# Exploring Host-Binding Machineries of Mycobacteriophages with AlphaFold2

Christian Cambillau, Adeline Goulet

# Exploring host-binding machineries of mycobacteriophages with AlphaFold2

Christian Cambillau[a,b] and Adeline Goulet[c]#

[a]School of Microbiology, University College Cork, Cork, Ireland

[b]AlphaGraphix, Formiguères, 66210, France

[c]Laboratoire d'Ingénierie des Systèmes Macromoléculaires (LISM), Institut de Microbiologie, Bioénergies et Biotechnologie (IM2B), CNRS and Aix-Marseille Université UMR7255, Marseille, France

Running title: Structural insights into mycobacteriophages' tail tips

#Address correspondence to Adeline Goulet, adeline.goulet@univ-amu.fr.

**Keywords:** bacteriophage, *Mycobacteria*, AlphaFold2, host-binding machineries, carbohydrate-binding module, receptor-binding protein, polyglycine helices

## Abstract

Although more than 12,000 bacteriophages infecting mycobacteria (or mycobacteriophages) have been isolated so far, there is a knowledge gap on their structure-function relationships. Here, we have explored the architecture of host-binding machineries from seven representative mycobacteriophages of the *Siphoviridae* family infecting *Mycobacterium smegmatis*, *Mycobacterium abscessus* and *Mycobacterium tuberculosis*, using AlphaFold2 (AF2). AF2 enables confident structural analyses of large and flexible biological assemblies resistant to experimental methods, thereby opening new avenues to shed light on phage structure and function. Our results highlight the modularity and structural diversity of siphophage host-binding machineries that recognize host-specific receptors at the onset of viral infection. Interestingly, the studied mycobacteriophages' host-binding machineries present unique features as compared with those of phages infecting other Gram-positive actinobacteria. Although they all assemble the classical Dit (Distal tail), Tal (Tail-associated lysin) and receptor-binding proteins, five of them contain two potential additional adhesion proteins. Moreover, we have identified brush-like domains formed of multiple polyglycine helices, which expose hydrophobic residues, as potential receptor-binding domains. These polyglycine-rich domains, which have been observed in only five native proteins, may be a hallmark of mycobacteriophages' host-binding machineries, and more common in nature than expected. Altogether, the unique composition of mycobacteriophages' host-binding machineries indicate they might have evolved to bind to the peculiar mycobacterial cell envelope rich in polysaccharides and mycolic acids. This work provides a rational framework to efficiently produce recombinant proteins or protein domains and test their host-binding function, and hence to shed light on molecular mechanisms used by mycobacteriophages to infect their host.


## Importance

Mycobacteria include both saprophytes, such as the model system *Mycobacterium smegmatis*, and pathogens, such as *Mycobacterium tuberculosis* and *Mycobacterium abscessus* that are poorly responsive to antibiotic treatments and pose a global public health problem. Mycobacteriophages have been collected at a very large scale over the last decade, and they have proven to be valuable tools for mycobacteria genetic manipulation, rapid diagnostics and infection treatment. Yet, molecular mechanisms used by mycobacteriophages to infect their host remain poorly understood. Therefore, exploring the structural diversity of mycobacteriophages' host-binding machineries is important not only to better understand viral diversity and bacteriophage-host interactions, but also to rationally develop biotechnological tools. With the powerful protein structure prediction software AlphaFold2, which was publicly released a year ago, it is now possible to get structural and functional insights on such challenging assemblies.

## Introduction

Mycobacteria are members of the Gram-positive *Actinobacteria* phylum ubiquitously found in terrestrial and aquatic ecosystems (1, 2). They include both saprophytes, such as the model system *Mycobacterium smegmatis*, and pathogens, such as *Mycobacterium tuberculosis* and *Mycobacterium abscessus* that are poorly responsive to antibiotic treatments and pose a global public health problem (3). One of the distinguishing features of *Mycobacteria* is their thick, waxy cell wall rich in mycolic acids that gives them many of their biological properties and acts as a protective barrier against harsh environmental conditions (4, 5). Viruses that infect mycobacteria, or mycobacteriophages, have been collected at a very large scale over the last decade thanks to the work of phage hunters enrolled in integrated research-education programs such as SEA-PHAGES (Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Sciences) (6). To date, more than 12, 000 mycobacteriophages have been identified and over 2,000 genomes have been sequenced, representing more than half of phages and genomic sequences available on Actinobacteriophages (https://phagesdb.org/, as for September 2022, (7)). Based on their remarkable genomic diversity, mycobacteriophages are classified into 31 clusters (clusters A-Z and AA-AE), further divided into subclusters, and 7 singletons (phages with no other close relatives) (8). This substantial collection of mycobacteriophages is highly valuable to better understand phage diversity and evolution, phage-host interactions, and to develop phage-based strategies to detect and treat mycobacterial infections (9–13). Yet, molecular mechanisms of mycobacteriophage infections are poorly understood. In particular, the architecture of mycobacteriophages' host-binding machineries and their mode of action to bind to the host cell surface at the onset of viral infection are not known.

While only one paper reports low-resolution structural data of the host-binding machinery of the *M. abscessus* siphophage Araucaria (14), many structural and functional analyses of host-binding machineries from siphophages infecting other Gram-positive bacteria have revealed molecular bases of their architecture and function (15–24). These machineries assemble an hexameric ring of the Dit protein (distal tail protein) bound to the last ring of major tail protein (MTP) and to a trimer of Tal (tail-associated lysin). Dit proteins are made up of a N-terminal domain, the fold of which is conserved in tail proteins such as MTP, head-to-tail joining proteins, and Dit of *Siphoviridae*, *Myoviridae* and phage tail-like system (23–28), and a C-terminal galectin domain (for an example see Fig. 2 DS6A). The Dit N-terminal domain is referred to as the belt domain because of its canonical belt loop assembling the ring. Tal proteins are composed of a N-terminal structural domain of ~350-400 residues, similar to the myophage T4 gp27 protein (29) (for an example see Fig. 3 Abinghost), that can be followed by a C-terminal extension of 1,000 to 2,000 residues (18, 19, 30). This extension is believed to play a role in cell wall polysaccharide/peptidoglycan degradation, e.g. the *Lactococcus lactis* P335 phage TP901-1 (31), or host binding, e.g. the *Bacillus subtilis* and *Escherichia coli* phages SPP1 and T5 (32, 33). This Dit-Tal common

core may be decorated by receptor-binding proteins (RBP) and other proteins resulting in a variety of host-binding machinery compositions and architectures. Moreover, Dit and Tal can be functionalized with carbohydrate-binding modules, involved in host cell surface polysaccharide binding, and lysin domains (16, 18, 19, 34, 35). In particular, the Tal C-terminal extensions of *Streptococcus thermophilus* and *Oenococcus oeni* phages contain one to five carbohydrate-binding modules (CBM). However, even though siphophages' host-binding machineries share common structural and functional domains, their compositional and architectural diversity leads to different host-binding mechanisms.

Since host-binding machineries bind to host-specific cell surface components, we reasoned that those of mycobacteriophages might differ from those of previously-mentioned phages. Therefore, in order to get structural insights into mycobacteriophages' host-binding machineries, we have performed AlphaFold2 (AF2) structure predictions of seven representative siphophages infecting *M. smegmatis*, *M. abscessus* and *M. tuberculosis*. AF2 overcomes bottlenecks of other structure prediction methods such as HHpred (36) and I-TASSER (37) that rely on the presence of structural homologs in the Protein Data Bank (PDB), and produces 3D structures of full-length proteins, multimeric assemblies and protein-protein complexes that equally well compare with experimental structures when AF2 confidence scores are carefully considered (38–41). It therefore positions as a powerful technique providing reliable structural information on flexible and multi-domains proteins that cannot be otherwise analysed using experimental techniques such as X-ray crystallography and cryo-electron microscopy (18, 19, 30). We have predicted the structures of Dits and Tals, and of potential RBPs and additional adhesion proteins assembling these machineries. Interestingly, these mycobacteriophage proteins share common features with those of other Gram-positive-infecting siphophages, such as the presence of CBMs, and also present unique features. In particular, brush-like polyglycine-rich domains with surface-exposed hydrophobic residues are present in potential RBPs and/or additional adhesion proteins, and could be involved in the recognition and binding of the mycobacterial cell wall. All in all, these mycobacteriophage protein structures confirm the mosaic and versatile nature of phage host-binding machineries evolved to bind to a specific host and to initiate the viral infection cycle.

## Results

We have selected seven mycobacteriophages that belong to different genomic clusters, infect different *Mycobacterium* species, and assemble host-binding machineries of different composition (Fig. 1A). Abinghost, Badfish, Butters and Wildcat infect *M. smegmatis* and belong to different clusters (clusters N and V for Butters and Wildcat, respectively) and subclusters (subclusters B1 and B3 for Badfish and Abinghost, respectively). Isca infects *M. abscessus* and belongs to the Subcluster A3. Araucaria also infects *M. abscessus* but it has not been assigned to a cluster yet. Lastly, DS6A infects *M. tuberculosis* and is a

singleton. Genes coding for proteins assembling siphophage host-binding machineries are, in most cases, comprised in sequence between the long *tmp* (gene coding for the tape measure protein that defines the length of the tail) and the *holin*/*lysin* cassette (42). Genome analyses in between the *tmp* and the first *lysin* of these phages have revealed different compositions of their host-binding machineries (Fig. 1B). Of note, *lysin A* and *lysin B* in Isca correspond to ORFs 9 and 10 upstream the *tmp*. We have found that three to five proteins of more than 150 residues assemble mycobacteriophages' host-binding machineries including Dit, Tal, potential RBP and potential additional adhesion proteins (Fig. 1A and B).

**Predicted structures of Dits**

The Dit size of the seven analysed mycobacteriophages varies from 323 residues for the shortest (DS6A) to 573 residues for the longest (Butters) (Fig. 1A). The Dit of DS6A is a classical Dit containing only the belt and galectin domains (Fig. 2). However, the galectin loops are slightly longer than those of Dit from siphophages infecting the lactic acid bacteria (LAB) *L. lactis*, *S. thermophilus*, and *O. oeni* (18, 19, 23). Even though the Dit of Abinghost, Araucaria and Badfish are ~150 residues longer than that of DS6A, their predicted structure show that they are also classical Dits devoid of any CBM insertion. Instead, they possess two long loops in the belt domain (residues 24-107 in the Abinghost's Dit, residues 25-76 in the Araucaria and Badfish's Dit) and the galectin domain (residues 296-343 in the Abinghost's, Araucaria's and Badfish's Dits) (Fig. 2). It is worth noting that these flexible loops are likely to adopt conformations different from those predicted by AF2 (Fig. S1). Moreover, their location at the periphery of the Dit ring indicate that they may be involved in establishing contacts with other components of the host-binding machinery or with the host cell surface. Lastly, the Dit of Isca, Butters and Wildcat present an additional domain inserted in the galectin domain. When submitted to the Dali server, the predicted structure of the Isca's Dit 110-residue inserted domain returned a hit with a domain of the myophage T4 gp11 protein found at the interface between the baseplate and the short tail fibers (Table 1). However, the inserted domain in the longest Butter's and Wildcat's Dits has been identified as a CBM family 66 (CBM66) according to the Carbohydrate-Active EnZymes database (CAZy) nomenclature (43) (Table 1).

**Predicted structures of Tals**

The Tal size of the seven analysed mycobacteriophages varies from 345 residues for the shortest (Araucaria) to 794 residues for the longest (Wildcat) (Fig. 1A). However, they all assemble bulky trimers devoid of elongated extensions like those observed in the Tals of some LAB-infecting siphophages (19, 21, 30) (Fig. 3). Abinghost, Araucaria and Badfish possess short Tals of similar size (347-370 residues, Fig. 1A). Although their predicted structures superimpose well onto each other (rmsd of 0.8-1.0 Å on Cα atoms) and are similar to that of the Tal of *Neisseria meningitidis* MC58 (Table 1), the Tal of Badfish harbours

longer N- and C-termini (Fig. 3). Butters, DS6A and Isca possess Tal of medium size (569-599 residues), yet their predicted structures show that they do not contain extra domains. Instead, they incorporate extra α-helices at the N- and C-terminal ends, as well as extra loops within the conserved backbone (Fig. 3 and S1). Lastly, Wildcat has a long, 794-residue Tal that contains an additional N-terminal 237-residue insertion including a lectin-like domain similar to that of a Trypanosoma sialidase (44) (Table1) a linker, and a long α-helix connected to the Tal central core (Fig. 3).

**Predicted structures of potential RBPs and additional adhesion proteins**

In siphophages infecting Gram-positive bacteria, *bona fide* RBPs, responsible for irreversible binding of a phage to host-specific receptors, are usually encoded by the *orf* adjacent to the *tal* (42, 45). Yet, the order of genes coding for components assembling siphophages' host-binding machineries may vary when *orfs* of ancillary proteins are inserted between the *tal* and *rbp* (42). Moreover, RBPs typically assemble as trimers. Based on these criteria, we have identified candidate *orfs* coding for long RBPs, from 619 residues in Isca to 855 residues in Wildcat (Fig. 1A), that assemble well-packed trimers (Fig. 4). It should be noted that, in Butters, we propose the *orf20*, and not the *tal*-adjacent *orf19*, as being the RBP-coding *orf*. We have made this choice based on the overall structural similarity between the Butters's ORF20 predicted trimer and other phages' RBP trimers (Fig. 4). In addition to RBPs, all mycobacteriophages, but Araucaria and Isca, also have potential additional adhesion proteins, which are adjacent to each other or separated by small ORFs (Fig. 1B). In order to get structural and functional insights into these proteins, we have predicted their structures as monomers, unless otherwise stated.

**Mycobacteriophages' potential RBPs are elongated, multi-domain proteins.** The predicted structures of trimers of potential RBPs form elongated assemblies from 250 Å for the shortest (Isca) to 450 Å for the longest (Araucaria) (Fig. S1). They are all composed of axial α-helical segments, giving them their rod shape, with unstructured segments and/or α-helices at their N-terminal end that may be involved in anchoring the RBPs to the central Dit-Tal core. A variety of globular domains are found at their C-terminal end and in their central region, except for the RBP of Araucaria, DS6A and Isca (Fig. 4).

***Araucaria and Isca.*** Araucaria and Isca possess only one RBP. The predicted structure of Araucaria's RBP is composed of a 281-residue α-helical rod, interrupted by a small β-prism (residues 31-56) and followed by three bulky domains. The first domain (residues 311-487), pointing out from the trimer axis, looks like a painter's brush in which a 10-stranded β-sandwich forms the brush handle, and a bundle of 10 left-handed polyglycine type II ($PG_{II}$) helices forms the brush hairs (Figs .4, 6 and S2). Interestingly, although $PG_{II}$ helix-containing domains have rarely been observed in nature (46–50), the C-terminal domain of the *Salmonella* phage S16 tail fiber adhesin also contains a brush domain involved in host recognition (51). Since we have identified brush domains in other mycobacteriophages' RBPs and additional adhesion

proteins, they will be analysed altogether in a next section. The Araucaria' brush domain is followed by a CBM, similar to a CBM66 from a fructosidase specifically hydrolysing levan (52) (Table 1), which is itself followed by a lectin domain similar to the N-terminal domain of the *Burkholderia cenocepacia* superlectin C (BC2L-C ) with dual carbohydrate specificity (53) (Table 1) (Fig. 4). Interestingly, this tandem of C-terminal CBM and lectin domain is similar to that found in a *Klebsellia* phage depolymerase that binds to and degrades host cell surface polysaccharides (54). Of note, the three lectin domains in Araucaria's RBP assemble a tight trimer as that of the *Klebsiella* phage depolymerase, which is critical for the trimer stability (54). These CBM and lectin domain are likely involved in the recognition of mycobacteria surface polysaccharides. Although the 240 Å-long Isca's RBP predicted structure is shorter than that of Araucaria and devoid of brush domain, it is overall similar to that of Araucaria. It starts with a 39-residue unstructured stretch followed by α-helical segments, and it is terminated by a CBM and a lectin domain. The CBM also returned a levanase CBM66 as a significant hit using Dali (Table 1). As for the lectin domain, which also forms a tightly packed trimer, it returned the BC2L-C lectin from *B. cenocepacia* as well as the receptor-binding domain of the *L. lactis* phage 1358 RBP (Table 1) as structural homologs, which is in agreement with a host-binding function for this protein.

***DS6A.*** DS6A is a singleton, which led to predicted structures with low confidence scores in some regions (Fig. S1). Therefore, we show only well-folded domains without junctions in between. In particular, a CBM and a lectin domain assemble the RBP C-terminal end (Fig. 4). The CBM also returned the levanase CBM66 as significant hit, and the lectin domain is similar to that of *L. lactis* phage 1358 RBD (Table 1).

***Abinghost, Badfish, Butters and Wildcat.*** The RBPs of Abinghost, Badfish, Butters and Wildcat all possess a central CBM similar to a CBM16 (Table 1). In the Abinghost's and Badfishs's RBPs, this CBM is followed by α-helical segments connected to a second CBM identified by Dali as a levanase CBM66 (Table 1), like that found in the Araucaria's and Isca's RBP. However, in the Abinghost's and Badfish's RBP, the C-terminal end is formed by a tightly packed, trimeric knob assembling small domains with a new fold composed of four β-strands and one α-helix (Fig. 4). On the other hand, the RBP of Butters is terminated by a brush domain (residues 486-738), and not by a lectin domain, the overall fold of which differ from that of the Araucaria RBP's brush domain. In particular, the handle comprises an 8-stranded, 2-layer β-sandwich, and the hairs, assembling 12 $PG_{II}$ helices, possess short β-sheets packed onto each other at their tip (Fig. 4 and 6). As for the RBP of Wildcat, the central CBM is followed by α-helical segments abutting to a well-packed trimer of a second CBM at the apex of the RBP. This second CBM shares structural similarities with a beta-agarase CBM6 (Table 1) (55), and is followed by a brush domain that superimposes well on that of the Araucaria's RBP, except that the latter has slightly longer loops at the brush hair tips. To note, the brush domains are packed against the central α-helices and placed above the second CBMs in the Wildcat RBP predicted structure. However, since a 12-residue linker connects the second CBM to the

brush domain (Fig. 4), its intrinsic flexibility may also place the brush domains below the second CBM at the RBP tip, or in a variety of positions in between being stacked against the helical segments and being aligned along the trimeric axis, pointing towards the RBP C-terminal end.

**Mycobacteriophages' potential additional adhesion proteins.** The predicted structures of potential additional adhesion proteins contain one or two brush domains, except those of Butters and Wildcat (Fig. 5). However, the potential RBPs of Butters and Wildcat contain a brush domain, as does the RBP of Araucaria that has a host-binding machinery devoid of additional adhesion proteins (Fig. 1A). This indicates that brush domains likely play a role in mycobacteria recognition and binding.

***Abinghost, Badfish*** The predicted structures of Abinghost's and Badfish's additional adhesion proteins (ORF32 and ORF33) are very close to each other (Fig. 5), which was also the case for their RBPs. This is in agreement with the fact that these phages belong to the same cluster B. Their adhesion protein 1 (ORF32) starts with a short N-terminal segment and a β-hairpin, which are followed by a tandem of brush domains. The first N-terminal brush domain comprises an 8-stranded, 2-layer β-sandwich and 8 $PG_{II}$ helices, while the second C-terminal brush domain comprises a 10-stranded, 3-layer β-sandwich and 9 $PG_{II}$ helices. As for their adhesion protein 2 (ORF33), they start with a long, poorly predicted N-terminal extension, likely involved in contacting other components of the host-binding machinery, followed by a β-helix domain commonly found in phage tail fibers (56), tail spikes (57), and RBP (58) and associated to polysaccharide binding, and end with a single brush domain similar to that of the Araucaria's RBP (Fig. 5 and S1).

***Butters.*** The Butter's adhesion protein 1 (ORF19) is composed of a N-terminal β-stranded knob followed by a β-barrel, and its C-terminal domain is similar to the *B. cenocepacia* BC2L-C lectin domain (Table 1), as it is also the case for the C-terminal lectin domains of Araucaria's and Isca's RBPs (Fig. 5). The presence of this C-terminal lectin domain led us to predict the structure of a trimeric assembly, which shows an arrangement of the three lectin domains similar to that observed in siphophages' adhesion proteins (23, 24, 59, 60). As for the Butters's adhesion protein 2 (ORF22), it is overall similar to that of Abinghost and Badfish (ORF33). It contains a β-strand-rich N-terminal region and a C-terminal brush domain (Fig. 5). However, this brush domain superimposes well on the C-terminal brush domain of Abinghost's and Badfish's adhesion protein 1 (ORF32) (rmsd =1.1 Å on Cα atoms).

***DS6A*** The DS6A's additional adhesion proteins (ORF25 and ORF27) also contain a C-terminal brush domain, similar to that of Araucaria's RBP (rmsd 1.1 Å on Cα atoms). As for their N-terminal regions, while the N-terminal jelly roll of ORF25 is predicted with high confidence, the confidence on the ORF27 N-terminal β-strand-rich predicted structure is weak (Fig. 5 and S1).

***Wildcat.*** Lastly, the predicted structures of the two Wildcat's additional adhesion proteins show features that distinguish them from the other mycobacteriophages' additional adhesion proteins. The adhesion protein 1 (ORF44) is made up of three domains: the N-terminal domain is a three-stranded β-sheet, the

central α/β domain has been identified as a penicillin-binding domain/β-lactamase/DD-peptidase structural homolog using Dali (61) (Table 1), and the two C-terminal domains are jelly-rolls packed one onto the other (Fig. 5). Canonical catalytic residues of β -lactamase/DD-peptidase are not found in the central domain, indicating that this protein does not have catalytic activity on host cell surface components. Interestingly, the second adhesion protein (ORF45) harbours a long, unstructured N-terminal segment followed by three consecutive domains, each being composed of two packed jelly-rolls, as those observed in the first adhesion protein (Fig. 5). Such domains have no structural homologs in the PDB.

**Mycobacteriophages' brush domains present solvent-exposed hydrophobic residues**

All mycobacteriophages under study, but Isca, have proteins in their host-binding machineries containing brush domains. These brush domains are formed of a β-sandwich as the brush handle, from which emerge $PG_{II}$ helices making the brush hairs and accounting for 36% to 47% of glycine residue (Fig. 6). The brush domain in the Araucaria's RBP is similar to those found in the Abinghost's and Badfish's adhesion protein 2 (ORF33), in both DS6A's adhesion proteins (ORF25 and ORF27), and in the Wildcat's RBP (ORF43). The handle region is composed of a 10-stranded, 3-layer β-sandwich, and the hairs are formed of ten $PG_{II}$ helices (Fig. 6). Yet, the loops at the brush hair tip are of variable lengths from one phage to another. Brush domains with a fold different from that described above are present in Butters's RBP (ORF20) and adhesion protein 2 (ORF22), and in the Abinghost's and Badfish's additional adhesion proteins 1 (ORF32). Variations occur in the handle, which are formed of a 2-layer β-sandwich as in the Butters' RBP and in the Abinghost's and Badfish's additional adhesion protein 1 (for the N-terminal brush domain), and/or in the hairs, which assemble a variable number of $PG_{II}$ helices with different packing. In these brush domains, variations are also observed at the brush hair tip including small parallel and anti-parallel β-sheets (Fig. 6). These structural variations could reflect the adaptation of phages to their host cell surface.

A striking feature of these brush domains is the distribution of their bulky hydrophobic residues (Val, Leu, Ile, Met, Phe, Trp). They are found not only buried within the brush handle, as expected, but also exposed at the surface of the brush hair (Fig. 6). They often form clusters of 4 to 11 residues at the brush hair tip (DS6A's ORF27 and ORF25, respectively), which is the region identified as being the host-binding domain of the long tail fibers of phages infecting Gram-negative bacteria and determining host specificity (51, 62, 63). Such solvent-exposed hydrophobic patches are not common in soluble proteins, except in domains interacting with lipids such as those found in phospholipases or lipases and therefore may be involved in host recognition.

**Insights into the architecture of whole mycobacteriophages' host-binding machineries**

In order to gain insights into the assembly of mycobacteriophages' host-binding machinery, we have selected in the actinobacteriophage database (phagesdb.org) a good quality negative staining electron micrograph (nsEM) of Badfish in which the tail distal end exhibits recognizable features. The tail itself is made of stacked hexamers of the MTP and is terminated by a bulky volume that can ascribed to the stacked Dit and Tal oligomers that are known to interact together (42). The thin and long extension following the Dit-Tal core could correspond to a RBP trimer with its C-terminal CBMs accounting for the enlargement visible at the end of the tail tip (Fig. 7A and B). This implies that the RBP trimer attaches to the Tal trimer via its N-terminal α-helices. To test this hypothesis, we have submitted three Tal monomers and three RBP monomers (residues 1 to 330 because of memory limitation) to AF2 to predict the structure of the Tal-RBP interface. Remarkably, the predicted structure of the Tal-RBP complex shows the RBPs' N-terminal α-helices nestled in the Tal trimer internal cavity (Fig. 7B, C and S1). The positively-charged surface of each RBP N-terminal helix interacts with a 1350-$Å^2$ negatively-charged patch of each Tal monomer (Fig. 7C). This predicted structure supports a direct interaction between the Badfish's Tal trimer and a trimer of its potential RBP. In the nsEM image of Badfish, we cannot identify signals accounting for the two potential additional adhesion proteins. Yet, it should be noted that negative staining with uranyl acetate can disrupt the structural integrity of protein complexes. Notably, the RBP of the *S. thermophilus* phage STP1 has not been observed in virions by nsEM for a long time, and hence was thought to be absent in this phage, because of a too high concentration of uranyl acetate (21, 64).

We have also selected in the database a nsEM image of Isca (Fig. 7D), which possesses an RBP but is devoid of additional adhesion proteins (Fig. 1). Our predicted structures of Isca's Dit, Tal and RBP match with the overall architecture of its host-binding machinery observed in virions. The two disks after the last MTP ring can be ascribed to the Dit and Tal oligomers, and hence, the thin extension terminated by a cone-shaped structure at the tail tip can be ascribed to a RBP trimer (Fig. 7D and E). Yet, in this case, the Tal-RBP interface could not be predicted by AF2, likely because of the long, unstructured N-terminal stretch of the Isca's RBP (39 residues).

## Discussion

Deciphering molecular mechanisms used by mycobacteriophages to infect their host is important not only to better understand viral diversity and evolution, phage-host interactions, but also to develop phage-based biotechnological tools. Our predicted structures of mycobacteriophages' host-binding machineries emphasize on the modularity and structural diversity of these macromolecular assemblies that tailed phages have evolved to adapt to the surface of bacteria present in their ecological niche and hence to enhance their ability to initiate contacts with potential hosts. Components of mycobacteriophages' host-binding machineries variably combine structural and functional domains, including CBMs, lectin domains and β-helices that have also been found in tailed phages infecting Gram-negative or Gram-positive bacteria to

interact with cell surface polysaccharides. Surprisingly, polyglycine-rich domains, which we named brush domains based on their peculiar shape, appear to be widespread in mycobacteriophages' host-binding machineries and likely play an important role in host binding. Brush domains have rarely been observed in nature, yet they appear as a versatile scaffold with functions related to protein-protein or protein-ligand interactions (46, 48–51). The unique example of a viral brush domain is that of the *Salmonella* phage S16 tail fiber adhesin that initiates host adsorption by binding to the outer membrane protein OmpC and lipopolysaccharides of the *Salmonella* cell wall (65). In particular, residues located in highly variable loops at the brush hair tip were shown to be involved in host binding (51). In mycobacteriophages, these brush domains present surface-exposed hydrophobic residues, and we therefore propose that they have likely evolved to interact with lipids of the waxy mycobacterial cell wall in which exceptionally long chain fatty acids, the mycolic acids, form an outer mycomembrane. Interestingly, these brush domains may be more common in nature than expected and could be the hallmark of mycobacteriophages. Further AF2 structure predictions of host-binding machineries of different phages will provide valuable information to better understand the function and evolution of these brush domains as well as to unveil unprecedented structural and functional domains.

The mycomembrane is a very unusual feature for bacteria that belong to the *Actinobacteria* phylum of Gram-positive (monoderm) bacteria, which raises questions related to the nature of cell receptors recognized by mycobacteriophages and to their host-binding mechanisms. The mycomembrane inner leaflet, composed of mycolic acids, is linked to arabinogalactan that is in turn covalently attached to peptidoglycan, while its outer leaflet is composed of free phospholipids, glycolipids and lipoglycans (66). The mycomembrane also contains porins and transporters and is covered by a capsule formed of proteins, polysaccharides and small amounts of lipids (66). Therefore, there is a variety of potential receptors in the mycobacterial cell wall for phage binding. Yet, our predicted structures indicate that mycobacteriophages would preferentially use cell surface polysaccharides and lipids rather than proteins to recognize and attach to their hosts. Indeed, CBM, lectin domains, β-helix domains and brush domains, likely involved in carbohydrate and lipid binding, are present in mycobacteriophages' potential RBP and additional adhesion proteins. This composition is in agreement with the fact that, although there are only few papers reporting the identification of mycobacteriophages' receptors, cell wall associated-glycolipids or phospholipids have been identified as cell surface receptors of the mycobacteriophages D4, D29, I3 and Phlei, the lipid or sugar moieties of which being important for host binding depending on the phage (67–69). Moreover, the extended RBPs that certainly plugged to the Tal trimer, like in our Badfish Tal-RBP complex structure prediction, recall the overall architecture and composition of the CBM-rich Tal C-terminal extension of *S. thermophilus* and *O. oeni* siphophages that bind to cell surface polysaccharides (18, 19). In the Araucaria low-resolution structure, the long fiber observed at the tail distal end, similar to that of the coliphages

Lambda and T5, suggested that this mycobacteriophage could bind to proteinaceous receptors (14). The nsEM image of Isca, which only contains RBP like Araucaria, also shows a single fiber attached to Tal (Fig. 7D). However, the predicted structures of Araucaria's and Isca's RBPs reveal that they encompass CBM, lectin domains and brush domains at the tip of the RBP trimeric assemblies, which does not support the possibility of binding to proteins. Mycobacteriophages would therefore be more similar to siphophages infecting Gram-positive bacteria that generally bind to cell wall polysaccharides like the model *L. lactis* phage p2 (70, 71) than siphophages infecting Gram-negative bacteria that irreversibly bind to membrane proteins like the model *E. coli* phages Lambda and T5 (72). The presence of multiple host-binding domains, either within the same protein (the Araucaria's and Isca's RBPs for instance), or in different host-binding proteins (the DS6A's and Butters' RBP and additional adhesion proteins for instance), offer the advantage to achieve strong interactions with the host in spite of low binding affinity of saccharides and lipids through avidity phenomena, and this could be strengthen with the presence of multiple copies of the additional adhesion proteins and their possible multimerization.

Mycobacteriophages of the actinobacteriophage database, but the *M. tuberculosis* phage DS6A, have been isolated on the non-pathogenic *M. smegmatis* $mc^2155$, and the host range of only few of them has been reported (12, 73). There is a correlation between phage clusters and host range, yet mycobacteriophages have the ability to expand or modify their host range through mutations in tail genes (73). However, such mutations in the mycobacteriophages Rosebush (phagesdb.org, cluster B2, ORF32, mutations L297R, W121G, G288D, W293R) and Halo (phagesdb.org, Cluster G, ORF22, mutations A306V, A604E) do not map to CBM, lectin domain or brush domain identified in our predicted structures (Fig. S3). Instead, they are found in positions that likely impact the trimeric assembly of these proteins, thereby changing their ability to interact with mycobacterial cell envelope and modifying host specificity. Interestingly, the Halo mutant that was isolated for its ability to solely infect *M. tuberculosis* did acquire the capacity to bind to *M. tuberculosis* and was still able to strongly bind, but not infect, *M. smegmatis* (73). Generally, host binding is a two-step process that includes the reversible binding of phages that search for specific receptors, followed by irreversible contacts that anchor the whole host-binding machinery to the host in a stable conformation allowing genome injection. In this process, the host-binding proteins and cell receptors mediating the reversible contacts are not necessarily the same as those mediating the irreversible attachment (74), and we hypothesize that the potential additional adhesion proteins, rich in brush domains, would be important for reversible contacts with the host while also strengthening the irreversible contact of the RBP. The observation that the Halo mutant binds both *M. tuberculosis* and *M. smegmatis* but only infects *M. tuberculosis* and our predicted structure of the ORF22 of the Halo host-binding machinery indicate that these mutations would rather impact the dynamics of the whole host-binding machineries leading to the irreversible attachment of phages before DNA ejection than the mere recognition of receptors.

The function of CBMs, lectin domains, and brush domains identified in mycobacteriophages needs to be validated experimentally, and our structure predictions provide a rational basis to produce recombinant proteins or protein domains and perform cell-binding assays. Moreover, structural analyses of whole host-binding machineries in virions alone and bound to the host cell wall will be essential to shed light on their overall architecture and on molecular mechanisms used by mycobacteriophages to recognize and bind to their host cell surface. In particular, deciphering the role, at the molecular level, and the dynamics of CBM insertions in the Dit-Tal core and of the CBM, lectin domains and brush domains in the potential RBP and additional adhesion proteins throughout host binding, will be important to better understand phage-host interactions and also offers great perspectives for host-range engineering (17, 63), which may be instrumental to rationally develop diagnostics tools and phage therapy protocols.

Lastly, our work illustrates the tremendous potential of AF2 to unveil virus structure and biology. AF2 structure predictions can be used to revisit viral genome annotations and to efficiently characterize the overwhelming number of newly discovered viruses, as exemplified by the human gut phageome (75), and the 'viral dark matter' - viral proteins of unknown function – that can represent up to 75% of phage encoded proteins.

## Material and Methods

### Phage selection

Seven mycobacteriophages that belong to different clusters and infect different hosts were selected for analysis in this study. Abinghost (cluster B3), Badfish (cluster B1), Butters (cluster N) and Wildcat (cluster V) infect *M. smegmatis*. Araucaria (not classified) and Isca (cluster A3) infect *M. abscessus*. DS6A is a singleton that infects *M. tuberculosis*. The Genbank accession numbers are as follow: Abinghost (MN444873), Araucaria (AHAS00000000), Badfish (KJ194580), Butters (KC576783), DS6A (JN698994), Isca (MN586063) and Wildcat (DQ398052).

### Protein structure predictions

We used a Github notebook (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb#scrollTo=XUo6foMQxwS2) to perform structure predictions of protein monomers and multimers. Due to memory limitations, the number of residues for monomer and multimer structure predictions had to be less than 1400, and we therefore split long sequences into smaller stretches. First, we ran structure predictions of monomers in order to determine sensible boundaries of stretches that were then submitted for structure prediction as trimers (Tal, RBP) or hexamers (Dit) to AlphaFold multimer (40). Moreover, we predicted structures of stretches with overlapping segments to allow assembly of full-length multimers using *Coot*

(76). The structure prediction of Badfish Tal-RBP complex was performed by submitting three sequences of Tal and three sequences of the potential RBP (residues 1-330). The pLDDT values of predicted structure, stored in the pdb file as B-factors, were plotted using Excel (Fig. S1 and S3).The final predicted protein or domain structures were submitted to the Dali server (77) to identify the closest structural homologs in the PDB. Visual representations of the structures were prepared with ChimeraX (78).

## Author Contributions

Conceptualization: A.G., C.C.; Investigation: C.C.; Resources: C.C.; Formal analysis: A.G. C.C.; Writing original draft: A.G., C.C.; Writing-review and editing: A.G., C.C.; Visualization: A.G., C.C.; Project administration: A.G.

## Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

## Conflict of Interest

C.C. is employee of AlphaGraphix (cambillau.alphagraphix@gmail.com).

## Acknowledgments

We acknowledge UCSF ChimeraX for molecular graphics that is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

## Supplemental Material

Supplemental material for this article is provided.

## Data Availability

All coordinates of predicted structures will be deposited in the open data repository Zenodo.

**Table 1. Dali hits of protein and domain predicted structures.**

| | Dali hits | | | |
|---|---|---|---|---|
| | **PDB ID** | **Z-score*** | **rmsd (Å)#** | **lali$** |
| **Abinghost** | | | | |
| Tal (ORF30) | 3d37 | 18.5 | 4 | 250/316 |
| RBP (ORF31) | | | | |
| CBM16 | 2zew | 15 | 2.6 | 140/147 |
| CBM66 | 4azz | 16.4 | 2.6 | 53/165 |
| | | | | |
| **Araucaria** | | | | |
| Tal (ORF71) | 3d37 | 18.4 | 3.8 | 247/316 |
| RBP (ORF72) | | | | |
| CBM66 | 4azz | 14.3 | 2.3 | 131/165 |
| lectin | 2wq4 | 13 | 2.3 | 117/131 |
| | | | | |
| **Badfish** | | | | |
| Tal (ORF30) | 3d37 | 18 | 4.2 | 248/316 |
| RBP (ORF31) | | | | |
| CBM16 | 2zey | 15.3 | 2.8 | 136/145 |
| CBM66 | 4azz | 16.4 | 2.6 | 53/165 |
| | | | | |
| **Butters** | | | | |
| Dit (ORF17) | | | | |
| CBM66 | 4azz | 15.4 | 2.8 | 54/165 |
| RBP (ORF20) | | | | |
| CBM16 | 2zew | 14.3 | 2.6 | 129/144 |
| Additional adhesion protein 1 (ORF 19) | | | | |
| lectin | 2wq4 | 8.9 | 2.5 | 109/134 |
| | | | | |
| **DS6A** | | | | |
| RBP (ORF23) | | | | |
| CBM66 | 4b1l | 10 | 3.3 | 126/162 |
| lectin | 4l92 | 6.3 | 3.5 | 121/391 |
| | | | | |
| **Isca** | | | | |
| Dit (ORF27) | | | | |
| domain of T4 gp11 | 1el6 | 5.4 | 3.2 | 83/208 |
| RBP (ORF31) | | | | |
| CBM66 | 4azz | 15.6 | 2.6 | 154/165 |
| lectin | 2wq4/4l92 | 9/8.9 | 2.7/2.8 | 114/134 ; 124/391 |
| | | | | |
| **Wildcat** | | | | |
| Dit (ORF39) | | | | |
| CBM66 | 4azz | 15.9 | 2.5 | 51/165 |
| Tal (ORF40) | 1mz5 | 15.8 | 2.5 | 165/622 |
| RBP (ORF43) | | | | |
| CBM16 | 2zew | 14.3 | 2.6 | 129/144 |
| CBM6 | 2cdo | 6.3 | 2.4 | 94/138 |
| Additional adhesion protein 1 (ORF 44) | | | | |
| penicillin-binding domain | 1ei5 | 29.6 | 3.5 | 312/518 |

* : the Z-score is a measure of structural similarity

#: root mean square deviation on Cα atoms

$: lali indicates the number of aligned residues in the predicted structure over the total number of residues in the PDB hit.

## Figures

**A**

| Phage | Abinghost | Araucaria | Badfish | Butters | DS6A | Isca | Wildcat |
|---|---|---|---|---|---|---|---|
| **Host** | *M. smegmatis* | *M. abscessus* | *M. smegmatis* | *M. smegmatis* | *M. tuberculosis* | *M. abscessus* | *M. smegmatis* |
| **Cluster** | Cluster B3 | n.d. | Cluster B1 | Cluster N | singleton | Cluster A3 | Cluster V |
| Dit | ORF 29 (476*) | ORF 70 (478*) | ORF 29 (477*) | ORF 17 (573*) | ORF 20 (323*) | ORF 26 (457*) | ORF 39 (531*) |
| Tal | ORF 30 (347*) | ORF 71 (345*) | ORF 30 (370*) | ORF 18 (569*) | ORF 21 (579*) | ORF 27 (599*) | ORF 40 (794*) |
| RBP | ORF 31 (741*) | ORF 72 (785*#) | ORF 31 (751*) | ORF 20 (738*#) | ORF 23 (702*) | ORF 31 (619*) | ORF 43 (855*#) |
| Additional adhesion protein 1 | ORF 32 (460*#) | | ORF 32 (448*#) | ORF 19 (280*) | ORF 25 (286*#) | | ORF 44 (658*) |
| Additional adhesion protein 2 | ORF 33 (400*#) | | ORF 33 (386*#) | ORF 22 (346*#) | ORF 27 (622*#) | | ORF 45 (707*) |
| Lysin A,B | ORFs 46-47 | n.d. | ORFs 49-50 | ORFs 26-27 | ORFs 30-31 | ORFs 9-10 | ORF 49, ORF 52 |

*: number of residues ; #: contains a brush domain

**B**

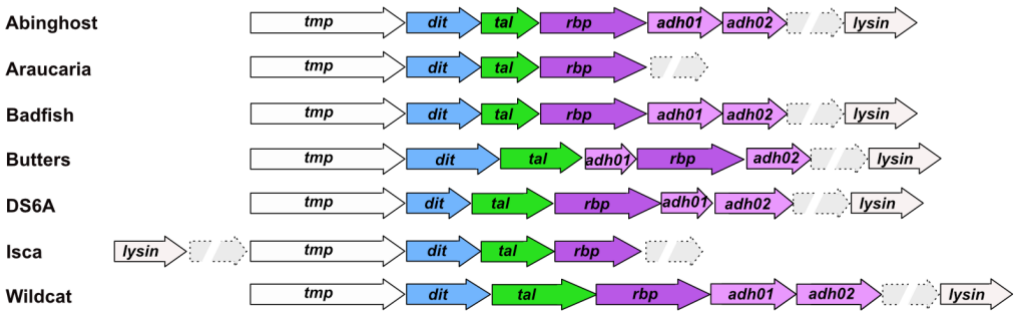



**Fig. 1. Diversity of mycobacteriophages analysed in this study. A.** Genomic properties (clusters), host spectrum, and host-binding machinery composition of the seven analysed mycobacteriophages. **B.** Schematics of the genomic region coding for proteins assembling host-binding machineries.

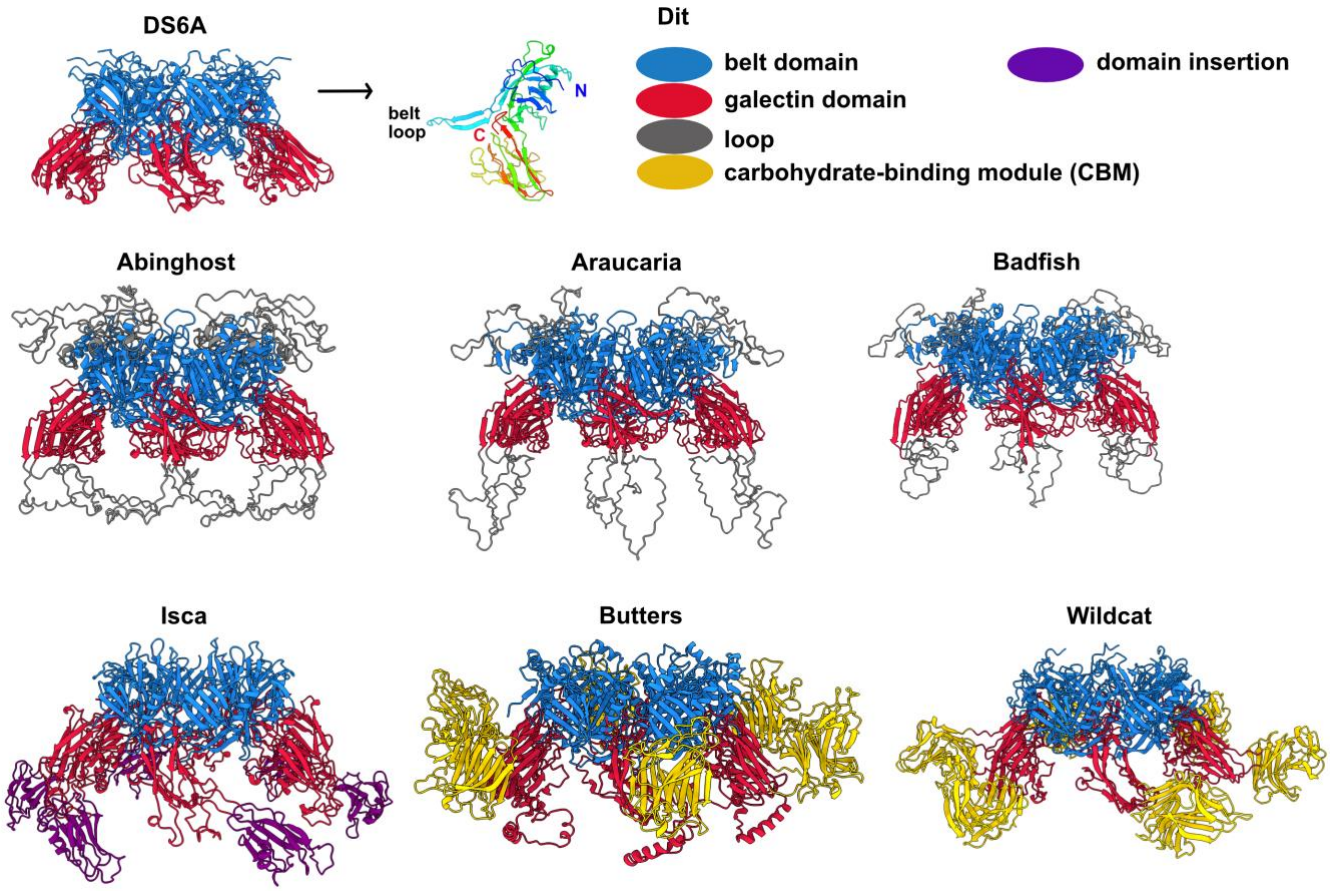



**Fig. 2. Dit predicted structures.** Ribbon representations of Dit predicted structures. One rainbow-colored DS6A Dit monomer is shown for clarity. While the Dit of Abinghost, Araucaria, Badfish and DS6A contain only the belt and galectin domains, the Dit of Butters, Isca and Wildcat have an extra domain inserted in the galectin domain. This extra-domain has been identified as a CBM for the Dit of Butters and Wildcat. The color code is indicated.

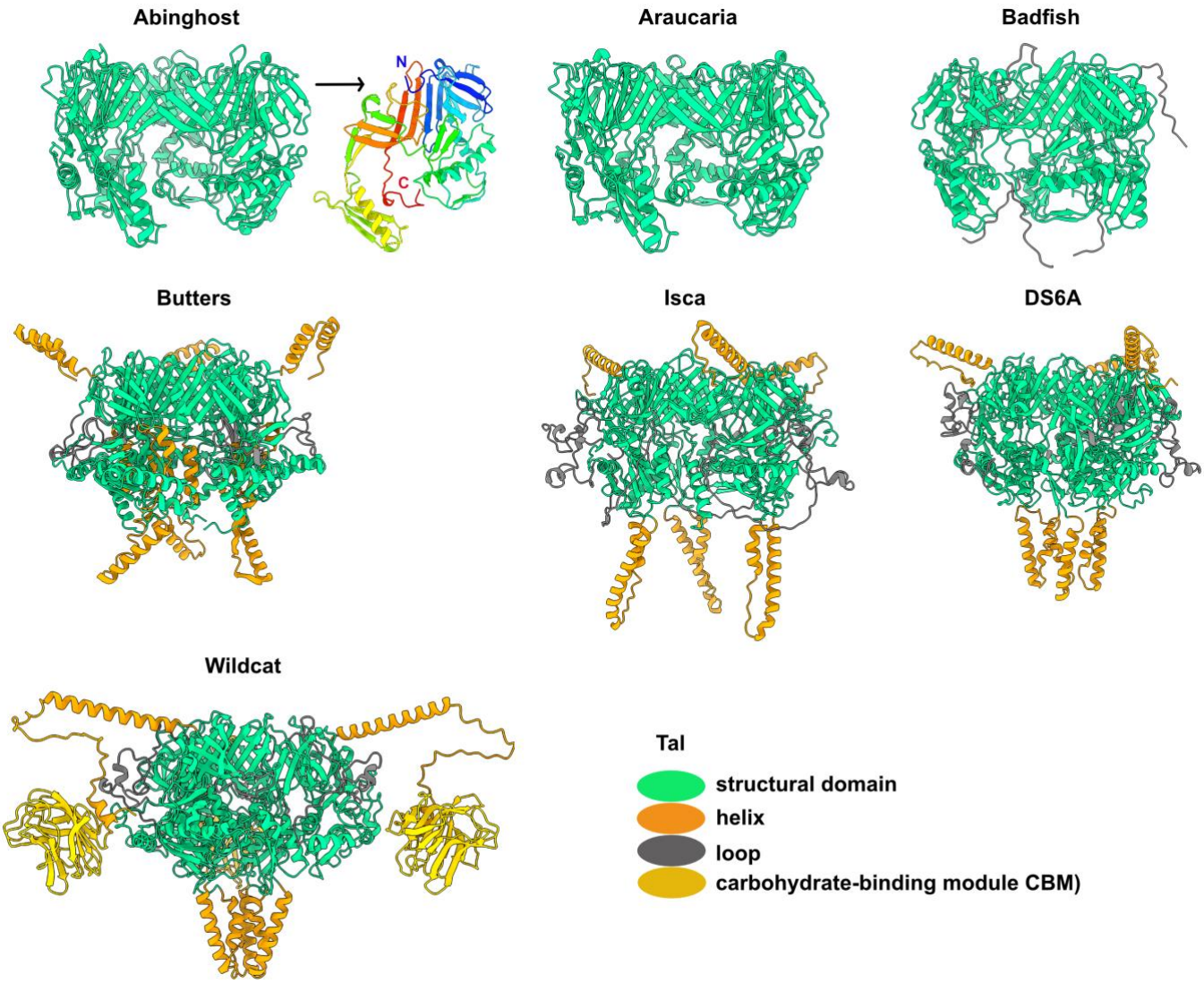
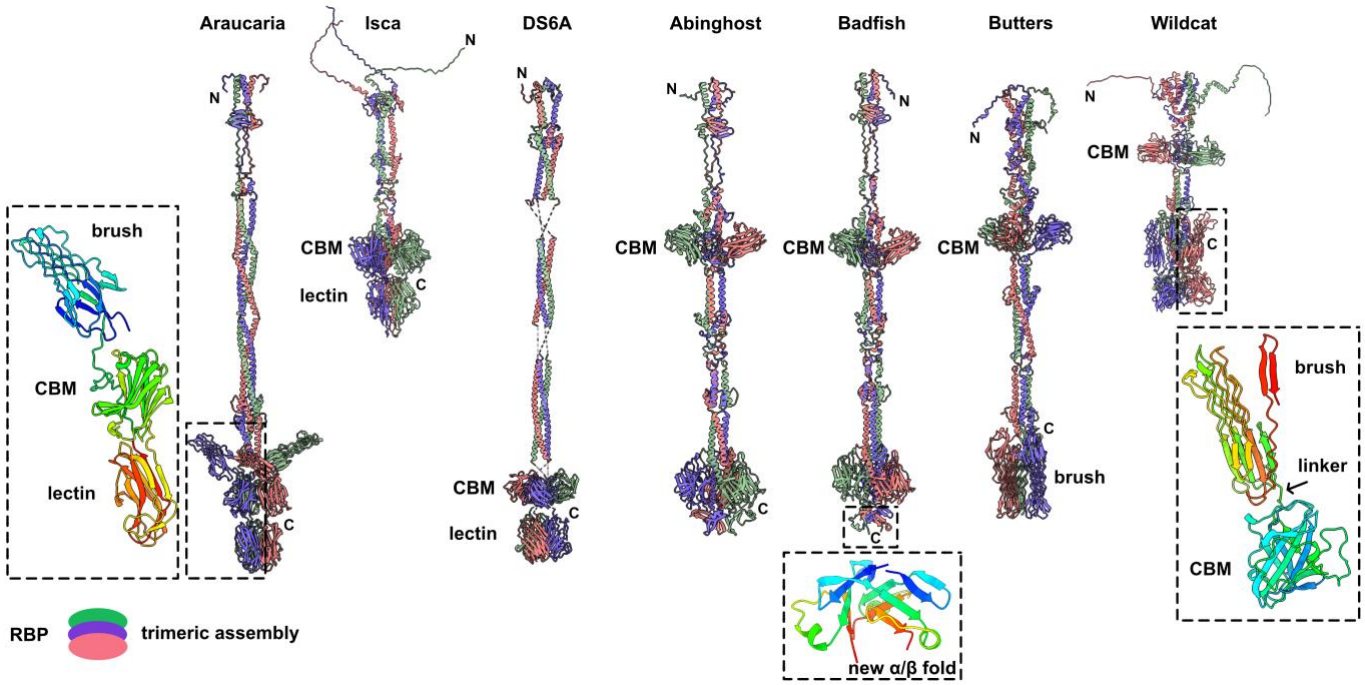



**Fig. 3. Tal predicted structures.** Ribbon representations of the Tal predicted structures. One rainbow-colored Abinghost Tal monomer is shown for clarity. The color code is indicated.




471

472 **Fig. 4. Predicted structures of potential RBPs reveal multi-domain assemblies.** Ribbon representations
473 of the RBP predicted structures (the N- and C-termini are indicated.). Trimeric assemblies are colored by
474 chain. The inset on the left-hand side highlights the brush domain, CBM and lectin domain at the
475 Araucaria's RBP C-terminal end (rainbow colored from the N-terminus to the C-terminus). The inset below
476 the Badfish's RBP highlights the trimerc knob (rainbow colored) at its C-terminal end. The inset on the
477 right-hand side highlights the brush domain and CBM, connected by a flexible linker, at the Wildcat's RBP
478 C-terminal end (rainbow colored).

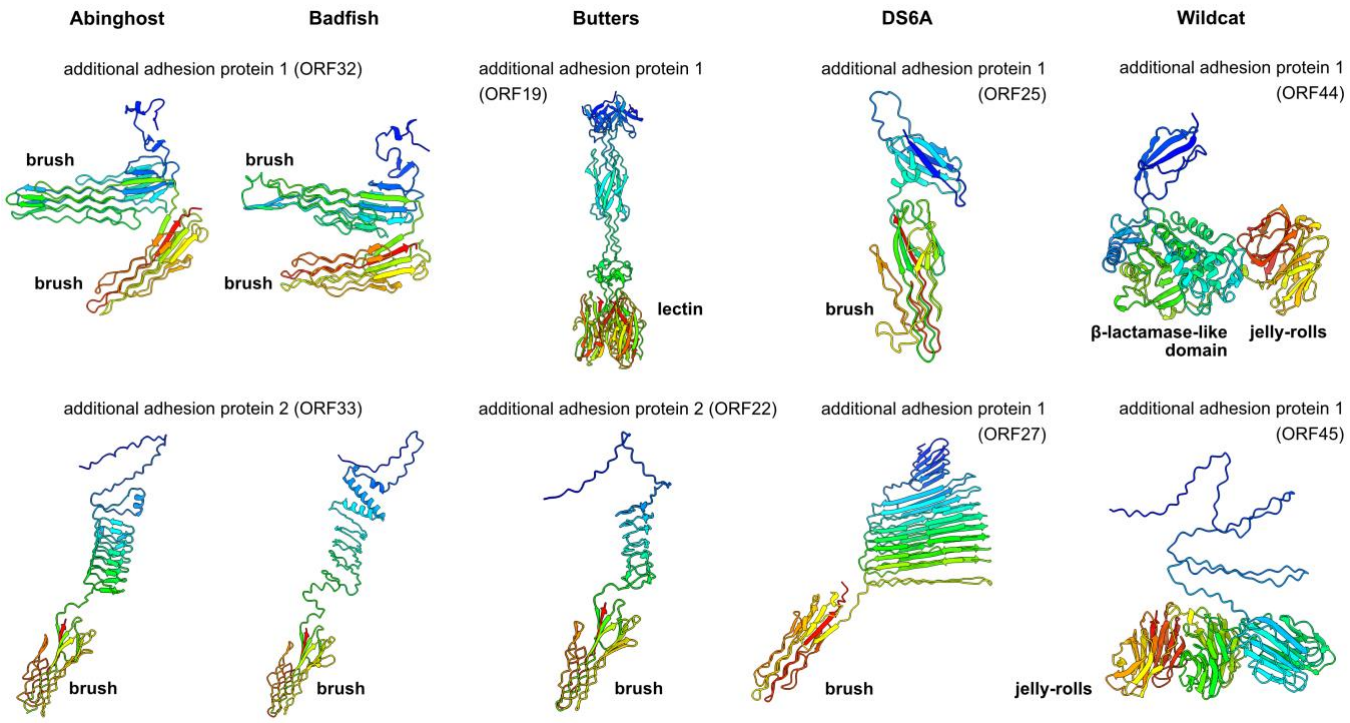



**Fig. 5. Additional adhesion protein predicted structures.** Ribbon representations of the additional adhesion protein predicted structures (rainbow colored from the N-terminus to the C-terminus).

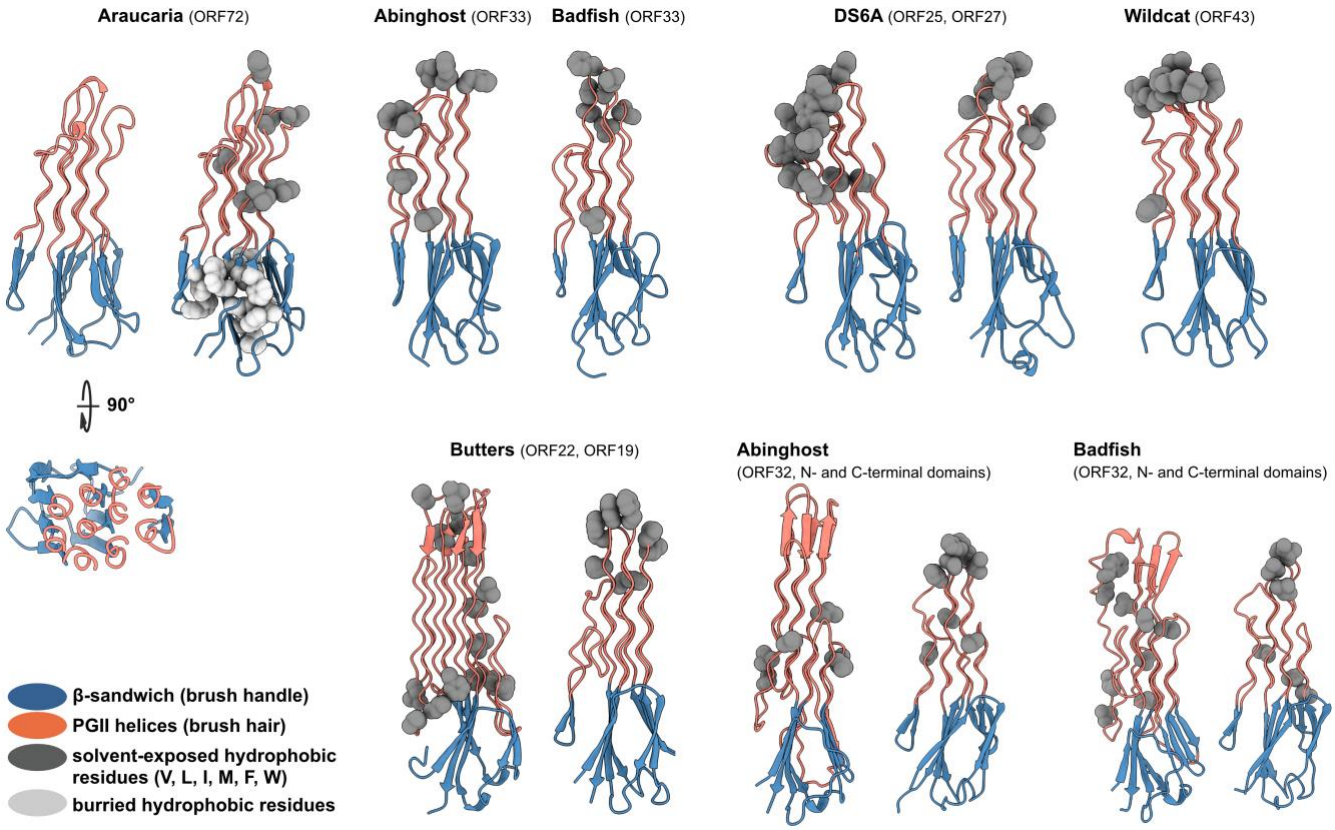



**Fig. 6. Mycobacteriophages' brush domains present solvent-exposed hydrophobic residues.** Ribbon representations of the brush domains identified in the mycobacteriophages' RBPs and additional adhesion proteins. Solvent-exposed and buried hydrophobic residues are shown as spheres. The color code is indicated.

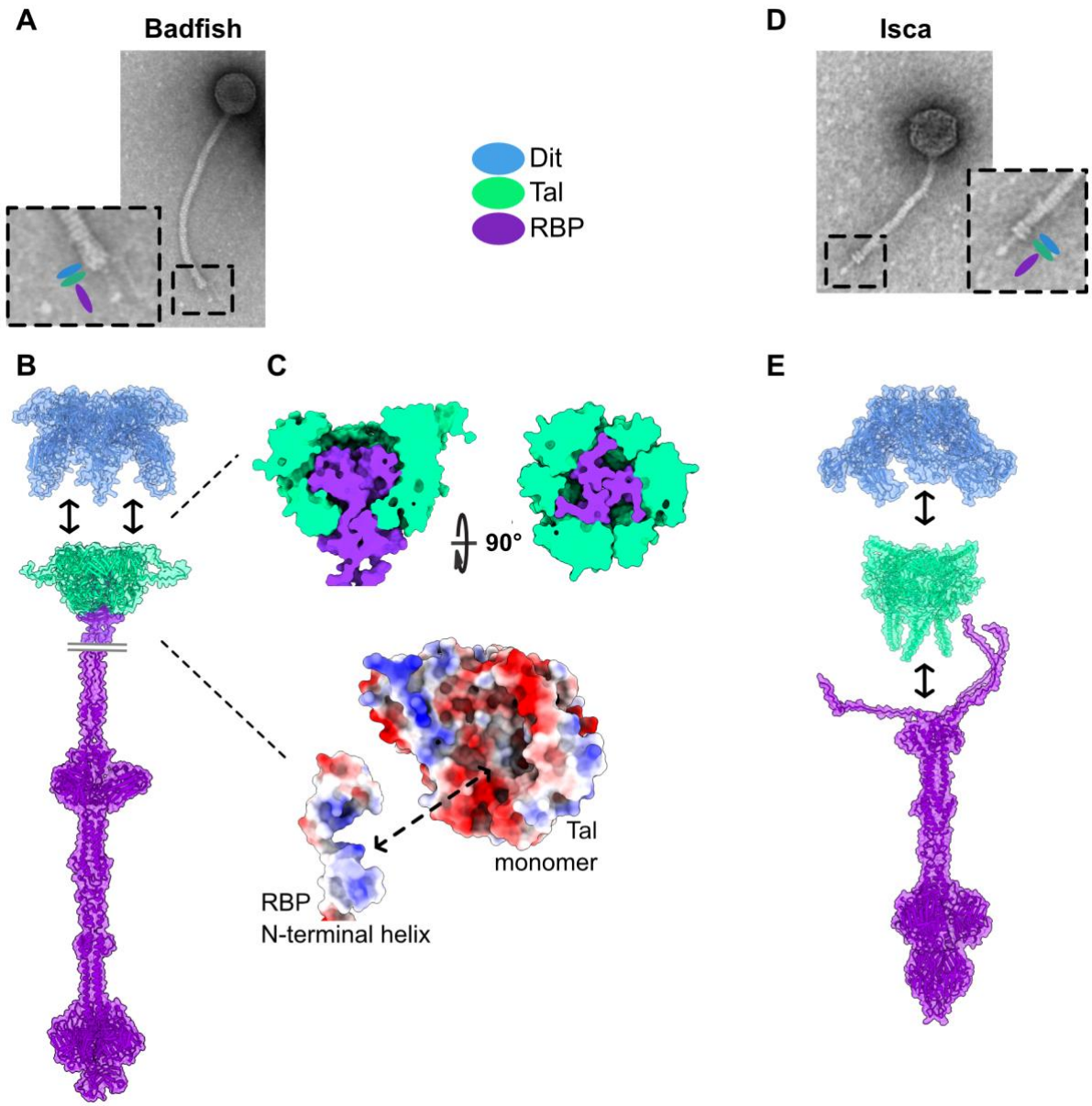



**Fig. 7. Architectures of mycobacteriophages' host-binding machineries. A.** nsEM picture of Badfish (from phagesdb.org) with its host-binding machinery at the tail distal end highlighted in a dotted box. In the inset, the colored circles indicate the relative position of Dit, Tal and RBP (the color code is indicated). **B.** Surface and ribbon representation of the predicted structures of the Badfish's Dit ring, Tal-RBP$_{(residues\ 1\text{-}330)}$ complex, and full-length RBP trimer. These predicted structures are compatible with the overall architecture of its host-binding machinery. **C.** Top. Orthogonal, slabbed views of the Tal-RBP complex showing three RBP N-terminal helices plugged into the Tal trimer internal cavity. Bottom. A positively-charged surface of one RBP N-terminal helix interacts with a negatively-charged surface of a Tal monomer. **D.** nsEM picture of Isca (from phagesdb.org) with its host-binding machinery at the tail distal end highlighted in a dotted box. In the inset, the colored circles indicate the relative position of Dit, Tal and RBP (the color code is indicated). **E.** Surface and ribbon representation of the predicted structures of the Isca's Dit, Tal, and RBP. These predicted structures are compatible with the overall architecture of its host-binding machinery.

## References


1. Prasanna AN, Mehra S. 2013. Comparative phylogenomics of pathogenic and non-pathogenic mycobacterium. PLoS One 8:e71248.
2. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Meier-Kolthoff JP, Klenk H-P, Clément C, Ouhdouch Y, van Wezel GP. 2016. Taxonomy, Physiology, and Natural Products of Actinobacteria. Microbiol Mol Biol Rev 80:1–43.
3. Broncano-Lavado A, Senhaji-Kacha A, Santamaría-Corral G, Esteban J, García-Quintanilla M. 2022. Alternatives to Antibiotics against Mycobacterium abscessus. Antibiotics (Basel) 11:1322.
4. Daffé M, Marrakchi H. 2019. Unraveling the Structure of the Mycobacterial Envelope. Microbiol Spectr 7.
5. Batt SM, Minnikin DE, Besra GS. 2020. The thick waxy coat of mycobacteria, a protective layer against antibiotics and the host's immune system. Biochemical Journal 477:1983–2006.
6. Hanauer DI, Graham MJ, SEA-PHAGES, Betancur L, Bobrownicki A, Cresawn SG, Garlena RA, Jacobs-Sera D, Kaufmann N, Pope WH, Russell DA, Jacobs WR, Sivanathan V, Asai DJ, Hatfull GF. 2017. An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. Proc Natl Acad Sci U S A 114:13531–13536.
7. Russell DA, Hatfull GF. 2017. PhagesDB: the actinobacteriophage database. Bioinformatics 33:784–786.
8. Hatfull GF. 2020. Actinobacteriophages: Genomics, Dynamics, and Applications. Annu Rev Virol 7:37–61.
9. Hatfull GF, Sarkis GJ. 1993. DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. Mol Microbiol 7:395–405.
10. Hatfull GF. 2018. Mycobacteriophages. Microbiol Spectr 6.
11. Sarkis GJ, Jacobs WR, Hatfull GF. 1995. L5 luciferase reporter mycobacteriophages: a sensitive tool for the detection and assay of live mycobacteria. Mol Microbiol 15:1055–1067.
12. Dedrick RM, Guerrero-Bustamante CA, Garlena RA, Russell DA, Ford K, Harris K, Gilmour KC, Soothill J, Jacobs-Sera D, Schooley RT, Hatfull GF, Spencer H. 2019. Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nat Med 25:730–733.
13. Little JS, Dedrick RM, Freeman KG, Cristinziano M, Smith BE, Benson CA, Jhaveri TA, Baden LR, Solomon DA, Hatfull GF. 2022. Bacteriophage treatment of disseminated cutaneous Mycobacterium chelonae infection. Nat Commun 13:2313.
14. Sassi M, Bebeacua C, Drancourt M, Cambillau C. 2013. The First Structure of a Mycobacteriophage, the Mycobacterium abscessus subsp. bolletii Phage Araucaria. Journal of Virology 87:8099–8109.
15. Bielmann R, Habann M, Eugster MR, Lurz R, Calendar R, Klumpp J, Loessner MJ. 2015. Receptor binding proteins of Listeria monocytogenes bacteriophages A118 and P35 recognize serovar-specific teichoic acids. Virology 477:110–118.
16. Dieterle M-E, Spinelli S, Sadovskaya I, Piuri M, Cambillau C. 2017. Evolved distal tail carbohydrate binding modules of Lactobacillus phage J-1: a novel type of anti-receptor widespread among lactic acid bacteria phages. Mol Microbiol 104:608–620.
17. Dunne M, Rupf B, Tala M, Qabrati X, Ernst P, Shen Y, Sumrall E, Heeb L, Plückthun A, Loessner MJ, Kilcher S. 2019. Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Reports 29:1336-1350.e4.
18. Goulet A, Joos R, Lavelle K, Van Sinderen D, Mahony J, Cambillau C. 2022. A structural discovery journey of streptococcal phages adhesion devices by AlphaFold2. Front Mol Biosci 9:960325.
19. Goulet A, Cambillau C. 2021. Structure and Topology Prediction of Phage Adhesion Devices Using AlphaFold2: The Case of Two Oenococcus oeni Phages. 10. Microorganisms, 2021/10/24 ed. 9.
20. Kizziah JL, Manning KA, Dearborn AD, Dokland T. 2020. Structure of the host cell recognition and penetration machinery of a Staphylococcus aureus bacteriophage. PLoS Pathog 16:e1008314.
21. Lavelle K, Goulet A, McDonnell B, Spinelli S, van Sinderen D, Mahony J, Cambillau C. 2020. Revisiting the host adhesion determinants of Streptococcus thermophilus siphophages. Microb Biotechnol 13:1765–1779.
22. Legrand P, Collins B, Blangy S, Murphy J, Spinelli S, Gutierrez C, Richet N, Kellenberger C,

554 Desmyter A, Mahony J, van Sinderen D, Cambillau C. 2016. The Atomic Structure of the Phage Tuc2009
555 Baseplate Tripod Suggests that Host Recognition Involves Two Different Carbohydrate Binding
556 Modules. mBio 7:e01781-15, /mbio/7/1/e01781-15.atom.
557 23. Sciara G, Bebeacua C, Bron P, Tremblay D, Ortiz-Lombardia M, Lichière J, van Heel M,
558 Campanacci V, Moineau S, Cambillau C. 2010. Structure of lactococcal phage p2 baseplate and its
559 mechanism of activation. Proc Natl Acad Sci USA 107:6852–6857.
560 24. Veesler D, Spinelli S, Mahony J, Lichiere J, Blangy S, Bricogne G, Legrand P, Ortiz-Lombardia
561 M, Campanacci V, van Sinderen D, Cambillau C. 2012. Structure of the phage TP901-1 1.8 MDa
562 baseplate suggests an alternative host adhesion mechanism. Proceedings of the National Academy of
563 Sciences of the United States of America, 2012/05/23 ed. 109:8954–8.
564 25. Pell LG, Kanelis V, Donaldson LW, Lynne Howell P, Davidson AR. 2009. The phage λ major tail
565 protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion
566 system. Proc Natl Acad Sci USA 106:4160–4165.
567 26. Cardarelli L, Pell LG, Neudecker P, Pirani N, Liu A, Baker LA, Rubinstein JL, Maxwell KL,
568 Davidson AR. 2010. Phages have adapted the same protein fold to fulfill multiple functions in virion
569 assembly. Proc Natl Acad Sci USA 107:14384–14389.
570 27. Veesler D, Robin G, Lichière J, Auzat I, Tavares P, Bron P, Campanacci V, Cambillau C. 2010.
571 Crystal structure of bacteriophage SPP1 distal tail protein (gp19.1): a baseplate hub paradigm in gram-
572 positive infecting phages. J Biol Chem 285:36666–36673.
573 28. Bárdy P, Füzik T, Hrebík D, Pantůček R, Thomas Beatty J, Plevka P. 2020. Structure and
574 mechanism of DNA delivery of a gene transfer agent. Nat Commun 11:3034.
575 29. Kanamaru S, Leiman PG, Kostyuchenko VA, Chipman PR, Mesyanzhinov VV, Arisaka F,
576 Rossmann MG. 2002. Structure of the cell-puncturing device of bacteriophage T4. Nature 415:553–557.
577 30. Goulet A and C C. 2022. Present impact of AlphaFold2 revolution on structural biology, and an
578 illustration with the structure prediction of the bacteriophage J-1 host adhesion device. Frontiers in
579 Molecular Biosciences in press.
580 31. Stockdale SR, Mahony J, Courtin P, Chapot-Chartier M-P, van Pijkeren J-P, Britton RA, Neve H,
581 Heller KJ, Aideh B, Vogensen FK, van Sinderen D. 2013. The lactococcal phages Tuc2009 and TP901-1
582 incorporate two alternate forms of their tail fiber into their virions for infection specialization. J Biol
583 Chem 288:5581–5590.
584 32. Linares R, Arnaud C-A, Degroux S, Schoehn G, Breyton C. 2020. Structure, function and
585 assembly of the long, flexible tail of siphophages. Curr Opin Virol 45:34–42.
586 33. Sao-Jose C, Lhuillier S, Lurz R, Melki R, Lepault J, Santos MA, Tavares P. 2006. The
587 ectodomain of the viral receptor YueB forms a fiber that triggers ejection of bacteriophage SPP1 DNA. J
588 Biol Chem, 2006/02/17 ed. 281:11464–70.
589 34. Hayes S, Vincentelli R, Mahony J, Nauta A, Ramond L, Lugli GA, Ventura M, van Sinderen D,
590 Cambillau C. 2018. Functional carbohydrate binding modules identified in evolved dits from siphophages
591 infecting various Gram-positive bacteria: Functional carbohydrate binding modules identified in evolved
592 dits. Mol Microbiol 110:777–795.
593 35. Hayes S, Mahony J, Vincentelli R, Ramond L, Nauta A, van Sinderen D, Cambillau C. 2019.
594 Ubiquitous Carbohydrate Binding Modules Decorate 936 Lactococcal Siphophage Virions. Viruses 11.
595 36. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas
596 AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred
597 Server at its Core. J Mol Biol 430:2237–2243.
598 37. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure
599 and function prediction. Nat Methods 12:7–8.
600 38. Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL,
601 Laskowski RA, Pozzati G, Shenoy A, Zhu W, Kundrotas P, Serra VR, Rodrigues CHM, Dunham AS,
602 Burke D, Borkakoti N, Velankar S, Frost A, Basquin J, Lindorff-Larsen K, Bateman A, Kajava AV,
603 Valencia A, Ovchinnikov S, Durairaj J, Ascher DB, Thornton JM, Davey NE, Stein A, Elofsson A, Croll
604 TI, Beltrao P. 2022. A structural biology community assessment of AlphaFold2 applications. Nat Struct
605 Mol Biol 29:1056–1067.
606 39. Gao M, Nakajima An D, Parks JM, Skolnick J. 2022. AF2Complex predicts direct physical
607 interactions in multimeric proteins with deep learning. Nat Commun 13:1744.

608 40. Evans R et al. 2021. Protein complex prediction with AlphaFold-Multimer. BioRxiv
609 https://doi.org/10.1101/2021.10.04.463034;
610 41. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates
611 R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B,
612 Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M,
613 Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P,
614 Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589.
615 42. Goulet A, Spinelli S, Mahony J, Cambillau C. 2020. Conserved and Diverse Traits of Adhesion
616 Devices from Siphoviridae Recognizing Proteinaceous or Saccharidic Receptors. Viruses, 2020/05/10 ed.
617 12.
618 43. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-
619 active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490-495.
620 44. Buschiazzo A, Tavares GA, Campetella O, Spinelli S, Cremona ML, París G, Amaya MF, Frasch
621 AC, Alzari PM. 2000. Structural basis of sialyltransferase activity in trypanosomal sialidases. EMBO J
622 19:16–24.
623 45. Dunne M, Hupfeld M, Klumpp J, Loessner M. 2018. Molecular Basis of Bacterial Host
624 Interactions by Gram-Positive Targeting Bacteriophages. Viruses 10:397.
625 46. Buglino J, Shen V, Hakimian P, Lima CD. 2002. Structural and biochemical analysis of the Obg
626 GTP binding protein. Structure 10:1581–1592.
627 47. Mus F, Eilers BJ, Alleman AB, Kabasakal BV, Wells JN, Murray JW, Nocek BP, DuBois JL,
628 Peters JW. 2017. Structural Basis for the Mechanism of ATP-Dependent Acetone Carboxylation. Sci Rep
629 7:7234.
630 48. Pentelute BL, Gates ZP, Tereshko V, Dashnau JL, Vanderkooi JM, Kossiakoff AA, Kent SBH.
631 2008. X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic
632 protein enantiomers. J Am Chem Soc 130:9695–9701.
633 49. Warkentin E, Weidenweber S, Schühle K, Demmer U, Heider J, Ermler U. 2017. A rare
634 polyglycine type II-like helix motif in naturally occurring proteins. Proteins 85:2017–2023.
635 50. Weidenweber S, Schühle K, Demmer U, Warkentin E, Ermler U, Heider J. 2017. Structure of the
636 acetophenone carboxylase core complex: prototype of a new class of ATP-dependent
637 carboxylases/hydrolases. Sci Rep 7:39674.
638 51. Dunne M, Denyes JM, Arndt H, Loessner MJ, Leiman PG, Klumpp J. 2018. Salmonella Phage
639 S16 Tail Fiber Adhesin Features a Rare Polyglycine Rich Domain for Host Recognition. Structure
640 26:1573-1582.e4.
641 52. Cuskin F, Flint JE, Gloster TM, Morland C, Baslé A, Henrissat B, Coutinho PM, Strazzulli A,
642 Solovyova AS, Davies GJ, Gilbert HJ. 2012. How nature can exploit nonspecific catalytic and
643 carbohydrate binding modules to create enzymatic specificity. Proc Natl Acad Sci U S A 109:20889–
644 20894.
645 53. Sulák O, Cioci G, Delia M, Lahmann M, Varrot A, Imberty A, Wimmerová M. 2010. A TNF-like
646 trimeric lectin domain from Burkholderia cenocepacia with specificity for fucosylated human histo-blood
647 group antigens. Structure 18:59–72.
648 54. Squeglia F, Maciejewska B, Łątka A, Ruggiero A, Briers Y, Drulis-Kawa Z, Berisio R. 2020.
649 Structural and Functional Studies of a Klebsiella Phage Capsule Depolymerase Tailspike: Mechanistic
650 Insights into Capsular Degradation. Structure 28:613-624.e4.
651 55. Henshaw J, Horne-Bitschy A, van Bueren AL, Money VA, Bolam DN, Czjzek M, Ekborg NA,
652 Weiner RM, Hutcheson SW, Davies GJ, Boraston AB, Gilbert HJ. 2006. Family 6 carbohydrate binding
653 modules in beta-agarases display exquisite selectivity for the non-reducing termini of agarose chains. J
654 Biol Chem 281:17099–17107.
655 56. Garcia-Doval C, Castón JR, Luque D, Granell M, Otero JM, Llamas-Saiz AL, Renouard M,
656 Boulanger P, van Raaij MJ. 2015. Structure of the Receptor-Binding Carboxy-Terminal Domain of the
657 Bacteriophage T5 L-Shaped Tail Fibre with and without Its Intra-Molecular Chaperone. Viruses 7:6424–
658 6440.
659 57. Steinbacher S, Seckler R, Miller S, Steipe B, Huber R, Reinemer P. 1994. Crystal Structure of P22
660 Tailspike Protein: Interdigitated Subunits in a Thermostable Trimer. Science 265:383–386.
661 58. Hawkins NC, Kizziah JL, Hatoum-Aslan A, Dokland T. 2022. Structure and host specificity of

Staphylococcus epidermidis bacteriophage Andhra. Sci Adv 8:eade0459.
59. Dunne M, Rupf B, Tala M, Qabrati X, Ernst P, Shen Y, Sumrall E, Heeb L, Pluckthun A, Loessner MJ, Kilcher S. 2019. Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. 5. Cell Rep, 2019/10/31 ed. 29:1336-1350 e4.
60. Farenc C, Spinelli S, Vinogradov E, Tremblay D, Blangy S, Sadovskaya I, Moineau S, Cambillau C. 2014. Molecular insights on the recognition of a Lactococcus lactis cell wall pellicle by the phage 1358 receptor binding protein. J Virol 88:7005–7015.
61. Bompard-Gilles C, Remaut H, Villeret V, Prangé T, Fanuel L, Delmarcelle M, Joris B, Frère J, Van Beeumen J. 2000. Crystal structure of a D-aminopeptidase from Ochrobactrum anthropi, a new member of the "penicillin-recognizing enzyme" family. Structure 8:971–980.
62. Drexler K, Riede I, Montag D, Eschbach M-L, Henning U. 1989. Receptor specificity of the Escherichia coli T-even type phage Ox2. Journal of Molecular Biology 207:797–803.
63. Trojet SN, Caumont-Sarcos A, Perrody E, Comeau AM, Krisch HM. 2011. The gp38 adhesins of the T4 superfamily: a complex modular determinant of the phage's host specificity. Genome Biol Evol 3:674–686.
64. Lavelle K, Murphy J, Fitzgerald B, Lugli GA, Zomer A, Neve H, Ventura M, Franz CM, Cambillau C, van Sinderen D, Mahony J. 2018. A Decade of Streptococcus thermophilus Phage Evolution in an Irish Dairy Plant. Appl Environ Microbiol 84:e02855-17.
65. Marti R, Zurfluh K, Hagens S, Pianezzi J, Klumpp J, Loessner MJ. 2013. Long tail fibres of the novel broad-host-range T-even bacteriophage S16 specifically recognize *Salmonella* OmpC: T4-like *Salmonella* phage S16. Molecular Microbiology 87:818–834.
66. Chiaradia L, Lefebvre C, Parra J, Marcoux J, Burlet-Schiltz O, Etienne G, Tropis M, Daffé M. 2017. Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane. Sci Rep 7:12807.
67. Chen J, Kriakov J, Singh A, Jacobs WR, Besra GS, Bhatt A. 2009. Defects in glycopeptidolipid biosynthesis confer phage I3 resistance in Mycobacterium smegmatis. Microbiology 155:4050–4057.
68. Furuchi A, Tokunaga T. 1972. Nature of the receptor substance of Mycobacterium smegmatis for D4 bacteriophage adsorption. J Bacteriol 111:404–411.
69. Imaeda T, Blas FS. 1969. Adsorption of Mycobacteriophage on Cell-wall Components. Journal of General Virology 5:493–498.
70. Bebeacua C, Tremblay D, Farenc C, Chapot-Chartier M-P, Sadovskaya I, van Heel M, Veesler D, Moineau S, Cambillau C. 2013. Structure, adsorption to host, and infection mechanism of virulent lactococcal phage p2. J Virol 87:12302–12312.
71. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. mBio 5:e00880-00814.
72. Goulet A, Spinelli S, Mahony J, Cambillau C. 2020. Conserved and Diverse Traits of Adhesion Devices from Siphoviridae Recognizing Proteinaceous or Saccharidic Receptors. Viruses 12.
73. Jacobs-Sera D, Marinelli LJ, Bowman C, Broussard GW, Guerrero Bustamante C, Boyle MM, Petrova ZO, Dedrick RM, Pope WH, Science Education Alliance Phage Hunters Advancing Genomics And Evolutionary Science Sea-Phages Program, Modlin RL, Hendrix RW, Hatfull GF. 2012. On the nature of mycobacteriophage diversity and host preference. Virology 434:187–201.
74. Bertozzi Silva J, Storms Z, Sauvageau D. 2016. Host receptors for bacteriophage adsorption. FEMS Microbiology Letters 363:fnw002.
75. Shkoporov AN, Hill C. 2019. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. Cell Host Microbe 25:195–209.
76. Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66:486–501.
77. Holm L. 2020. Using Dali for Protein Structure Comparison. Methods Mol Biol 2112:29–42.
78. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30:70–82.