Evolutionary plasticity of bacterial surface layer protein 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 the genomic S-layer cassette, is shared across *C. difficile* isolates *via* homologous recombination[37], and its highly

variable sequence enables its use for typing (S-layer cassette type, SLCT) of different clinical isolates[37,38]. A homologue of *Cd*SlpA, *Peptostreptococcus anaerobius* surface protein PCWBR2 was recently implicated in promoting colorectal carcinogenesis[39]. Despite their medical significance, our understanding of the structural and functional landscape of S-layers formed by these diverse orthologues from closely related species is limited and current experimental structural information is available for only two *C. difficile* sequence types[40].

Here, to address these major gaps in our understanding of S-layer biology and to circumvent the inherent challenges in studying these cellular structures, we focus on narrower phylogenetic scales, where S-layer evolution is more tractable. We experimentally characterise novel S-layer types and present evidence of independent S-layer losses and gains across a bacterial family *Peptostreptococcaceae*. We reveal the role of modular evolution in shaping S-layer diversity and enabling anchoring to variable cell wall architectures. We present novel structures of *C. difficile* S-layers and uncover how, at the narrowest phylogenetic scales, extreme sequence divergence, recombination, and adaptive evolution drive extensive structural changes of the crystalline lattice surrounding the bacterial cells, without compromising its role in protection against the human immune system. Physiological tests and CRISPRi experiments demonstrate that the resistance to the innate immunity effectors emerged specifically in pathogenic lineages. Our results provide insights into the mechanisms of adaptation of host-exposed cell surfaces of S-layer-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.

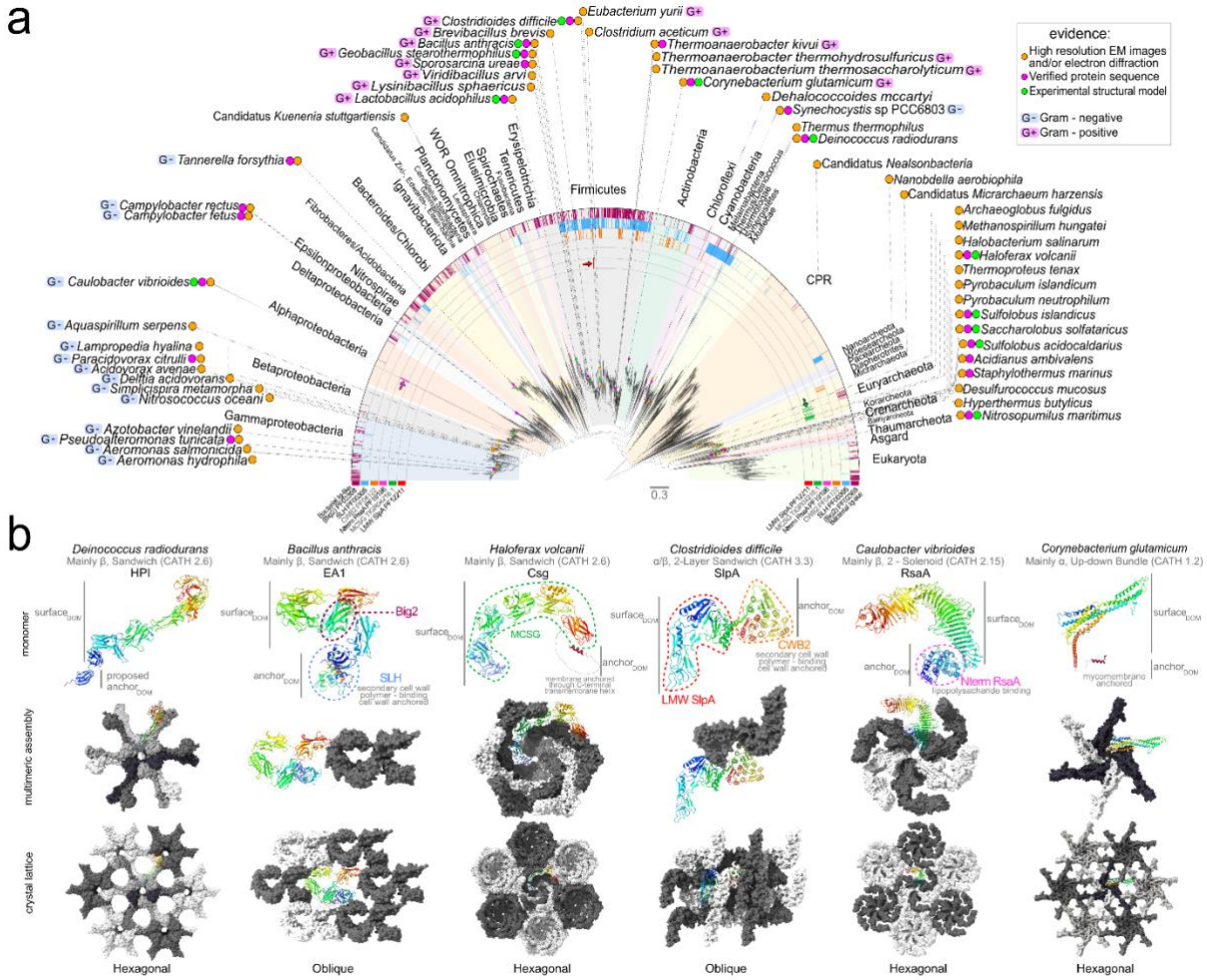

Fig. 1: Diversity of S-layer protein lattices across the Tree of Life.

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.

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

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

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

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

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

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

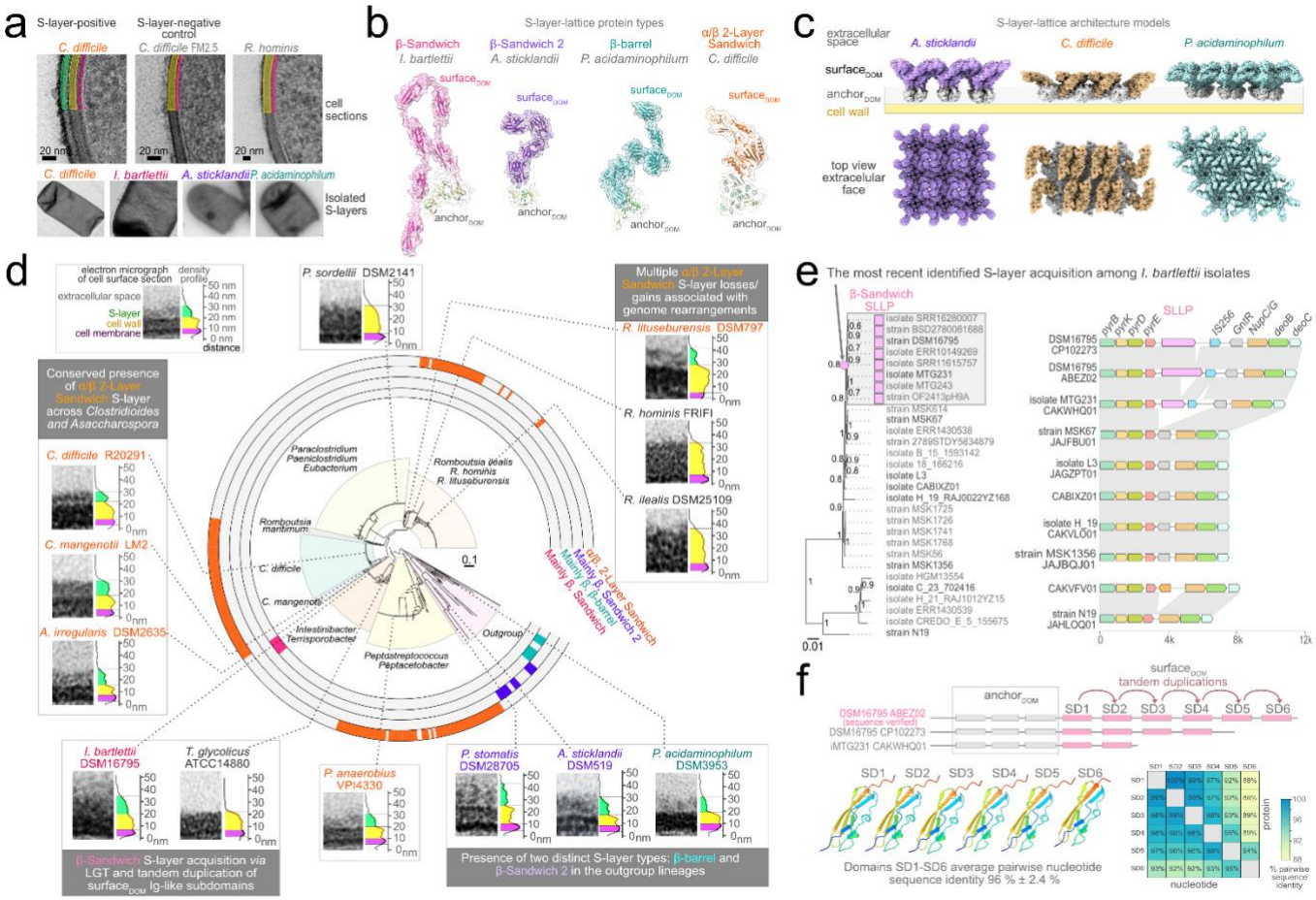

Fig. 2: Identification of new types of S-layer and mechanism of emergence of novel S-layer protein.

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.

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

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

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

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

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

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

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

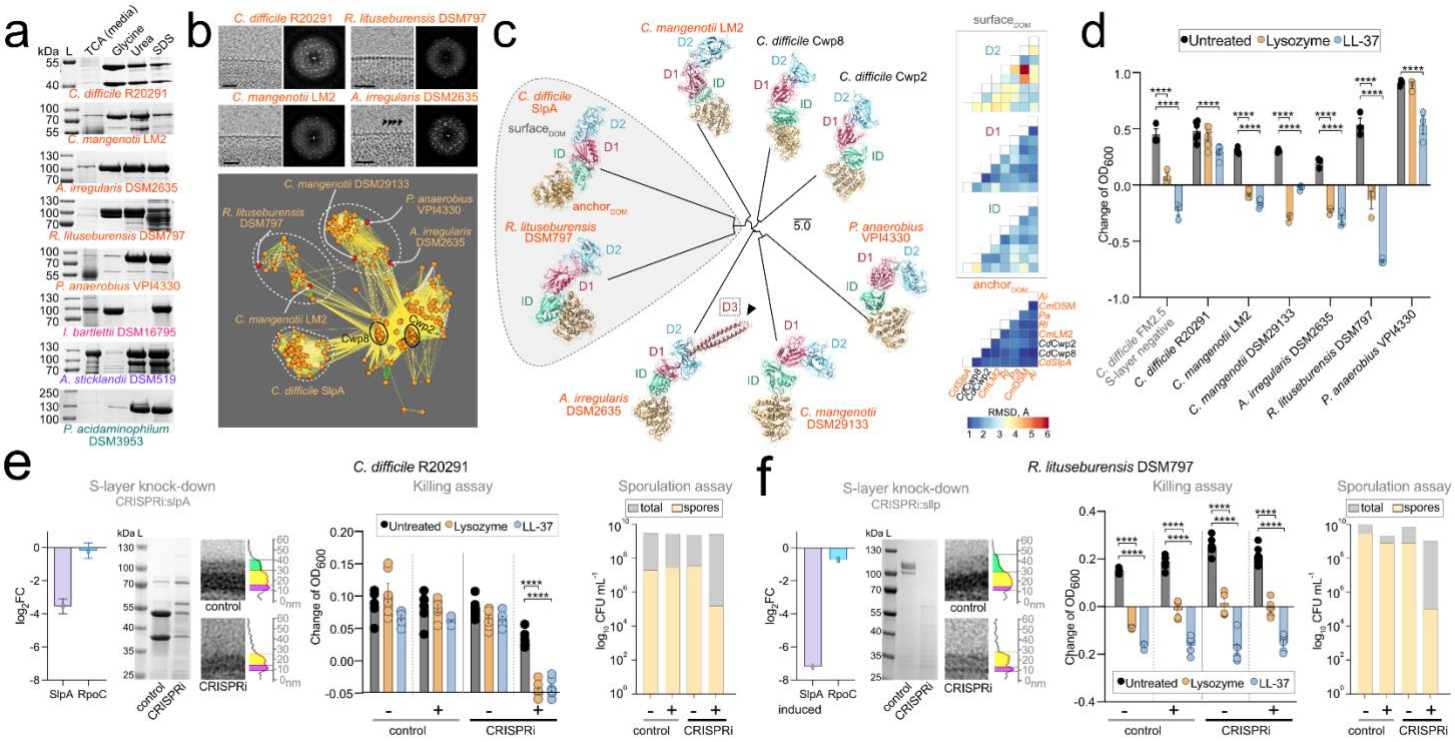

Fig. 3: Variable biochemical and physiological properties within a single structural type of S-layer.

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.

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

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

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

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

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

experimentally, *Terrisporobacter glycolicus* and *Intestinibacter bartlettii,* only *I. bartlettii* exhibited evidence of S-layer presence (Fig. 2d, and Supplementary Fig. 4) and its SLLP contained the SH3 anchoring domain (Fig. 4a). Although the presence of SH3 domains is conserved across all *Terrisporobacter/Intestinibacter*, the SH3-anchored β-sandwich SLLP homologues were specific to a distinct group of *I. bartlettii* isolates (Fig. 2e) suggesting that this novel S-layer has evolved subsequently to the major changes to the cell envelope architecture in the *Terrisporobacter/Intestinibacter* ancestor.

For the anchoring domains to perform their function in the attachment of S-layers and other surface proteins to the cell, specific secondary cell wall polymers (SCWP) need to be synthesised and incorporated into the cell wall[51]. In protein similarity network analyses of all proteins encoded in the 117 *Romboutsia*- and *Paeniclostridium*-genomes which have undergone multiple anchoring system swaps/gains/losses, we identified only a single protein without anchor enriched in CWB2-positive or SLH-positive species (Fig. 4d). The CWB2 system-associated protein was a homologue of a polysaccharide synthase (CD2777) involved in the synthesis of SCWP for CWB2 anchoring in *C. difficile*[47]*.* The SLH system-associated protein was a homologue of TagA, involved in SLH-specific SCWP biosynthesis in Firmicutes, including *B. anthracis*[52] and *Paenibacillus alvei*[51]. Therefore, enzymes required for the biosynthesis of SCWP anchors co-distribute with their cognate anchoring domains. The presence of these secondary cell wall polymer biosynthetic enzymes was also detected in organisms where both SLH and CWB2 proteins were encoded in the same genome (Fig. 4d). However, in those cases, the inferred SCWP type reflected that of the more abundant anchoring domain, while the second domain was found in only a few proteins (Fig. 4d), suggesting residual presence from incomplete anchoring domain swap, an indication that this can be a stepwise process.

For S-layers to assemble on the cell surface, in addition to the presence of specific cell surface anchors, a protein secretion system is needed to transport SLLP monomers from the cytosol across the cell membrane. An SLLP-specific SecA2 secretion system is utilised in the assembly of α/β-sandwich *C. difficile* S-layer[53]. However, our analyses revealed that neither of the *Peptostreptococcaceae* genomes encoding SH3- or CuOX-anchored β-sandwich SLLPs encoded any identifiable orthologues of SecA2. This suggests that, in *Peptostreptococcaceae*, SecA2 is associated predominantly with α/β-sandwich CWB2-anchored S-layers and some homologues of β-barrel SLH-anchored SLLPs*,* and that β-sandwich SH3- and CuOX-anchored SLLPs can utilise a more broadly distributed canonical SecA or a not yet identified alternative secretion system. Therefore, S-layers within a single bacterial family are not only assembled from different building blocks but also utilise different mechanisms for their assembly and the evolution of S-layer needs to be coordinated with associated cell wall modification and protein secretion machineries.

Together, our analyses revealed that S-layer anchor$_{DOM}$s are part of broader cell wall anchoring systems shaped by modular evolution. Entire systems and anchor$_{DOM}$ of individual proteins were swapped, acquired, or lost on multiple independent occasions, resulting in a mosaic of different anchors and functional domains often found in the genomes of closely related species. Homologues of enzymes known to be involved in the biosynthesis of CWB2- and SLH-specific secondary cell wall polymer anchors are commonly swapped alongside the SLH/CWB2 anchoring systems. This indicates that, at least in *Peptostreptococcaceae*, evolution of S-layer anchoring mechanisms is coordinated with that of cell wall chemistry. The capacity for utilisation of pre-existing swappable anchor$_{DOM}$ modules can also provide an explanation for the diversity of anchors observed in S-layers sampled from across the Tree of Life, including those utilising the same surface$_{DOM}$-fold (Fig. 1b).

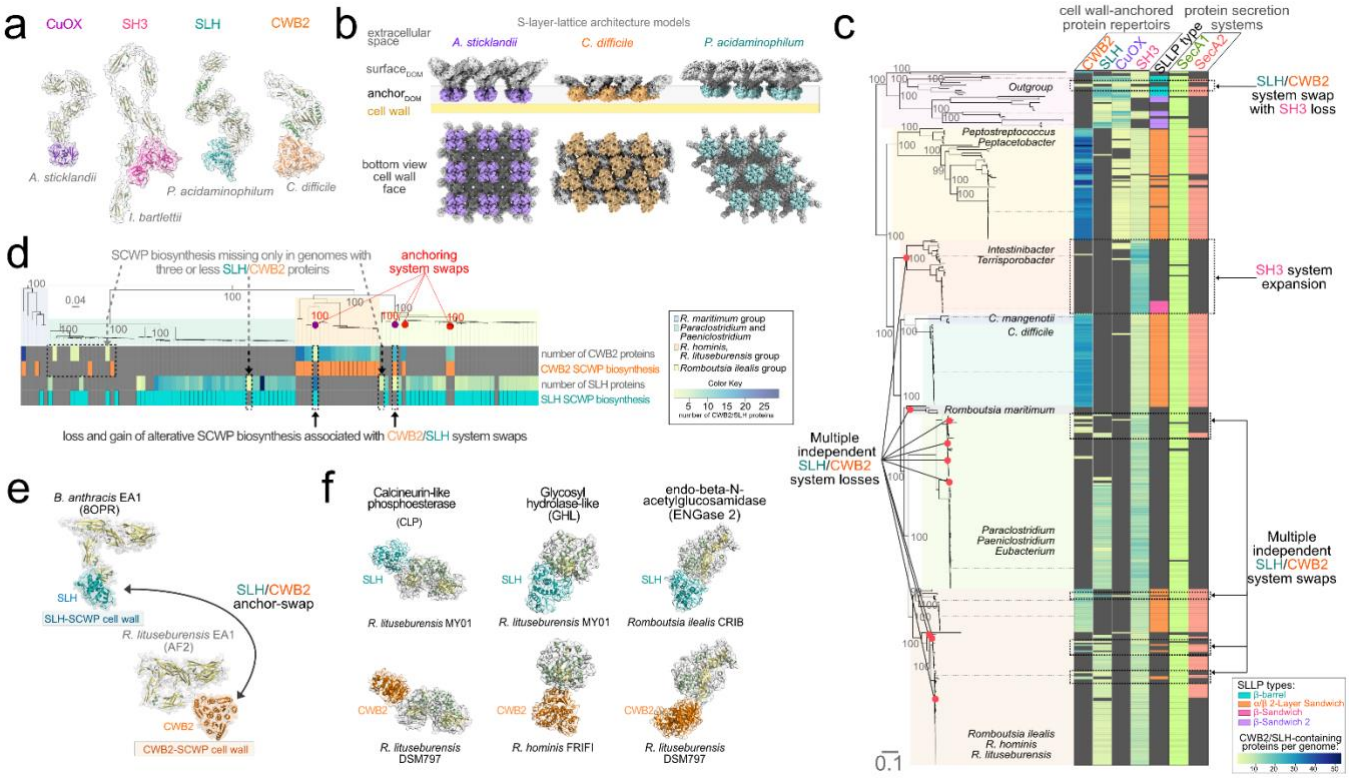

Fig. 4: Modular architecture and evolution of the S-layer and other cell envelope systems.

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.

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

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

swap in SLLP homologues. We were able to experimentally verify the presence of an EA1 homologue in *C. mangenotii* LM2 surface protein preparations, where it constituted a minor S-layer component, alongside its dominant CWB2-α/β-sandwich SLLP (Supplementary Fig. 7h,i).

Together, our data demonstrate that the observed diversity of S-layers utilising the same $anchor_{DOM}$- and $surface_{DOM}$-types can be generated through extreme sequence divergence, homologous recombination, and LGT between close and more distant relatives. Our analyses also revealed patterns of accelerated evolution in S-layer-associated proteins and proteins involved in S-layer assembly.

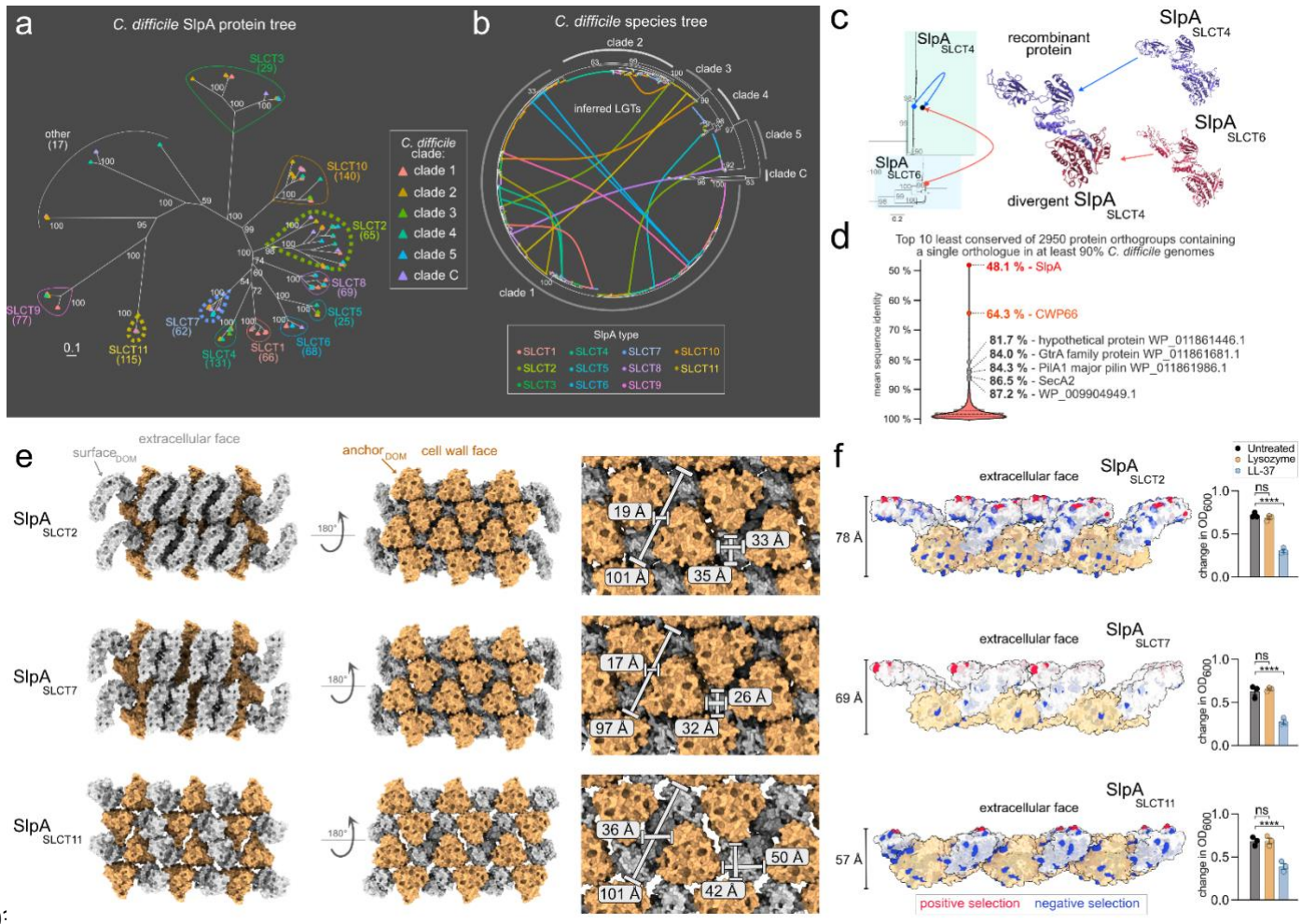

Fig. 5: Evolutionary mechanisms shaping the diversity of S-layers across clinical isolates of a pathogen.

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.

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

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

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

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

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

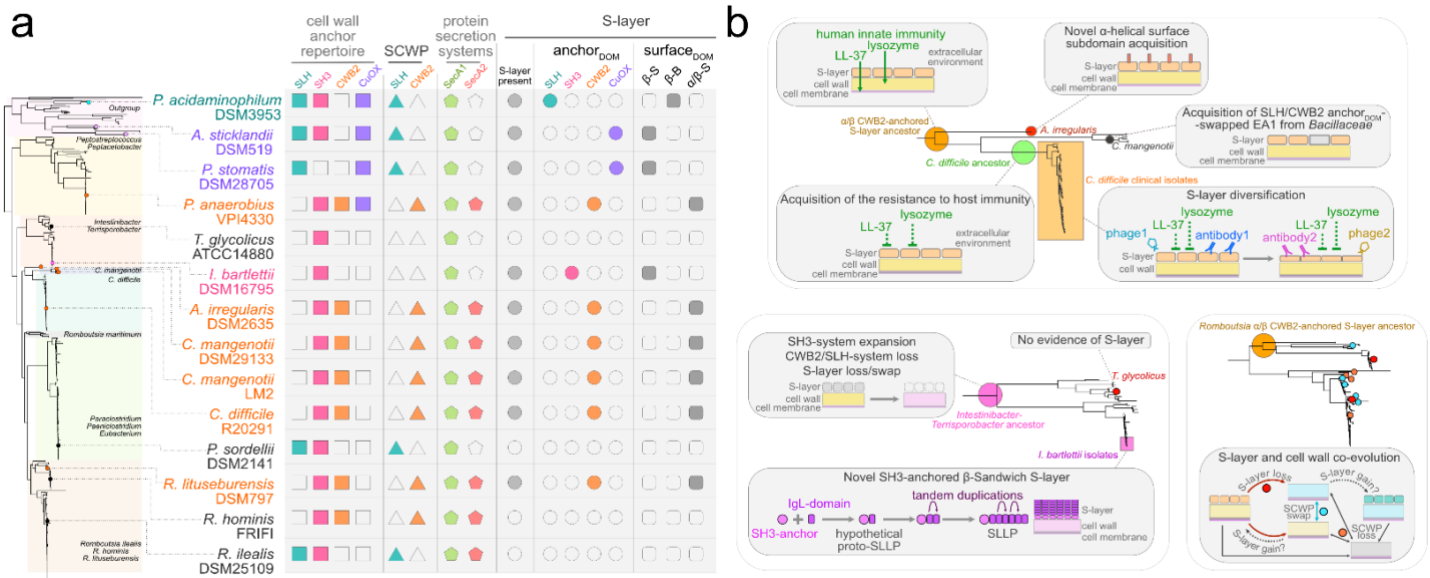

Fig. 6: Multiple evolutionary pathways to S-layer exoskeleton gain, loss and diversification in a single family of Bacteria.

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).

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

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

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

782

783 **Materials and Methods**

784

785 **Inference of the Tree of Life**

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

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

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

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

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

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

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

891 Initial genome alignments of *Romboutsia* and *Intestinibacter* genome assemblies downloaded from
892 NCBI were performed with Mauve[112] to identify regions syntenic to those encoding SLLPs that we
893 experimentally verified in S-layer-positive species. Final pairwise alignments for the identified genomic
894 regions were generated using minimap2[113] with parameters (-X -N 50 -p 0.1 -c) and visualised in R using
895 gggenomes[114] package.
896
897 **Identification of non-$anchor_{DOM}$ proteins co-distributing with anchoring systems in *Romboutsia***

898 All protein sequences encoded in the available *Romboutsia* and *Paeniclostridium* genome
899 assemblies (118), excluding those containing CWB2 (PF04122.15) and SLH (PF00395.23) anchoring domains
900 identified using hmmsearch (-E $1e^{-5}$), were analysed using protein similarity network analysis. Analysed

901 genomes were classified into two categories: CWB2-containing (23 assemblies encoding more than two
902 CWB2 proteins) and SLH-containing (65 assemblies encoding more than two SLH proteins). Non- anchor$_{DOM}$
903 proteins co-distributing with their respective anchor$_{DOM}$ systems (Supplementary Data 10) were identified
904 as protein networks containing sequences from more than 90% of the respective anchor$_{DOM}$ -containing
905 species (>21 CWB2 or >58 SLH) and no sequences from species containing the other anchor$_{DOM}$ system.
906
907 ***Clostridioides difficile* genome assemblies used in bioinformatic analyses**

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

941 To identify a set of one-to-one universal orthologues for *C. difficile* core genome species tree
942 inference, an initial dataset of 866 genome assemblies sampled across all six clades and containing all SLCT
943 representatives was analysed using sequence similarity network analyses, resulting in 4,089 networks.
944 Network decomposition following walktrap analyses resulted in a dataset of networks including 605
945 networks of one-to-one orthologues with average sequence identities ranging between 85.4 % and 100 %.
946 Of these, the final set of 147 one-to-one orthologues (Supplementary Data 3) with sequence identity level

947 below 98% was used for core genome species tree inference. The identified sequences were aligned with
948 MAFFT[89], trimmed with trimAL (gappyout mode)[90] and resulting trimmed alignments were concatenated
949 with module Nexus of the Biopython package[91] to a final alignment of 866 concatenated sequences of
950 37,580 amino acid sites, which was used to generate the *C. difficile* core genome species tree
951 (Supplementary Data 11) with IQ-TREE[92] under LG + G4 model with 1,000 ultrafast bootstraps[93].
952
953 **Identification *of Clostridioides difficile* orthogroups**

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

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

987 Initially, amino acid sequences encoded by the 972 non-redundant nucleotide *slpA* genes identified
988 across all 18,035 NCBI genomes, were aligned with MAFFT[89] and used to construct phylogeny with IQ-
989 TREE[92] under best fitting (according to AIC and BIC)[132] WAG+F+I+G4 model with 1,000 ultrafast
990 bootstraps[93]. Due to the high level of sequence divergence, the accuracy of SlpA alignments within highly
991 variable regions (mainly within LMW) was low. To account for that, we separately analysed SlpAs belonging

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

1030 Evidence of recombination across the SlpAs sampled from all available *C. difficille* genomes was
1031 detected using RDP4.101 and RDP5 Beta 5.53[146], with no differences in the results obtained with each
1032 version. The 972 non-redundant nucleotide *slpA* genes identified across all 18,035 NCBI genomes and their
1033 corresponding MAFFT[89] amino acid alignment were used to generate nucleotide alignment with a local
1034 version of PAL2NAL_v14[133]. The resulting nucleotide alignment was used as an input in a full exploratory
1035 recombination scan using RDP[147], GENECONV[148], Bootscan[149], Maxchi[150], Chimaera[151], SiSscan[152], and
1036 3Seq[153] methods with default parameters. Of all 40 candidate recombination events detected using at least
1037 one of the seven methods, 9 candidate events were statistically significant in all seven methods, of which

1038 both inferred parental sequences were 'known' in four events (i.e. their sequences were present in the
1039 tested dataset). Of the four strongly supported recombination events with both parent sequences
1040 available, only a single unambiguous recombination event was identified following the manual verification
1041 within SlpA protein phylogeny and nucleotide sequence alignments of the parent and recombinant
1042 sequences. Recombinant sequence: GenBank=CP068558.1, locus_tag=JMW52_06215,
1043 protein_id=QQY54376.1; Major parent: GenBank=PRZC01000013.1, locus_tag=C4007_03100,
1044 protein_id=MDB2714594.1;
1045 Minor parent: GenBank=ABLTPH020000028.1, locus_tag=REJ96_016885, protein_id=MDS6367270.1.
1046 Significance values from RDP4 analysis: RDP (2.6 E-58), GENECONV (5.9 E-54), Bootscan (1.5 E-60), Maxchi
1047 (4.3 E-24), Chimaera (4.1 E-23), SiSscan (4.9 E-35), 3Seq (5.6 E-11).
1048

1049 **Other bioinformatics methods**

1050 Unless stated otherwise, protein phylogenies were inferred using sequences aligned with MAFFT[89]
1051 in IQ-TREE[92] with 1,000 ultrafast bootstraps[93] under best fitting model selected with ModelFinder[132]
1052 according to Bayesian information criterion implemented in IQ-TREE. Multiple sequence alignments were
1053 inspected using Jalview[154]. Phylogenetic trees were inspected in Archaeopteryx[155]. Other general
1054 bioinformatics packages used for data analyses and visualisation included: ggplot2[156], dplyr[157], ggmsa[158],
1055 ggtree[159], ape[160], treeio[161], ggtreeExtra[162], ggnewscale[163].
1056

1057 **General microbiology methods**

1058 *Peptostreptococaccceae* strains were selected based on phylogenetic analyses and obtained from
1059 reference culture collection (DSMZ, German Collection of Microorganisms and Cell Cultures GmbH) or
1060 individual laboratory sources. All bacterial strains and specific culture conditions used in this study are
1061 described in Supplementary Table 2. Bacteria were cultured in DG250 or A35 workstations (Don Whitley
1062 Scientific) in anaerobic conditions (10% H_2, 10% CO_2, 80% N_2).
1063

1064 **Bacterial growth**

1065 To investigate the presence of S-layer and expression of S-layer proteins in different culture
1066 conditions, all strains were cultured in strain-specific media following DSMZ strain collection protocols
1067 (Supplementary Table 2) and three media commonly used for the model *Peptostreptococcacea* bacterium
1068 *C. difficile*: TY (3% tryptose, 2% yeast extract), BHIS (brain heart infusion, 0.5% yeast extract, 0.1% L-
1069 cysteine), and chemically defined minimal medium (CDMM[164]). Bacterial growth was tested at 34 °C and 37
1070 °C for all media and strains, and additionally at 28 °C for *C. mangenotii* LM2. All cultures were started by
1071 inoculating fresh media with overnight pre-cultures, in Hungate-type tubes to an OD_{600nm} of 0.08 measured
1072 in WPA CO8000 Cell Density Meter. To test the presence of S-layer and SLLPs at different growth phases,
1073 samples were taken at early-, middle-, and late-exponential, and early-, and late-stationary phases time
1074 points and OD_{600nm} values indicated in the results. Supplementation of growth media with $CaCl_2$ to a final
1075 concentration of 500 μM was carried out to assess the impact of calcium on S-layer proteins in culture.
1076

1077 **CRISPR interference for targeted knockdown of *sllp* gene expression**

1078 *E. coli-C. difficile* shuttle vector for CRISPR interference, pIA33[165], kindly provided by Prof. Craig
1079 Ellermeier, was used to target expression of genes encoding identified SLLP in *R. lituseburensis* and *C.*
1080 *difficile*. The sgRNA was analysed using CASPER[166] (https://github.com/TrinhLab/CASPERapp), generated by
1081 amplification using oligonucleotide pairs: pIA33_ SLLP_2_Rlitu_F and pIA33_Insert_R; pIA33_SlpA1_F and
1082 pIA33_Insert_R listed in Supplementary Table 3 using pIA33 template, to substitute sgRNA-rfp, and

1083 introduced into MscI+PstI-digested pIA33 backbone by isothermal assembly[167]. Negative-control plasmid
1084 (pIA34[165]) containing non-targeting sgRNA sequence that does not recognize any sequence in the genome,
1085 was used to control for non-specific plasmid effect. CRISPRi::*sll*p plasmids were introduced into
1086 *Peptostreptococacceae* host by conjugation[168]. Transformants were selected on TY thiamphenicol 10 µg mL^{-1} agar, supplemented with cycloserine (250 µg mL^{-1}).

1088 Induction of dCas9-opt activity was mediated by supplementation of growth media with xylose (1%
1089 w/v) and the interference was observed for *C. difficile slpA* (as shown before[165]) and for *R. lituseburensis*
1090 *sllp*, as demonstrated by SDS-PAGE of surface layer extracts, RT-qPCR, and electron microscopy.

1091 Viability of CRISPRi::*sllp* knockdown strains was assessed by spot assay, where a 5 µl aliquots of
1092 serial 10-fold dilutions of xylose induced and non-induced overnight cultures, were spotted onto TY-agar.
1093 Following 24-hour incubation, CFU mL^{-1} were enumerated.

1094 A RT-qPCR was performed to confirm gene silencing, using oligonucleotide pairs:
1095 qCdR20291_SlpA_F and qCdR20291_SlpA_R; qCdR20291_RpoC_F and qCdR20291_RpoC_R;
1096 qCdR20291_GyrA_F and qCdR20291_GyrA_R; qRlitu_SLLP_F and qRlitu_SLLP_R; qRlitu_RpoC_F and
1097 qRlitu_RpoC_R; qRlitu_GyrA_F and qRlitu_GyrA_R, designed with PrimerQuest (Integrated DNA
1098 Technologies) listed in Supplementary Table 3. RNA extraction was carried out with Monarch Total RNA
1099 Miniprep kit (NEB) from 8 h cultures. DNAse treatment was carried out with TURBO DNA-free™ Kit
1100 (Invitrogen). cDNA preparation was performed using the SuperScript III VILO cDNA synthesis kit. RT-qPCR
1101 was performed using Luna Universal qPCR master Mix (NEB) in LightCycler96 (Roche). Samples of two
1102 biological replicates were analysed in 3 technical replicates. Results were calculated by the comparative
1103 cycle threshold method[169] and normalized to the gyrA transcript. Relative expression of *sllp* and *rpoC*
1104 control was presented as a log2 fold change, relative to non-induced CRISPRi::*sllp* expression level.
1105

1106 **Broth-based killing assay**

1107 Human innate immune effector (lysozyme; 500 µg mL^{-1}, or LL-37; 5 µg mL^{-1}) was added to mid-
1108 exponential cultures (OD_{600nm} ~0.5) and its impact on the growth of strains expressing intact SLLP or of
1109 CRISPRi *sllp* knock-down was assessed by monitoring OD_{600nm}. Two biological and three technical replicates
1110 were analysed in microplate reader (Stratus, Cerillo). OD change was estimated from a difference between
1111 OD_{600nm} at endpoint (8 h timepoint for *C. difficile* and 5 h for other *Peptostreptococacceae*) and OD_{600nm}
1112 measured upon supplementation with the additive.
1113

1114 **Sporulation assay**

1115 CRISPRi::*sllp* knock-down strains of *C. difficile* and *R. lituseburensis* presented sporulation defects
1116 compared to negative control when assessed by electron microscopy of whole cells. To estimate
1117 sporulation efficiency, the knock-down strains were cultured overnight, in biological duplicate. Aliquots of
1118 50 µl were spotted onto selective TY-agar supplemented or non-supplemented with xylose and cultured for
1119 24 h. Growth was collected from the plate and resuspended in 500 µl PBS. Cell suspension was split and
1120 one aliquot of 250 µl was heat-treated at 70 °C for 20 min. Both non-treated and heat-treated aliquots
1121 were serially 10-fold diluted in PBS, and 5 µl aliquots of dilutions were spotted onto TY, 0.1% taurocholate,
1122 0.1% cysteine agar plates. Following overnight incubation, CFU mL^{-1} was enumerated. The heat-resistant
1123 CFUs are represented as a percentage of total viable (non-treated) CFUs.
1124

1125 **Transmission electron microscopy (TEM) of bacterial cell sections**

1126 Stationary phase bacterial cultures were harvested by centrifugation at 2,000 x*g* for 10 min at room
1127 temperature. Cell pellets were collected in 1 ml of fixative (2% glutaraldehyde in 0.1 M sodium cacodylate
1128 buffer, pH 7.4) and stored at 4 °C until further processing, performed at Newcastle University Electron

1129 Microscopy Research Services (EMRS). Samples were enrobed in low melting point agarose and post-fixed
1130 in 1% osmium tetroxide, dehydrated in gradient acetone, and embedded in epoxy resin.
1131 Cell surface glycopolymers were visualised using a published protocol[47]. Briefly, bacterial colonies
1132 were collected from 24 h agar plate cultures and pre-fixed with 2% glutaraldehyde and 10 mM L-lysine in
1133 CB-RR (0.1 M sodium cacodylate buffer, pH 7.4, 0.1% ruthenium red) for 10 min at room temperature.
1134 Subsequently, pellets were fixed in 2% glutaraldehyde in CB-RR for 2 h at room temperature, washed with
1135 CB-RR, and post-fixed in 1% osmium tetroxide in CB-RR. Samples were then enrobed in agarose,
1136 dehydrated in a gradient of ethanol, and embedded in the polyhydroxy aromatic acrylic resin LR white.
1137 Ultrathin sections of around 70 nm were deposited on copper grids and stained with uranyl acetate
1138 and lead citrate. Grids were then visualised using a Hitachi HT7800 120 kV transmission electron
1139 microscope (EMRS, Newcastle University).
1140 Line profile analysis of cell envelope across cell sections was performed with Fiji[170].
1141
1142 **Electron microscopy of isolated S-layer lattice fragments**

1143 S-layer preparation for electron microscopy was performed as previously described[40] with minor
1144 modifications. Mid-exponential cultures were harvested by centrifugation at 2,000 x*g* for 15 min at room
1145 temperature. Cell pellets were washed twice in ice-cold deionised water (0.25x of culture volume) and
1146 disrupted by pulse-sonication for 2 min on ice. Unbroken cells were removed by centrifugation at 800 x*g* for
1147 10 minutes at 4 °C and the supernatant was further centrifuged at 3,000 x*g* for 10 minutes. Harvested
1148 pellets containing S-layer fragments were washed with ice-cold 1 M NaCl (to an effective OD_{600nm} of 100, of
1149 initial culture) and resuspended in cold 2% Triton X-100 (to an effective OD_{600nm} of 200, of initial culture).
1150 Preparations were loaded on glow-discharged, carbon-coated 300 mesh copper EM grids (Agar Scientific),
1151 stained with 2% uranyl acetate, and assessed using a Hitachi HT7800 transmission electron microscope
1152 operating at 120 kV (EMRS, Newcastle University).
1153 For electron cryo-microscopy, S-layer fragments were diluted 10-fold in cold deionised water and
1154 added onto plasma-treated Quantifoil Cu R2/2, 200 mesh grids. The grids were blotted for 5 seconds at 4
1155 °C, 90% relative humidity, vitrified in liquid ethane, using Vitrobot (FEI), and stored in liquid nitrogen.
1156 Cryo-EM grid preparation and imaging were performed at the York Structural Biology Laboratory
1157 (YSBL) on a Glacios electron microscope equipped with a Falcon 4 direct electron detector (Thermo Fisher
1158 Scientific). Micrographs were taken at 200 kW with a magnification of 120,000 and -2µm defocus.
1159 Images were viewed and Fourier transforms were generated using 3dMOD[171] and Fiji[170].

1160
1161 **Identification of S-layer lattice proteins**

1162 As the main building blocks of known bacterial S-layers, SLLPs typically constitute one of the most
1163 abundant cellular polypeptides and the dominant protein in S-layers, aiding their identification from
1164 dominant bands in gel electrophoresis of S-layer preparations.
1165 To identify the *Peptostreptococcaceae* SLLPs, bacteria with evidence of S-layer presence in electron
1166 micrographs of S-layer fragments and whole cell sections were cultured to stationary phase and harvested
1167 by centrifugation at 4,000 x*g* for 10 min at room temperature. Additional experiments with sampling from
1168 across the growth cycle were performed for *C. mangenotii* LM2, which contains both CWB2-SLLP and
1169 CWB2-EA1, and for the strains where no evidence of S-layer was found in the electron micrographs.
1170 Harvested wet cell pellets were washed with a HEPES-NaCl buffer (50 mM HEPES, 100 mM NaCl, pH 7.5),
1171 followed by a critical step of the careful removal of all HEPES-NaCl liquid to ensure consistent extraction
1172 conditions in the subsequent steps. Non-covalently bound proteins, including S-layer proteins, were
1173 extracted from the cell surface using three different chemical methods. Preparations using 0.2 M glycine pH

1174 2.2 were carried out as previously described[40]. Two additional methods of extraction: 6 M urea, 50 mM Tris
1175 pH 8.0, 1 mM phenyl-methyl-sulphonide-fluoride (PMSF); and 1% sodium dodecyl sulphate (SDS), 50 mM
1176 Tris–HCl pH 8.0, 1 mM PMSF were used following the published protocols[172,173]. For each method, the
1177 washed cell pellets were resuspended in extraction solution to an effective OD_{600nm} of 50 and incubated at
1178 room temperature for 30 min, followed by centrifugation at 21,000 x*g* for 10 min. A 2 M Tris base was used
1179 to neutralise the pH of low-pH glycine preparations. Extracts were then mixed with a 0.25x volume of
1180 sample buffer (8% SDS, 400 mM DTT, 40% glycerol, 200 mM Tris-HCl pH 6.8, 0.4% bromophenol blue) and
1181 separated via 6-20% gradient SDS-polyacrylamide gel electrophoresis (SDS-PAGE).
1182 Recovered spent media were used to assess secreted protein content and cell autolysis, by
1183 precipitation with 10% (v/v) of trichloroacetic acid (TCA), for 30 min on ice. Following centrifugation at
1184 21,000 x*g* for 10 min at 4 °C, the precipitate was washed 3 times with ice-cold 90% acetone. The resulting
1185 air-dried pellet was resuspended in PBS to an effective OD_{600nm} of 50, mixed with a 0.25x volume of sample
1186 buffer (8% SDS, 400 mM dithiothreitol (DTT), 40% glycerol, 200 mM Tris-HCl pH 6.8, 0.4% bromophenol
1187 blue) and separated via 6-20% gradient SDS-PAGE.
1188 To identify proteins in S-layer preparations, trypsin digests of the respective bands excised from
1189 SDS-polyacrylamide gels were analysed by LC-MS/MS in positive ion mode on a Finnigan LTQ ion trap mass
1190 spectrometer (Thermo Fisher Scientific, Bremen, Germany). Protein identification was performed using the
1191 Mascot MS/MS search engine program (Matrix Science Ltd, London) against a custom protein sequence
1192 database (*Peptostreptococcaceae*, containing 179,609 unique sequences created on 26/01/2023 from
1193 UniProt).
1194
1195 **X-ray crystallography of SlpA**

1196 TY broth cultures (16 h) of *C. difficile* strains were centrifuged at room temperature at 5,000 x*g* and
1197 the resulting pellets were washed with 0.1x culture volume of 50 mM M HEPES pH 7.4, 150 mM NaCl
1198 buffer. S-layer extraction was performed by resuspending the washed pellet in 0.01x culture volume of 0.2
1199 M glycine-HCl pH 2.2, as described above. S-layer extract was filtered and resolved onto a Superdex 200
1200 26/600 column using an ÄKTA Pure FPLC system (Cytiva) in 50 mM Tris-HCl pH 7.5, 150 mM NaCl buffer.
1201 Purified proteins were concentrated to 10 mg mL^{-1} and subjected to crystallization with the sitting drop
1202 vapour-diffusion method at 20 °C with a range of commercial screens.
1203 Although $SlpA_{SLCT11}$ was purified from Ox247 (wild type strain[37]) and Ox247Δorf2 (mutant defective
1204 in glycosylation of SlpA[174], denoted SLCT11*), only the non-glycosylated protein resulted in successful
1205 crystallisation in 4 mM oxometalates, 50 mM Tris (base); 50 mM bicine 8.5, 20% v/v glycerol; 10% w/v PEG
1206 4000 (Morpheus Fusion, Molecular Dimensions). $SlpA_{SLCT2}$ purified from Opt2472 strain[122] produced
1207 diffraction quality crystals in 1.6 M Sodium citrate tribasic (Wizard Classic 3, Molecular Dimensions).
1208 Diffraction data were collected on I24 beamline (λ = 0.71 Å) at the Diamond Light Source
1209 Synchrotron (Didcot, UK; mx24948-136) at 100 K. The data was processed with xia2[175] ($SlpA_{SLCT11*}$) or
1210 DIALS[176] ($SlpA_{SLCT2}$), scaled with Aimless[177] within CCP4 Cloud[178]. Initial phases were obtained by molecular
1211 replacement in Phaser[179] with models of the CWB2 anchoring domain using the $SlpA_{CD630}$ (PDB ID: 7ACY)
1212 template for $SlpA_{SLCT11*}$ or the AlphaFold2[145] predicted model for $SlpA_{SLCT2}$. The generated solution model
1213 was then subjected to automatic model building with ModelCraft[180], followed by iterative cycles of manual
1214 building with Coot[181] and refinement in Refmac5[182].
1215 Final models were obtained after iterative cycles of manual model building with Coot and
1216 refinement in Phenix_refine[183] and PDBREDO[184]. Validation of final models was performed using Coot and
1217 Phenix internal tools, as well as MolProbity[185] web server. Data collection and refinement statistics are
1218 summarized in Supplementary Table 4.
1219

1220 **Other structural analyses**

1221 Structural prediction for proteins of interest was performed with ColabFold v1.5.5: AlphaFold2
1222 using MMseqs2[107]. Comparative structural searches were carried out with Foldseek[186] and Dali[187] servers.
1223 Structural analyses including crystal contacts, distances, angles and dimensions were performed with
1224 PDBSum[188], PISA[189] and with plugins implemented in PyMOL Molecular Graphics System (Schrödinger, LLC)
1225 or UCSF ChimeraX[190]. Graphical representations of S-layer proteins and crystal latices were generated using
1226 UCSF ChimeraX. SymProFold pipeline was used to predict S-layer lattice assemblies[35] following the standard
1227 protocol (https://github.com/symprofold/SymProFold_Tutorial_Data). AlphaFold3 server, using the
1228 PDB2025 (up to 3 February 2025) template option, was used to generate structural predictions for full-
1229 length and surfaceDOMs models for experimentally identified SLLPs across *Peptotreptococcaceae.*
1230 Individual subdomains – $anchor_{DOM}$, $surface_{DOM}$, D1, D2 and ID - were manually extracted from the ful-
1231 length predictions and aligned using the LSQ function in COOT. Calculated RMSD for pairwise
1232 superimpositions in Coot were represented as a heatmap using Rstudio.
1233

1234 **References**

1235

1236 1 Messner, P. & Sleytr, U. B. Crystalline bacterial cell-surface layers. *Adv Microb Physiol* **33**, 213-275
1237 (1992). https://doi.org:10.1016/s0065-2911(08)60218-0
1238 2 Bharat, T. A. M., von Kugelgen, A. & Alva, V. Molecular Logic of Prokaryotic Surface Layer
1239 Structures. *Trends Microbiol* **29**, 405-415 (2021). https://doi.org:10.1016/j.tim.2020.09.009
1240 3 Pum, D., Toca-Herrera, J. L. & Sleytr, U. B. S-layer protein self-assembly. *Int J Mol Sci* **14**, 2484-2501
1241 (2013). https://doi.org:10.3390/ijms14022484
1242 4 Messner, P., Schäffer, C., Egelseer, E.-M. & Sleytr, U. B. in *Prokaryotic Cell Wall Compounds:*
1243 *Structure and Biochemistry* (eds Helmut König, Harald Claus, & Ajit Varma) 53-109 (Springer Berlin
1244 Heidelberg, 2010).
1245 5 Fagan, R. P. & Fairweather, N. F. Biogenesis and functions of bacterial S-layers. *Nat Rev Microbiol*
1246 **12**, 211-222 (2014). https://doi.org:10.1038/nrmicro3213
1247 6 Missiakas, D. & Schneewind, O. Assembly and Function of the Bacillus anthracis S-Layer. *Annu Rev*
1248 *Microbiol* **71**, 79-98 (2017). https://doi.org:10.1146/annurev-micro-090816-093512
1249 7 Bradshaw, W. J., Roberts, A. K., Shone, C. C. & Acharya, K. R. The structure of the S-layer of
1250 Clostridium difficile. *J Cell Commun Signal* **12**, 319-331 (2018). https://doi.org:10.1007/s12079-017-
1251 0429-z
1252 8 Trautner, C. & Vermaas, W. F. The sll1951 gene encodes the surface layer protein of Synechocystis
1253 sp. strain PCC 6803. *J Bacteriol* **195**, 5370-5380 (2013). https://doi.org:10.1128/JB.00615-13
1254 9 Lasa, I., Caston, J. R., Fernandez-Herrero, L. A., de Pedro, M. A. & Berenguer, J. Insertional
1255 mutagenesis in the extreme thermophilic eubacteria Thermus thermophilus HB8. *Mol Microbiol* **6**,
1256 1555-1564 (1992). https://doi.org:10.1111/j.1365-2958.1992.tb00877.x
1257 10 Kirk, J. A. *et al.* New class of precision antimicrobials redefines role of Clostridium difficile S-layer in
1258 virulence and viability. *Sci Transl Med* **9** (2017). https://doi.org:10.1126/scitranslmed.aah6813
1259 11 Fioravanti, A., Mathelie-Guinlet, M., Dufrene, Y. F. & Remaut, H. The Bacillus anthracis S-layer is an
1260 exoskeleton-like structure that imparts mechanical and osmotic stabilization to the cell wall. *PNAS*
1261 *Nexus* **1**, pgac121 (2022). https://doi.org:10.1093/pnasnexus/pgac121
1262 12 Ali, S. *et al.* Slr4, a newly identified S-layer protein from marine Gammaproteobacteria, is a major
1263 biofilm matrix component. *Mol Microbiol* **114**, 979-990 (2020). https://doi.org:10.1111/mmi.14588
1264 13 Wong, L. L. *et al.* Surface-layer protein is a public-good matrix exopolymer for microbial community
1265 organisation in environmental anammox biofilms. *ISME J* **17**, 803-812 (2023).
1266 https://doi.org:10.1038/s41396-023-01388-y
1267 14 Shimotahira, N. *et al.* The surface layer of Tannerella forsythia contributes to serum resistance and
1268 oral bacterial coaggregation. *Infect Immun* **81**, 1198-1206 (2013).
1269 https://doi.org:10.1128/IAI.00983-12
1270 15 Kern, J. W. & Schneewind, O. BslA, a pXO1-encoded adhesin of Bacillus anthracis. *Mol Microbiol* **68**,
1271 504-515 (2008). https://doi.org:10.1111/j.1365-2958.2008.06169.x
1272 16 Merrigan, M. M. *et al.* Surface-layer protein A (SlpA) is a major contributor to host-cell adherence
1273 of Clostridium difficile. *PLoS One* **8**, e78404 (2013). https://doi.org:10.1371/journal.pone.0078404
1274 17 Sakakibara, J. *et al.* Loss of adherence ability to human gingival epithelial cells in S-layer protein-
1275 deficient mutants of Tannerella forsythensis. *Microbiology (Reading)* **153**, 866-876 (2007).
1276 https://doi.org:10.1099/mic.0.29275-0
1277 18 Plaut, R. D. *et al.* Genetic evidence for the involvement of the S-layer protein gene sap and the
1278 sporulation genes spo0A, spo0B, and spo0F in Phage AP50c infection of Bacillus anthracis. *J*
1279 *Bacteriol* **196**, 1143-1154 (2014). https://doi.org:10.1128/JB.00739-13
1280 19 Royer, A. L. M. *et al.* Clostridioides difficile S-Layer Protein A (SlpA) Serves as a General Phage
1281 Receptor. *Microbiol Spectr* **11**, e0389422 (2023). https://doi.org:10.1128/spectrum.03894-22
1282 20 Sara, M. & Sleytr, U. B. Molecular sieving through S layers of Bacillus stearothermophilus strains. *J*
1283 *Bacteriol* **169**, 4092-4098 (1987). https://doi.org:10.1128/jb.169.9.4092-4098.1987
1284 21 von Kugelgen, A. *et al.* Membraneless channels sieve cations in ammonia-oxidizing marine archaea.
1285 *Nature* (2024). https://doi.org:10.1038/s41586-024-07462-5

1286 22 Sleytr, U. B., Egelseer, E. M., Ilk, N., Pum, D. & Schuster, B. S-Layers as a basic building block in a
1287 molecular construction kit. *FEBS J* **274**, 323-334 (2007). https://doi.org:10.1111/j.1742-
1288 4658.2006.05606.x
1289 23 An, B. *et al.* Engineered Living Materials For Sustainability. *Chem Rev* **123**, 2349-2419 (2023).
1290 https://doi.org:10.1021/acs.chemrev.2c00512
1291 24 Yuliawati, Y. *et al.* Potency of surface layer protein from Lactobacillus sp. as drug nanocarriers.
1292 *Journal of Applied Pharmaceutical Science* (2024). https://doi.org:10.7324/JAPS.2024.199203
1293 25 Zhang, B. *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). https://doi.org:10.1186/s12934-024-02409-x
1296 26 Decout, A. *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 https://doi.org:10.1038/s41467-024-55233-7
1299 27 Li, J. *et al.* Exopolysaccharides and Surface-Layer Proteins Expressed by Biofilm-State
1300 Lactiplantibacillus plantarum Y42 Play Crucial Role in Preventing Intestinal Barrier and Immunity
1301 Dysfunction of Balb/C Mice Infected by Listeria monocytogenes ATCC 19115. *J Agric Food Chem* **72**,
1302 8581-8594 (2024). https://doi.org:10.1021/acs.jafc.4c00460
1303 28 Konstantinov, S. R. *et al.* S layer protein A of Lactobacillus acidophilus NCFM regulates immature
1304 dendritic cell and T cell functions. *Proc Natl Acad Sci U S A* **105**, 19474-19479 (2008).
1305 https://doi.org:10.1073/pnas.0810305105
1306 29 Wang, H., Zhang, L., Li, Q., Xu, S. & Lu, R. Surface-layer protein produced by Lactobacillus crispatus
1307 JCM 2009 ameliorates lipopolysaccharide-induced inflammation through autophagy cross-talk with
1308 the NF-kappaB signaling pathway. *Int J Biol Macromol* **166**, 633-640 (2021).
1309 https://doi.org:10.1016/j.ijbiomac.2020.10.221
1310 30 Ormsby, M. J. *et al.* An intact S-layer is advantageous to Clostridioides difficile within the host. *PLoS*
1311 *Pathog* **19**, e1011015 (2023). https://doi.org:10.1371/journal.ppat.1011015
1312 31 Wang, Y. T., Oh, S. Y., Hendrickx, A. P., Lunderberg, J. M. & Schneewind, O. Bacillus cereus G9241 S-
1313 layer assembly contributes to the pathogenesis of anthrax-like disease in mice. *J Bacteriol* **195**, 596-
1314 605 (2013). https://doi.org:10.1128/JB.02005-12
1315 32 Fioravanti, A. *et al.* Structure of S-layer protein Sap reveals a mechanism for therapeutic
1316 intervention in anthrax. *Nat Microbiol* **4**, 1805-1814 (2019). https://doi.org:10.1038/s41564-019-
1317 0499-1
1318 33 Hunault, L. *et al.* Anti-S-layer monoclonal antibodies impact Clostridioides difficile physiology. *Gut*
1319 *Microbes* **16**, 2301147 (2024). https://doi.org:10.1080/19490976.2023.2301147
1320 34 Johnston, E., Isbilir, B., Alva, V., Bharat, T. A. M. & Doye, J. P. K. Punctuated and continuous
1321 structural diversity of S-layers across the prokaryotic tree of life. PREPRINT (Version 1) available at.
1322 *bioRxiv*, 2024.2005.2028.596244 (2024). https://doi.org:10.1101/2024.05.28.596244
1323 35 Buhlheller, C. *et al.* SymProFold: Structural prediction of symmetrical biological assemblies. *Nat*
1324 *Commun* **15**, 8152 (2024). https://doi.org:10.1038/s41467-024-52138-3
1325 36 Sogues, A. *et al.* Cryo-EM structure and polar assembly of the PS2 S-layer of Corynebacterium
1326 glutamicum. *bioRxiv* (2024). https://doi.org:10.1101/2024.09.05.611363
1327 37 Dingle, K. E. *et al.* Recombinational switching of the Clostridium difficile S-layer and a novel
1328 glycosylation gene cluster revealed by large-scale whole-genome sequencing. *J Infect Dis* **207**, 675-
1329 686 (2013). https://doi.org:10.1093/infdis/jis734
1330 38 Ni Eidhin, D., Ryan, A. W., Doyle, R. M., Walsh, J. B. & Kelleher, D. Sequence and phylogenetic
1331 analysis of the gene for surface layer protein, slpA, from 14 PCR ribotypes of Clostridium difficile. *J*
1332 *Med Microbiol* **55**, 69-83 (2006). https://doi.org:10.1099/jmm.0.46204-0
1333 39 Long, X. *et al.* Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates
1334 tumour immunity. *Nat Microbiol* **4**, 2319-2330 (2019). https://doi.org:10.1038/s41564-019-0541-3
1335 40 Lanzoni-Mangutchi, P. *et al.* Structure and assembly of the S-layer in C. difficile. *Nat Commun* **13**,
1336 970 (2022). https://doi.org:10.1038/s41467-022-28196-w
1337 41 Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the
1338 uncultured microbial majority. *Nat Rev Microbiol* **19**, 225-240 (2021).
1339 https://doi.org:10.1038/s41579-020-00458-8

1340 42 Silale, A. *et al.* Dual function of OmpM as outer membrane tether and nutrient uptake channel in
1341 diderm Firmicutes. *Nat Commun* **14**, 7152 (2023). https://doi.org:10.1038/s41467-023-42601-y
1342 43 Messner, P., Hollaus, F. & Sleytr, U. B. Paracrystalline Cell Wall Surface Layers of Different Bacillus
1343 stearothermophilus Strains. *International Journal of Systematic Bacteriology* **34**, 202-210 (1984).
1344 https://doi.org:10.1099/00207713-34-2-202
1345 44 Usenik, A. *et al.* The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the
1346 Clostridium difficile Cell Wall Proteins Cwp8 and Cwp6. *Structure* **25**, 514-521 (2017).
1347 https://doi.org:10.1016/j.str.2016.12.018
1348 45 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3.
1349 *Nature* **630**, 493-500 (2024). https://doi.org:10.1038/s41586-024-07487-w
1350 46 Willing, S. E. *et al.* Clostridium difficile surface proteins are anchored to the cell wall using CWB2
1351 motifs that recognise the anionic polymer PSII. *Mol Microbiol* **96**, 596-608 (2015).
1352 https://doi.org:10.1111/mmi.12958
1353 47 Chu, M. *et al.* A Clostridium difficile Cell Wall Glycopolymer Locus Influences Bacterial Shape,
1354 Polysaccharide Production and Virulence. *PLoS Pathog* **12**, e1005946 (2016).
1355 https://doi.org:10.1371/journal.ppat.1005946
1356 48 Schmidt, A., Muller, N., Schink, B. & Schleheck, D. A proteomic view at the biochemistry of
1357 syntrophic butyrate oxidation in Syntrophomonas wolfei. *PLoS One* **8**, e56905 (2013).
1358 https://doi.org:10.1371/journal.pone.0056905
1359 49 Pinheiro, B. A. *et al.* Functional insights into the role of novel type I cohesin and dockerin domains
1360 from Clostridium thermocellum. *Biochem J* **424**, 375-384 (2009).
1361 https://doi.org:10.1042/BJ20091152
1362 50 Artzi, L. *et al.* Cellulosomics of the cellulolytic thermophile Clostridium clariflavum. *Biotechnol*
1363 *Biofuels* **7**, 100 (2014). https://doi.org:10.1186/1754-6834-7-100
1364 51 Hager, F. F. *et al.* Functional Characterization of Enzymatic Steps Involved in Pyruvylation of
1365 Bacterial Secondary Cell Wall Polymer Fragments. *Front Microbiol* **9**, 1356 (2018).
1366 https://doi.org:10.3389/fmicb.2018.01356
1367 52 Oh, S. Y., Lunderberg, J. M., Chateau, A., Schneewind, O. & Missiakas, D. Genes Required for
1368 Bacillus anthracis Secondary Cell Wall Polysaccharide Synthesis. *J Bacteriol* **199** (2017).
1369 https://doi.org:10.1128/JB.00613-16
1370 53 Fagan, R. P. & Fairweather, N. F. Clostridium difficile has two parallel and essential Sec secretion
1371 systems. *J Biol Chem* **286**, 27483-27493 (2011). https://doi.org:10.1074/jbc.M111.263889
1372 54 Ausiello, C. M. *et al.* Surface layer proteins from Clostridium difficile induce inflammatory and
1373 regulatory cytokines in human monocytes and dendritic cells. *Microbes Infect* **8**, 2640-2646 (2006).
1374 https://doi.org:10.1016/j.micinf.2006.07.009
1375 55 Ryan, A. *et al.* A role for TLR4 in Clostridium difficile infection and the recognition of surface layer
1376 proteins. *PLoS Pathog* **7**, e1002076 (2011). https://doi.org:10.1371/journal.ppat.1002076
1377 56 Waligora, A. J. *et al.* Characterization of a cell surface protein of Clostridium difficile with adhesive
1378 properties. *Infect Immun* **69**, 2144-2153 (2001). https://doi.org:10.1128/IAI.69.4.2144-2153.2001
1379 57 Zhou, Q. *et al.* The cwp66 Gene Affects Cell Adhesion, Stress Tolerance, and Antibiotic Resistance in
1380 Clostridioides difficile. *Microbiol Spectr* **10**, e0270421 (2022).
1381 https://doi.org:10.1128/spectrum.02704-21
1382 58 Promadej, N., Fiedler, F., Cossart, P., Dramsi, S. & Kathariou, S. Cell wall teichoic acid glycosylation
1383 in Listeria monocytogenes serotype 4b requires gtcA, a novel, serogroup-specific gene. *J Bacteriol*
1384 **181**, 418-425 (1999). https://doi.org:10.1128/JB.181.2.418-425.1999
1385 59 Chateau, A. *et al.* Distinct Pathways Carry Out alpha and beta Galactosylation of Secondary Cell
1386 Wall Polysaccharide in Bacillus anthracis. *J Bacteriol* **202** (2020). https://doi.org:10.1128/JB.00191-
1387 20
1388 60 Rismondo, J., Percy, M. G. & Grundling, A. Discovery of genes required for lipoteichoic acid
1389 glycosylation predicts two distinct mechanisms for wall teichoic acid glycosylation. *J Biol Chem* **293**,
1390 3293-3306 (2018). https://doi.org:10.1074/jbc.RA117.001614
1391 61 Endo, T., Ikeo, K. & Gojobori, T. Large-scale search for genes on which positive selection may
1392 operate. *Mol Biol Evol* **13**, 685-690 (1996). https://doi.org:10.1093/oxfordjournals.molbev.a025629

1393 62 Blanc, G. *et al.* Molecular evolution of rickettsia surface antigens: evidence of positive selection.
1394 *Mol Biol Evol* **22**, 2073-2083 (2005). https://doi.org:10.1093/molbev/msi199
1395 63 Hunault, L. *et al.* A monoclonal antibody collection for C. difficile typing ? *Gut Pathog* **16**, 4 (2024).
1396 https://doi.org:10.1186/s13099-023-00592-7
1397 64 Chandra, H., Kovall, R. A., Yadav, J. S. & Sun, X. Host Immune Responses to Surface S-Layer Proteins
1398 (SLPs) of Clostridioides difficile. *Microorganisms* **11** (2023).
1399 https://doi.org:10.3390/microorganisms11020380
1400 65 Sara, M. & Sleytr, U. B. Comparative studies of S-layer proteins from Bacillus stearothermophilus
1401 strains expressed during growth in continuous culture under oxygen-limited and non-oxygen-
1402 limited conditions. *J Bacteriol* **176**, 7182-7189 (1994). https://doi.org:10.1128/jb.176.23.7182-
1403 7189.1994
1404 66 Avall-Jaaskelainen, S. & Palva, A. Lactobacillus surface layers and their applications. *FEMS Microbiol*
1405 *Rev* **29**, 511-529 (2005). https://doi.org:10.1016/j.femsre.2005.04.003
1406 67 Moore, A. D., Bjorklund, A. K., Ekman, D., Bornberg-Bauer, E. & Elofsson, A. Arrangements in the
1407 modular evolution of proteins. *Trends Biochem Sci* **33**, 444-451 (2008).
1408 https://doi.org:10.1016/j.tibs.2008.05.008
1409 68 Pereira-Leal, J. B., Levy, E. D. & Teichmann, S. A. The origins and evolution of functional modules:
1410 lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci* **361**, 507-517 (2006).
1411 https://doi.org:10.1098/rstb.2005.1807
1412 69 von Kugelgen, A. *et al.* Interdigitated immunoglobulin arrays form the hyperstable surface layer of
1413 the extremophilic bacterium Deinococcus radiodurans. *Proc Natl Acad Sci U S A* **120**, e2215808120
1414 (2023). https://doi.org:10.1073/pnas.2215808120
1415 70 Sagmeister, T. *et al.* The molecular architecture of Lactobacillus S-layer: Assembly and attachment
1416 to teichoic acids. *Proc Natl Acad Sci U S A* **121**, e2401686121 (2024).
1417 https://doi.org:10.1073/pnas.2401686121
1418 71 Tikhonova, T. V. *et al.* Extracellular Fe(III) reductase structure reveals a modular organization
1419 enabling S-layer insertion and electron transfer to insoluble substrates. *Structure* **31**, 174-184 e173
1420 (2023). https://doi.org:10.1016/j.str.2022.12.010
1421 72 Navarre, W. W. & Schneewind, O. Surface proteins of gram-positive bacteria and mechanisms of
1422 their targeting to the cell wall envelope. *Microbiol Mol Biol Rev* **63**, 174-229 (1999).
1423 https://doi.org:10.1128/MMBR.63.1.174-229.1999
1424 73 Lopez, R., Garcia, E., Garcia, P. & Garcia, J. L. The pneumococcal cell wall degrading enzymes: a
1425 modular design to create new lysins? *Microb Drug Resist* **3**, 199-211 (1997).
1426 https://doi.org:10.1089/mdr.1997.3.199
1427 74 Megrian, D., Taib, N., Jaffe, A. L., Banfield, J. F. & Gribaldo, S. Ancient origin and constrained
1428 evolution of the division and cell wall gene cluster in Bacteria. *Nat Microbiol* **7**, 2114-2127 (2022).
1429 https://doi.org:10.1038/s41564-022-01257-y
1430 75 Megrian, D., Taib, N., Witwinowski, J., Beloin, C. & Gribaldo, S. One or two membranes? Diderm
1431 Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol Microbiol* **113**, 659-671 (2020).
1432 https://doi.org:10.1111/mmi.14469
1433 76 Witwinowski, J. *et al.* An ancient divide in outer membrane tethering systems in bacteria suggests a
1434 mechanism for the diderm-to-monoderm transition. *Nat Microbiol* **7**, 411-422 (2022).
1435 https://doi.org:10.1038/s41564-022-01066-3
1436 77 Lagkouvardos, I. *et al.* The Mouse Intestinal Bacterial Collection (miBC) provides host-specific
1437 insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol* **1**, 16131
1438 (2016). https://doi.org:10.1038/nmicrobiol.2016.131
1439 78 Hochberg, G. K. A. *et al.* A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503-
1440 508 (2020). https://doi.org:10.1038/s41586-020-3021-2
1441 79 Lukes, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. & Gray, M. W. How a neutral evolutionary
1442 ratchet can build cellular complexity. *IUBMB Life* **63**, 528-537 (2011).
1443 https://doi.org:10.1002/iub.489
1444 80 Isbilir, B., Yeates, A., Alva, V. & Bharat, T. A. M. Mapping the ultrastructural topology of the
1445 corynebacterial cell surface. *bioRxiv* (2024). https://doi.org:10.1101/2024.09.05.611374

1446 81 Roberts, K. Crystalline glycoprotein cell walls of algae: their stucture, composition and assembly.
1447 *Philos Trans R Soc Lond B Biol Sci* **268**, 129-146 (1974). https://doi.org:10.1098/rstb.1974.0021
1448 82 Baudelet, P.-H., Ricochon, G., Linder, M. & Muniglia, L. A new insight into cell walls of Chlorophyta.
1449 *Algal Research* **25**, 333-371 (2017). https://doi.org:10.1016/j.algal.2017.04.008
1450 83 Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon.
1451 *Nature* **613**, 332-339 (2023). https://doi.org:10.1038/s41586-022-05550-y
1452 84 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519-
1453 525 (2020). https://doi.org:10.1038/s41586-019-1916-6
1454 85 Sexton, D. L. *et al.* The cell envelope of Thermotogae suggests a mechanism for outer membrane
1455 biogenesis. *Proc Natl Acad Sci U S A* **120**, e2303275120 (2023).
1456 https://doi.org:10.1073/pnas.2303275120
1457 86 Hashimi, A. & Tocheva, E. I. Cell envelope diversity and evolution across the bacterial tree of life.
1458 *Nat Microbiol* **9**, 2475-2487 (2024). https://doi.org:10.1038/s41564-024-01812-9
1459 87 Sendra, K. M. *et al.* An ancient metalloenzyme evolves through metal preference modulation. *Nat*
1460 *Ecol Evol* **7**, 732-744 (2023). https://doi.org:10.1038/s41559-023-02012-0
1461 88 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
1462 https://doi.org:10.1371/journal.pcbi.1002195
1463 89 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements
1464 in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
1465 https://doi.org:10.1093/molbev/mst010
1466 90 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment
1467 trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
1468 https://doi.org:10.1093/bioinformatics/btp348
1469 91 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and
1470 bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
1471 https://doi.org:10.1093/bioinformatics/btp163
1472 92 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic
1473 algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2015).
1474 https://doi.org:10.1093/molbev/msu300
1475 93 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the
1476 Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522 (2018).
1477 https://doi.org:10.1093/molbev/msx281
1478 94 Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).
1479 https://doi.org:10.1093/nar/gks1195
1480 95 Halary, S., McInerney, J. O., Lopez, P. & Bapteste, E. EGN: a wizard for construction of gene and
1481 genome similarity networks. *BMC Evol Biol* **13**, 146 (2013). https://doi.org:10.1186/1471-2148-13-
1482 146
1483 96 Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of
1484 diversity. *BMC Bioinformatics* **19**, 164 (2018). https://doi.org:10.1186/s12859-018-2164-8
1485 97 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for
1486 large alignments. *PLoS One* **5**, e9490 (2010). https://doi.org:10.1371/journal.pone.0009490
1487 98 Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences
1488 using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*
1489 **313**, 903-919 (2001). https://doi.org:10.1006/jmbi.2001.5080
1490 99 Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0
1491 database: a significant proteome update and a new webserver. *Nucleic Acids Res* **47**, D490-D494
1492 (2019). https://doi.org:10.1093/nar/gky1130
1493 100 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419
1494 (2021). https://doi.org:10.1093/nar/gkaa913
1495 101 Sillitoe, I. *et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Res* **49**, D266-
1496 D273 (2021). https://doi.org:10.1093/nar/gkaa1079
1497 102 Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein
1498 family model curation. *Nucleic Acids Res* **49**, D1020-D1028 (2021).
1499 https://doi.org:10.1093/nar/gkaa1105

1500 103 Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41**, D387-395 (2013).
1501 https://doi.org:10.1093/nar/gks1234
1502 104 Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res* **34**, D227-230 (2006).
1503 https://doi.org:10.1093/nar/gkj063
1504 105 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J*
1505 *Mol Biol* **215**, 403-410 (1990). https://doi.org:10.1016/S0022-2836(05)80360-2
1506 106 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**,
1507 1236-1240 (2014). https://doi.org:10.1093/bioinformatics/btu031
1508 107 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682
1509 (2022). https://doi.org:10.1038/s41592-022-01488-1
1510 108 Csardi, G. & Nepusz, T. The igraph software. *Complex syst* **1695**, 1-9 (2006).
1511 109 Fang, H. & Gough, J. The 'dnet' approach promotes emerging research on cancer patient survival.
1512 *Genome Med* **6**, 64 (2014). https://doi.org:10.1186/s13073-014-0064-8
1513 110 Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical
1514 Computing. *(No Title)* (2013).
1515 111 Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data
1516 integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).
1517 https://doi.org:10.1093/bioinformatics/btq675
1518 112 Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved
1519 genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).
1520 https://doi.org:10.1101/gr.2289704
1521 113 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100
1522 (2018). https://doi.org:10.1093/bioinformatics/bty191
1523 114 Hackl, T. & Ankenbrand, M. J. gggenomes: a grammar of graphics for comparative genomics. *R*
1524 *package version 0.9* **5** (2022).
1525 115 Bletz, S., Janezic, S., Harmsen, D., Rupnik, M. & Mellmann, A. Defining and Evaluating a Core
1526 Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of Clostridium difficile. *J*
1527 *Clin Microbiol* **56** (2018). https://doi.org:10.1128/JCM.01987-17
1528 116 Shen, E. *et al.* Subtyping analysis reveals new variants and accelerated evolution of Clostridioides
1529 difficile toxin B. *Commun Biol* **3**, 347 (2020). https://doi.org:10.1038/s42003-020-1078-y
1530 117 Roy Chowdhury, P. *et al.* Comparative genomic analysis of toxin-negative strains of Clostridium
1531 difficile from humans and animals with symptoms of gastrointestinal disease. *BMC Microbiol* **16**, 41
1532 (2016). https://doi.org:10.1186/s12866-016-0653-3
1533 118 Mansfield, M. J. *et al.* Phylogenomics of 8,839 Clostridioides difficile genomes reveals
1534 recombination-driven evolution and diversification of toxin A and B. *PLoS Pathog* **16**, e1009181
1535 (2020). https://doi.org:10.1371/journal.ppat.1009181
1536 119 Knight, D. R., Elliott, B., Chang, B. J., Perkins, T. T. & Riley, T. V. Diversity and Evolution in the
1537 Genome of Clostridium difficile. *Clin Microbiol Rev* **28**, 721-741 (2015).
1538 https://doi.org:10.1128/CMR.00127-14
1539 120 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-
1540 redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-
1541 504 (2005). https://doi.org:10.1093/nar/gki025
1542 121 Sayers, E. W. *et al.* Database resources of the national center for biotechnology information.
1543 *Nucleic Acids Res* **50**, D20-D26 (2022). https://doi.org:10.1093/nar/gkab1112
1544 122 Dingle, K. E. *et al.* Effects of control interventions on Clostridium difficile infection in England: an
1545 observational study. *Lancet Infect Dis* **17**, 411-421 (2017). https://doi.org:10.1016/S1473-
1546 3099(16)30514-X
1547 123 Frentrup, M. *et al.* A publicly accessible database for Clostridioides difficile genome sequences
1548 supports tracing of transmission chains and epidemics. *Microb Genom* **6** (2020).
1549 https://doi.org:10.1099/mgen.0.000410
1550 124 Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia
1551 pestis phylogeny, and Escherichia core genomic diversity. *Genome Res* **30**, 138-152 (2020).
1552 https://doi.org:10.1101/gr.251678.119

125 Harvala, H., Alm, E., Akerlund, T. & Rizzardi, K. Emergence and spread of moxifloxacin-resistant Clostridium difficile ribotype 231 in Sweden between 2006 and 2015. *New Microbes New Infect* **14**, 58-66 (2016). https://doi.org:10.1016/j.nmni.2016.09.002

126 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org:10.1186/1471-2105-11-119

127 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019). https://doi.org:10.1186/s13059-019-1832-y

128 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015). https://doi.org:10.1186/s13059-015-0721-2

129 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60 (2015). https://doi.org:10.1038/nmeth.3176

130 Bansal, M. S., Kellis, M., Kordi, M. & Kundu, S. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* **34**, 3214-3216 (2018). https://doi.org:10.1093/bioinformatics/bty314

131 Knight, D. R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10** (2021). https://doi.org:10.7554/eLife.64325

132 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017). https://doi.org:10.1038/nmeth.4285

133 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612 (2006). https://doi.org:10.1093/nar/gkl315

134 McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013). https://doi.org:10.1093/nar/gkt376

135 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019). https://doi.org:10.1093/bioinformatics/bty633

136 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007). https://doi.org:10.1093/molbev/msm088

137 Yang, Z. & Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* **25**, 568-579 (2008). https://doi.org:10.1093/molbev/msm284

138 Alvarez-Carretero, S., Kapli, P. & Yang, Z. Beginner's Guide on the Use of PAML to Detect Positive Selection. *Mol Biol Evol* **40** (2023). https://doi.org:10.1093/molbev/msad041

139 Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107-1118 (2005). https://doi.org:10.1093/molbev/msi097

140 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**, 1891-1901 (2006). https://doi.org:10.1093/molbev/msl051

141 Murrell, B. *et al.* FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol* **30**, 1196-1205 (2013). https://doi.org:10.1093/molbev/mst030

142 Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol Biol Evol* **32**, 1365-1371 (2015). https://doi.org:10.1093/molbev/msv035

143 Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D. & Muse, S. V. 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). https://doi.org:10.1093/molbev/msaa037

144 Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764 (2012). https://doi.org:10.1371/journal.pgen.1002764

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

146 Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* **7**, veaa087 (2021). https://doi.org:10.1093/ve/veaa087

147 Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563 (2000). https://doi.org:10.1093/bioinformatics/16.6.562
148 Padidam, M., Sawyer, S. & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225 (1999). https://doi.org:10.1006/viro.1999.0056
149 Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98-102 (2005). https://doi.org:10.1089/aid.2005.21.98
150 Smith, J. M. Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129 (1992). https://doi.org:10.1007/BF00182389
151 Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**, 13757-13762 (2001). https://doi.org:10.1073/pnas.241370698
152 Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582 (2000). https://doi.org:10.1093/bioinformatics/16.7.573
153 Lam, H. M., Ratmann, O. & Boni, M. F. Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Mol Biol Evol* **35**, 247-251 (2018). https://doi.org:10.1093/molbev/msx263
154 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009). https://doi.org:10.1093/bioinformatics/btp033
155 Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**, 356 (2009). https://doi.org:10.1186/1471-2105-10-356
156 Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives* **17**, 160-167 (2019). https://doi.org:10.1080/15366367.2019.1565254
157 Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A grammar of data manipulation. *R package version* **1.1.4** (2023).
158 Zhou, L. *et al.* ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief Bioinform* **23** (2022). https://doi.org:10.1093/bib/bbac222
159 Yu, G. *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). https://doi.org:10.1111/2041-210x.12628
160 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004). https://doi.org:10.1093/bioinformatics/btg412
161 Wang, L. G. *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). https://doi.org:10.1093/molbev/msz240
162 Xu, S. *et al.* ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Mol Biol Evol* **38**, 4039-4042 (2021). https://doi.org:10.1093/molbev/msab166
163 Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. *R package version 0.4* **5** (2020).
164 Yamakawa, K., Kamiya, S., Meng, X. Q., Karasawa, T. & Nakamura, S. Toxin production by Clostridium difficile in a defined medium with limited amino acids. *J Med Microbiol* **41**, 319-323 (1994). https://doi.org:10.1099/00222615-41-5-319
165 Muh, U., Pannullo, A. G., Weiss, D. S. & Ellermeier, C. D. A Xylose-Inducible Expression System and a CRISPR Interference Plasmid for Targeted Knockdown of Gene Expression in Clostridioides difficile. *J Bacteriol* **201** (2019). https://doi.org:10.1128/JB.00711-18
166 Mendoza, B. J. & Trinh, C. T. 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). https://doi.org:10.1093/bioinformatics/btx564
167 Rabe, B. A. & Cepko, C. A Simple Enhancement for Gibson Isothermal Assembly. *BioRxiv* (2020). https://doi.org:10.1101/2020.06.14.150979

168 Kirk, J. A. & Fagan, R. P. Heat shock increases conjugation efficiency in Clostridium difficile. *Anaerobe* **42**, 1-5 (2016). https://doi.org:10.1016/j.anaerobe.2016.06.009
169 Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* **3**, 1101-1108 (2008). https://doi.org:10.1038/nprot.2008.73
170 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012). https://doi.org:10.1038/nmeth.2019
171 Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J Struct Biol* **116**, 71-76 (1996). https://doi.org:10.1006/jsbi.1996.0013
172 Fagan, R. & Fairweather, N. Dissecting the cell surface. *Methods Mol Biol* **646**, 117-134 (2010). https://doi.org:10.1007/978-1-60327-365-7_8
173 Kotiranta, A., Haapasalo, M., Lounatmaa, K. & Kari, K. Crystalline surface protein of Peptostreptococcus anaerobius. *Microbiology (Reading)* **141 (Pt 5)**, 1065-1073 (1995). https://doi.org:10.1099/13500872-141-5-1065
174 Richards, E. *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). https://doi.org:10.1074/jbc.RA118.004530
175 Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186-190 (2009). https://doi.org:10.1107/s0021889809045701
176 Winter, G. *et al.* DIALS as a toolkit. *Protein Sci* **31**, 232-250 (2022). https://doi.org:10.1002/pro.4224
177 Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214 (2013). https://doi.org:10.1107/S0907444913000061
178 Agirre, J. *et al.* The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **79**, 449-461 (2023). https://doi.org:10.1107/S2059798323003595
179 McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007). https://doi.org:10.1107/S0021889807021206
180 Bond, P. S. & Cowtan, K. D. ModelCraft: an advanced automated model-building pipeline using Buccaneer. *Acta Crystallogr D Struct Biol* **78**, 1090-1098 (2022). https://doi.org:10.1107/S2059798322007732
181 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004). https://doi.org:10.1107/S0907444904019158
182 Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367 (2011). https://doi.org:10.1107/S0907444911001314
183 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019). https://doi.org:10.1107/S2059798319011471
184 Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220 (2014). https://doi.org:10.1107/S2052252514009324
185 Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315 (2018). https://doi.org:10.1002/pro.3330
186 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243-246 (2024). https://doi.org:10.1038/s41587-023-01773-0
187 Holm, L. DALI and the persistence of protein shape. *Protein Sci* **29**, 128-140 (2020). https://doi.org:10.1002/pro.3749
188 Laskowski, R. A., Jablonska, J., Pravda, L., Varekova, R. S. & Thornton, J. M. PDBsum: Structural summaries of PDB entries. *Protein Sci* **27**, 129-134 (2018). https://doi.org:10.1002/pro.3289
189 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797 (2007). https://doi.org:10.1016/j.jmb.2007.05.022
190 Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023). https://doi.org:10.1002/pro.4792

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.