Kern, O. Schneewind, BslA, a pXO1-encoded adhesin of Bacillus anthracis. *Mol Microbiol* **68**,
1274 504-515 (2008).
1275 16. M. M. Merrigan *et al.*, Surface-layer protein A (SlpA) is a major contributor to host-cell adherence
1276 of Clostridium difficile. *PLoS One* **8**, e78404 (2013).
1277 17. J. Sakakibara *et al.*, Loss of adherence ability to human gingival epithelial cells in S-layer protein-
1278 deficient mutants of Tannerella forsythensis. *Microbiology (Reading)* **153**, 866-876 (2007).
1279 18. R. D. Plaut *et al.*, Genetic evidence for the involvement of the S-layer protein gene sap and the
1280 sporulation genes spo0A, spo0B, and spo0F in Phage AP50c infection of Bacillus anthracis. *J*
1281 *Bacteriol* **196**, 1143-1154 (2014).
1282 19. A. L. M. Royer *et al.*, Clostridioides difficile S-Layer Protein A (SlpA) Serves as a General Phage
1283 Receptor. *Microbiol Spectr* **11**, e0389422 (2023).
1284 20. M. Sara, U. B. Sleytr, Molecular sieving through S layers of Bacillus stearothermophilus strains. *J*
1285 *Bacteriol* **169**, 4092-4098 (1987).
1286 21. A. von Kugelgen *et al.*, Membraneless channels sieve cations in ammonia-oxidizing marine archaea.
1287 *Nature*, (2024).
1288 22. U. B. Sleytr, E. M. Egelseer, N. Ilk, D. Pum, B. Schuster, S-Layers as a basic building block in a
1289 molecular construction kit. *FEBS J* **274**, 323-334 (2007).
1290 23. B. An *et al.*, Engineered Living Materials For Sustainability. *Chem Rev* **123**, 2349-2419 (2023).
1291 24. Y. Yuliawati *et al.*, Potency of surface layer protein from Lactobacillus sp. as drug nanocarriers.
1292 *Journal of Applied Pharmaceutical Science*, (2024).

1293 25. B. Zhang *et al.*, Display of porcine epidemic diarrhea virus spike protein B-cell linear epitope on
1294 Lactobacillus mucosae G01 S-layer surface induce a robust immunogenicity in mice. *Microb Cell*
1295 *Fact* **23**, 142 (2024).
1296 26. A. Decout *et al.*, Lactobacillus crispatus S-layer proteins modulate innate immune response and
1297 inflammation in the lower female reproductive tract. *Nat Commun* **15**, 10879 (2024).
1298 27. J. Li *et al.*, Exopolysaccharides and Surface-Layer Proteins Expressed by Biofilm-State
1299 Lactiplantibacillus plantarum Y42 Play Crucial Role in Preventing Intestinal Barrier and Immunity
1300 Dysfunction of Balb/C Mice Infected by Listeria monocytogenes ATCC 19115. *J Agric Food Chem* **72**,
1301 8581-8594 (2024).
1302 28. S. R. Konstantinov *et al.*, S layer protein A of Lactobacillus acidophilus NCFM regulates immature
1303 dendritic cell and T cell functions. *Proc Natl Acad Sci U S A* **105**, 19474-19479 (2008).
1304 29. H. Wang, L. Zhang, Q. Li, S. Xu, R. Lu, Surface-layer protein produced by Lactobacillus crispatus JCM
1305 2009 ameliorates lipopolysaccharide-induced inflammation through autophagy cross-talk with the
1306 NF-kappaB signaling pathway. *Int J Biol Macromol* **166**, 633-640 (2021).
1307 30. M. J. Ormsby *et al.*, An intact S-layer is advantageous to Clostridioides difficile within the host. *PLoS*
1308 *Pathog* **19**, e1011015 (2023).
1309 31. Y. T. Wang, S. Y. Oh, A. P. Hendrickx, J. M. Lunderberg, O. Schneewind, Bacillus cereus G9241 S-
1310 layer assembly contributes to the pathogenesis of anthrax-like disease in mice. *J Bacteriol* **195**, 596-
1311 605 (2013).
1312 32. A. Fioravanti *et al.*, Structure of S-layer protein Sap reveals a mechanism for therapeutic
1313 intervention in anthrax. *Nat Microbiol* **4**, 1805-1814 (2019).
1314 33. L. Hunault *et al.*, Anti-S-layer monoclonal antibodies impact Clostridioides difficile physiology. *Gut*
1315 *Microbes* **16**, 2301147 (2024).
1316 34. E. Johnston, B. Isbilir, V. Alva, T. A. M. Bharat, J. P. K. Doye, Punctuated and continuous structural
1317 diversity of S-layers across the prokaryotic tree of life. PREPRINT (Version 1) available at. *bioRxiv*,
1318 2024.2005.2028.596244 (2024).
1319 35. C. Buhlheller *et al.*, SymProFold: Structural prediction of symmetrical biological assemblies. *Nat*
1320 *Commun* **15**, 8152 (2024).
1321 36. A. Sogues *et al.*, Cryo-EM structure and polar assembly of the PS2 S-layer of Corynebacterium
1322 glutamicum. *bioRxiv*, (2024).
1323 37. K. E. Dingle *et al.*, Recombinational switching of the Clostridium difficile S-layer and a novel
1324 glycosylation gene cluster revealed by large-scale whole-genome sequencing. *J Infect Dis* **207**, 675-
1325 686 (2013).
1326 38. D. Ni Eidhin, A. W. Ryan, R. M. Doyle, J. B. Walsh, D. Kelleher, Sequence and phylogenetic analysis
1327 of the gene for surface layer protein, slpA, from 14 PCR ribotypes of Clostridium difficile. *J Med*
1328 *Microbiol* **55**, 69-83 (2006).
1329 39. X. Long *et al.*, Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates
1330 tumour immunity. *Nat Microbiol* **4**, 2319-2330 (2019).
1331 40. P. Lanzoni-Mangutchi *et al.*, Structure and assembly of the S-layer in C. difficile. *Nat Commun* **13**,
1332 970 (2022).
1333 41. W. H. Lewis, G. Tahon, P. Geesink, D. Z. Sousa, T. J. G. Ettema, Innovations to culturing the
1334 uncultured microbial majority. *Nat Rev Microbiol* **19**, 225-240 (2021).
1335 42. A. Silale *et al.*, Dual function of OmpM as outer membrane tether and nutrient uptake channel in
1336 diderm Firmicutes. *Nat Commun* **14**, 7152 (2023).
1337 43. P. Messner, F. Hollaus, U. B. Sleytr, Paracrystalline Cell Wall Surface Layers of Different Bacillus
1338 stearothermophilus Strains. *International Journal of Systematic Bacteriology* **34**, 202-210 (1984).
1339 44. A. Usenik *et al.*, The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the
1340 Clostridium difficile Cell Wall Proteins Cwp8 and Cwp6. *Structure* **25**, 514-521 (2017).
1341 45. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3.
1342 *Nature* **630**, 493-500 (2024).
1343 46. S. E. Willing *et al.*, Clostridium difficile surface proteins are anchored to the cell wall using CWB2
1344 motifs that recognise the anionic polymer PSII. *Mol Microbiol* **96**, 596-608 (2015).
1345 47. M. Chu *et al.*, A Clostridium difficile Cell Wall Glycopolymer Locus Influences Bacterial Shape,
1346 Polysaccharide Production and Virulence. *PLoS Pathog* **12**, e1005946 (2016).

1347 48. A. Schmidt, N. Muller, B. Schink, D. Schleheck, A proteomic view at the biochemistry of syntrophic
1348 butyrate oxidation in Syntrophomonas wolfei. *PLoS One* **8**, e56905 (2013).
1349 49. B. A. Pinheiro *et al.*, Functional insights into the role of novel type I cohesin and dockerin domains
1350 from Clostridium thermocellum. *Biochem J* **424**, 375-384 (2009).
1351 50. L. Artzi *et al.*, Cellulosomics of the cellulolytic thermophile Clostridium clariflavum. *Biotechnol*
1352 *Biofuels* **7**, 100 (2014).
1353 51. F. F. Hager *et al.*, Functional Characterization of Enzymatic Steps Involved in Pyruvylation of
1354 Bacterial Secondary Cell Wall Polymer Fragments. *Front Microbiol* **9**, 1356 (2018).
1355 52. S. Y. Oh, J. M. Lunderberg, A. Chateau, O. Schneewind, D. Missiakas, Genes Required for Bacillus
1356 anthracis Secondary Cell Wall Polysaccharide Synthesis. *J Bacteriol* **199**, (2017).
1357 53. R. P. Fagan, N. F. Fairweather, Clostridium difficile has two parallel and essential Sec secretion
1358 systems. *J Biol Chem* **286**, 27483-27493 (2011).
1359 54. C. M. Ausiello *et al.*, Surface layer proteins from Clostridium difficile induce inflammatory and
1360 regulatory cytokines in human monocytes and dendritic cells. *Microbes Infect* **8**, 2640-2646 (2006).
1361 55. A. Ryan *et al.*, A role for TLR4 in Clostridium difficile infection and the recognition of surface layer
1362 proteins. *PLoS Pathog* **7**, e1002076 (2011).
1363 56. A. J. Waligora *et al.*, Characterization of a cell surface protein of Clostridium difficile with adhesive
1364 properties. *Infect Immun* **69**, 2144-2153 (2001).
1365 57. Q. Zhou *et al.*, The cwp66 Gene Affects Cell Adhesion, Stress Tolerance, and Antibiotic Resistance in
1366 Clostridioides difficile. *Microbiol Spectr* **10**, e0270421 (2022).
1367 58. N. Promadej, F. Fiedler, P. Cossart, S. Dramsi, S. Kathariou, Cell wall teichoic acid glycosylation in
1368 Listeria monocytogenes serotype 4b requires gtcA, a novel, serogroup-specific gene. *J Bacteriol*
1369 **181**, 418-425 (1999).
1370 59. A. Chateau *et al.*, Distinct Pathways Carry Out alpha and beta Galactosylation of Secondary Cell
1371 Wall Polysaccharide in Bacillus anthracis. *J Bacteriol* **202**, (2020).
1372 60. J. Rismondo, M. G. Percy, A. Grundling, Discovery of genes required for lipoteichoic acid
1373 glycosylation predicts two distinct mechanisms for wall teichoic acid glycosylation. *J Biol Chem* **293**,
1374 3293-3306 (2018).
1375 61. T. Endo, K. Ikeo, T. Gojobori, Large-scale search for genes on which positive selection may operate.
1376 *Mol Biol Evol* **13**, 685-690 (1996).
1377 62. G. Blanc *et al.*, Molecular evolution of rickettsia surface antigens: evidence of positive selection.
1378 *Mol Biol Evol* **22**, 2073-2083 (2005).
1379 63. L. Hunault *et al.*, A monoclonal antibody collection for C. difficile typing ? *Gut Pathog* **16**, 4 (2024).
1380 64. H. Chandra, R. A. Kovall, J. S. Yadav, X. Sun, Host Immune Responses to Surface S-Layer Proteins
1381 (SLPs) of Clostridioides difficile. *Microorganisms* **11**, (2023).
1382 65. M. Sara, U. B. Sleytr, Comparative studies of S-layer proteins from Bacillus stearothermophilus
1383 strains expressed during growth in continuous culture under oxygen-limited and non-oxygen-
1384 limited conditions. *J Bacteriol* **176**, 7182-7189 (1994).
1385 66. S. Avall-Jaaskelainen, A. Palva, Lactobacillus surface layers and their applications. *FEMS Microbiol*
1386 *Rev* **29**, 511-529 (2005).
1387 67. A. D. Moore, A. K. Bjorklund, D. Ekman, E. Bornberg-Bauer, A. Elofsson, Arrangements in the
1388 modular evolution of proteins. *Trends Biochem Sci* **33**, 444-451 (2008).
1389 68. J. B. Pereira-Leal, E. D. Levy, S. A. Teichmann, The origins and evolution of functional modules:
1390 lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci* **361**, 507-517 (2006).
1391 69. A. von Kugelgen *et al.*, Interdigitated immunoglobulin arrays form the hyperstable surface layer of
1392 the extremophilic bacterium Deinococcus radiodurans. *Proc Natl Acad Sci U S A* **120**, e2215808120
1393 (2023).
1394 70. T. Sagmeister *et al.*, The molecular architecture of Lactobacillus S-layer: Assembly and attachment
1395 to teichoic acids. *Proc Natl Acad Sci U S A* **121**, e2401686121 (2024).
1396 71. T. V. Tikhonova *et al.*, Extracellular Fe(III) reductase structure reveals a modular organization
1397 enabling S-layer insertion and electron transfer to insoluble substrates. *Structure* **31**, 174-184 e173
1398 (2023).
1399 72. W. W. Navarre, O. Schneewind, Surface proteins of gram-positive bacteria and mechanisms of their
1400 targeting to the cell wall envelope. *Microbiol Mol Biol Rev* **63**, 174-229 (1999).

1401 73. R. Lopez, E. Garcia, P. Garcia, J. L. Garcia, The pneumococcal cell wall degrading enzymes: a
1402 modular design to create new lysins? *Microb Drug Resist* **3**, 199-211 (1997).
1403 74. D. Megrian, N. Taib, A. L. Jaffe, J. F. Banfield, S. Gribaldo, Ancient origin and constrained evolution
1404 of the division and cell wall gene cluster in Bacteria. *Nat Microbiol* **7**, 2114-2127 (2022).
1405 75. D. Megrian, N. Taib, J. Witwinowski, C. Beloin, S. Gribaldo, One or two membranes? Diderm
1406 Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol Microbiol* **113**, 659-671 (2020).
1407 76. J. Witwinowski *et al.*, An ancient divide in outer membrane tethering systems in bacteria suggests a
1408 mechanism for the diderm-to-monoderm transition. *Nat Microbiol* **7**, 411-422 (2022).
1409 77. I. Lagkouvardos *et al.*, The Mouse Intestinal Bacterial Collection (miBC) provides host-specific
1410 insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol* **1**, 16131
1411 (2016).
1412 78. G. K. A. Hochberg *et al.*, A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503-
1413 508 (2020).
1414 79. J. Lukes, J. M. Archibald, P. J. Keeling, W. F. Doolittle, M. W. Gray, How a neutral evolutionary
1415 ratchet can build cellular complexity. *IUBMB Life* **63**, 528-537 (2011).
1416 80. B. Isbilir, A. Yeates, V. Alva, T. A. M. Bharat, Mapping the ultrastructural topology of the
1417 corynebacterial cell surface. *bioRxiv*, (2024).
1418 81. K. Roberts, Crystalline glycoprotein cell walls of algae: their stucture, composition and assembly.
1419 *Philos Trans R Soc Lond B Biol Sci* **268**, 129-146 (1974).
1420 82. P.-H. Baudelet, G. Ricochon, M. Linder, L. Muniglia, A new insight into cell walls of Chlorophyta.
1421 *Algal Research* **25**, 333-371 (2017).
1422 83. T. Rodrigues-Oliveira *et al.*, Actin cytoskeleton and complex cell architecture in an Asgard archaeon.
1423 *Nature* **613**, 332-339 (2023).
1424 84. H. Imachi *et al.*, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519-
1425 525 (2020).
1426 85. D. L. Sexton *et al.*, The cell envelope of Thermotogae suggests a mechanism for outer membrane
1427 biogenesis. *Proc Natl Acad Sci U S A* **120**, e2303275120 (2023).
1428 86. A. Hashimi, E. I. Tocheva, Cell envelope diversity and evolution across the bacterial tree of life. *Nat*
1429 *Microbiol* **9**, 2475-2487 (2024).
1430 87. K. M. Sendra *et al.*, An ancient metalloenzyme evolves through metal preference modulation. *Nat*
1431 *Ecol Evol* **7**, 732-744 (2023).
1432 88. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
1433 89. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in
1434 performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
1435 90. S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment
1436 trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
1437 91. P. J. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and
1438 bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
1439 92. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic
1440 algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2015).
1441 93. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the
1442 Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522 (2018).
1443 94. D. A. Benson *et al.*, GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).
1444 95. S. Halary, J. O. McInerney, P. Lopez, E. Bapteste, EGN: a wizard for construction of gene and
1445 genome similarity networks. *BMC Evol Biol* **13**, 146 (2013).
1446 96. F. Menardo *et al.*, Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of
1447 diversity. *BMC Bioinformatics* **19**, 164 (2018).
1448 97. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large
1449 alignments. *PLoS One* **5**, e9490 (2010).
1450 98. J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a
1451 library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**, 903-
1452 919 (2001).

99. A. P. Pandurangan, J. Stahlhacke, M. E. Oates, B. Smithers, J. Gough, The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res* **47**, D490-D494 (2019).
100. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419 (2021).
101. I. Sillitoe *et al.*, CATH: increased structural coverage of functional space. *Nucleic Acids Res* **49**, D266-D273 (2021).
102. W. Li *et al.*, RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* **49**, D1020-D1028 (2021).
103. D. H. Haft *et al.*, TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41**, D387-395 (2013).
104. N. Hulo *et al.*, The PROSITE database. *Nucleic Acids Res* **34**, D227-230 (2006).
105. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
106. P. Jones *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
107. M. Mirdita *et al.*, ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022).
108. G. Csardi, T. Nepusz, The igraph software. *Complex syst* **1695**, 1-9 (2006).
109. H. Fang, J. Gough, The 'dnet' approach promotes emerging research on cancer patient survival. *Genome Med* **6**, 64 (2014).
110. R. C. Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing. *(No Title)*, (2013).
111. M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).
112. A. C. Darling, B. Mau, F. R. Blattner, N. T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).
113. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
114. T. Hackl, M. J. Ankenbrand, gggenomes: a grammar of graphics for comparative genomics. *R package version 0.9* **5**, (2022).
115. S. Bletz, S. Janezic, D. Harmsen, M. Rupnik, A. Mellmann, Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of Clostridium difficile. *J Clin Microbiol* **56**, (2018).
116. E. Shen *et al.*, Subtyping analysis reveals new variants and accelerated evolution of Clostridioides difficile toxin B. *Commun Biol* **3**, 347 (2020).
117. P. Roy Chowdhury *et al.*, Comparative genomic analysis of toxin-negative strains of Clostridium difficile from humans and animals with symptoms of gastrointestinal disease. *BMC Microbiol* **16**, 41 (2016).
118. M. J. Mansfield *et al.*, Phylogenomics of 8,839 Clostridioides difficile genomes reveals recombination-driven evolution and diversification of toxin A and B. *PLoS Pathog* **16**, e1009181 (2020).
119. D. R. Knight, B. Elliott, B. J. Chang, T. T. Perkins, T. V. Riley, Diversity and Evolution in the Genome of Clostridium difficile. *Clin Microbiol Rev* **28**, 721-741 (2015).
120. K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-504 (2005).
121. E. W. Sayers *et al.*, Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-D26 (2022).
122. K. E. Dingle *et al.*, Effects of control interventions on Clostridium difficile infection in England: an observational study. *Lancet Infect Dis* **17**, 411-421 (2017).
123. M. Frentrup *et al.*, A publicly accessible database for Clostridioides difficile genome sequences supports tracing of transmission chains and epidemics. *Microb Genom* **6**, (2020).
124. Z. Zhou *et al.*, The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res* **30**, 138-152 (2020).

125. H. Harvala, E. Alm, T. Akerlund, K. Rizzardi, Emergence and spread of moxifloxacin-resistant Clostridium difficile ribotype 231 in Sweden between 2006 and 2015. *New Microbes New Infect* **14**, 58-66 (2016).
126. D. Hyatt *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
127. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
128. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
129. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60 (2015).
130. M. S. Bansal, M. Kellis, M. Kordi, S. Kundu, RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* **34**, 3214-3216 (2018).
131. D. R. Knight *et al.*, Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10**, (2021).
132. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017).
133. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612 (2006).
134. H. McWilliam *et al.*, Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013).
135. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019).
136. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).
137. Z. Yang, R. Nielsen, Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* **25**, 568-579 (2008).
138. S. Alvarez-Carretero, P. Kapli, Z. Yang, Beginner's Guide on the Use of PAML to Detect Positive Selection. *Mol Biol Evol* **40**, (2023).
139. Z. Yang, W. S. Wong, R. Nielsen, Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107-1118 (2005).
140. S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, S. D. Frost, Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**, 1891-1901 (2006).
141. B. Murrell *et al.*, FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol* **30**, 1196-1205 (2013).
142. B. Murrell *et al.*, Gene-wide identification of episodic selection. *Mol Biol Evol* **32**, 1365-1371 (2015).
143. S. R. Wisotsky, S. L. Kosakovsky Pond, S. D. Shank, S. V. Muse, Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Mol Biol Evol* **37**, 2430-2439 (2020).
144. B. Murrell *et al.*, Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764 (2012).
145. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
146. D. P. Martin *et al.*, RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* **7**, veaa087 (2021).
147. D. Martin, E. Rybicki, RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563 (2000).
148. M. Padidam, S. Sawyer, C. M. Fauquet, Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225 (1999).
149. D. P. Martin, D. Posada, K. A. Crandall, C. Williamson, A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98-102 (2005).
150. J. M. Smith, Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129 (1992).
151. D. Posada, K. A. Crandall, Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**, 13757-13762 (2001).

152. M. J. Gibbs, J. S. Armstrong, A. J. Gibbs, Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582 (2000).
153. H. M. Lam, O. Ratmann, M. F. Boni, Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Mol Biol Evol* **35**, 247-251 (2018).
154. A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, G. J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
155. M. V. Han, C. M. Zmasek, phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**, 356 (2009).
156. R. A. M. Villanueva, Z. J. Chen, ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives* **17**, 160-167 (2019).
157. H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, dplyr: A grammar of data manipulation. *R package version* **1.1.4**, (2023).
158. L. Zhou *et al.*, ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief Bioinform* **23**, (2022).
159. G. Yu *et al.*, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28-36 (2016).
160. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).
161. L. G. Wang *et al.*, Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol Biol Evol* **37**, 599-603 (2020).
162. S. Xu *et al.*, ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Mol Biol Evol* **38**, 4039-4042 (2021).
163. E. Campitelli, ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. *R package version 0.4* **5**, (2020).
164. K. Yamakawa, S. Kamiya, X. Q. Meng, T. Karasawa, S. Nakamura, Toxin production by Clostridium difficile in a defined medium with limited amino acids. *J Med Microbiol* **41**, 319-323 (1994).
165. U. Muh, A. G. Pannullo, D. S. Weiss, C. D. Ellermeier, A Xylose-Inducible Expression System and a CRISPR Interference Plasmid for Targeted Knockdown of Gene Expression in Clostridioides difficile. *J Bacteriol* **201**, (2019).
166. B. J. Mendoza, C. T. Trinh, Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms. *Bioinformatics* **34**, 16-23 (2018).
167. B. A. Rabe, C. Cepko, A Simple Enhancement for Gibson Isothermal Assembly. *BioRxiv*, (2020).
168. J. A. Kirk, R. P. Fagan, Heat shock increases conjugation efficiency in Clostridium difficile. *Anaerobe* **42**, 1-5 (2016).
169. T. D. Schmittgen, K. J. Livak, Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* **3**, 1101-1108 (2008).
170. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).
171. J. R. Kremer, D. N. Mastronarde, J. R. McIntosh, Computer visualization of three-dimensional image data using IMOD. *J Struct Biol* **116**, 71-76 (1996).
172. R. Fagan, N. Fairweather, Dissecting the cell surface. *Methods Mol Biol* **646**, 117-134 (2010).
173. A. Kotiranta, M. Haapasalo, K. Lounatmaa, K. Kari, Crystalline surface protein of Peptostreptococcus anaerobius. *Microbiology (Reading)* **141 (Pt 5)**, 1065-1073 (1995).
174. E. Richards *et al.*, The S-layer protein of a Clostridium difficile SLCT-11 strain displays a complex glycan required for normal cell growth and morphology. *J Biol Chem* **293**, 18123-18137 (2018).
175. G. Winter, xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186-190 (2009).
176. G. Winter *et al.*, DIALS as a toolkit. *Protein Sci* **31**, 232-250 (2022).
177. P. R. Evans, G. N. Murshudov, How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214 (2013).
178. J. Agirre *et al.*, The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **79**, 449-461 (2023).
179. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).

180. P. S. Bond, K. D. Cowtan, ModelCraft: an advanced automated model-building pipeline using Buccaneer. *Acta Crystallogr D Struct Biol* **78**, 1090-1098 (2022).
181. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004).
182. G. N. Murshudov *et al.*, REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367 (2011).
183. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019).
184. R. P. Joosten, F. Long, G. N. Murshudov, A. Perrakis, The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220 (2014).
185. C. J. Williams *et al.*, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315 (2018).
186. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243-246 (2024).
187. L. Holm, DALI and the persistence of protein shape. *Protein Sci* **29**, 128-140 (2020).
188. R. A. Laskowski, J. Jablonska, L. Pravda, R. S. Varekova, J. M. Thornton, PDBsum: Structural summaries of PDB entries. *Protein Sci* **27**, 129-134 (2018).
189. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797 (2007).
190. E. C. Meng *et al.*, UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023).

Acknowledgements

This work was supported by a Medical Research Council grant [MR/V032151/1] awarded to P.S.S and A.B.-S. and a Research Excellence Development Award [NU-018171] from the Faculty of Medical Sciences, Newcastle University awarded to A.B.-S. We thank Dr Joseph Gray for expert mass spectrometry analysis and protein identification, and Prof Craig Ellermeir for providing plasmids used in CRISPRi. Our thanks to Tracey Davey and Ross Laws from Newcastle University Electron Microscopy Research Services, supported by BB/R013942/1 for TEM sample processing; York Structural Biology Laboratory, University of York, supported by the Wellcome Trust (206161/Z/17/Z) for supporting cryo-EM data collection; Newcastle University Structural Biology Laboratory for facilitating crystallisation and data collection. We thank Diamond Light Source for access to beamline I24 (mx24948-136), through the "Macromolecular Crystallography at Newcastle, Durham, Lincoln and Durham" BAG). The contents of this work are solely the responsibilities of the authors and do not reflect the official views of any of the funders, who had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author contributions

K.M.S., A.B.-S. and P.S.S. are the corresponding authors. K.M.S. and A.B.-S conceptualised and planned the study. K.M.S. wrote the manuscript with contributions from A.B.-S. and P.S.S. K.M.S. performed all bioinformatic analyses. A.B.-S and K.M.S. performed experimental work. A.B.-S crystallized *Cd*SlpA. A.B.-S. and P.S.S. determined and analysed protein structures. P.S.S. and A.B.-S. secured funding. K.M.S. and A.B.-S. managed the project.

Competing financial interests

The authors declare no competing interests.

Data availability

The authors declare that the main data supporting the findings of this study are available within the article and Supplementary information. Data sets generated and analysed in this study are provided as Supplementary Data Files. The atomic coordinate files for X-ray structures have been deposited in the Protein Data Bank (PDB) with the accession codes 9F8E, 9F8F.