**Structural annotation of the conserved carbohydrate esterase vb\_24B\_21 from Shiga toxin-encoding bacteriophage Φ24B**

Barbara Franke1, Marta Veses-Garcia2, Kay Diederichs1, Heather Allison2,

Daniel J. Rigden2\*, Olga Mayans1\*

*1Department of Biology, University of Konstanz, 78457 Konstanz, Germany;* 2*Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK.*

\* To whom correspondence should be addressed:

Prof. Olga Mayans,

Tel: +49 (0)7531 882212, O[lga.Mayans@uni-konstanz.de](mailto:olga.mayans@uni-konstanz.de)

Prof. Daniel J. Rigden,

Tel: +44 (0)151 7954467, Drigden@liverpool.ac.uk

**Abstract**

Shiga toxin-encoding bacteriophagestransfer Shiga toxin genes to *Escherichia coli* and are responsible for the emergence of pathogenic bacterial strains that cause severe foodborne human diseases.Gene *vb\_24B\_21* is the most highly conserved gene across sequenced Shiga bacteriophages. Protein vb\_24B\_21(also termed 933Wp42 and NanS-p) is a carbohydrate esterase with homology to the *E. coli* chromosomally encoded NanS that deacetylates sialic acid in the intestinal mucus. To assist the functional characterization of vb\_24B\_21, we have studied its molecular structure by homology modelling its esterase domain and by elucidating the crystal structure of its uncharacterized C-terminal domain at the atomic resolution of 0.97 Å. Our modelling confirms that NanS from the *E. coli* host is the closest structurally characterized homolog to the esterase domain of vb\_24B\_21. Like NanS, vb\_24B\_21 has an atypical active site,comprising a simple catalytic dyad Ser-His and a divergent oxyanion hole. The crystal structure of the C-terminal domain reveals a lectin-like, jelly-roll -sandwich fold. The domain displays a prominent cleft that bioinformatics analysis predicts to be a carbohydrate binding site without catalytic properties. Taken together, the data indicate that vb\_24B\_21 contains a NanS-like esterase domain followed by a carbohydrate-binding module of yet undetermined binding specificity.

**Keywords:** carbohydrate deacetylase, jelly-roll domain, carbohydrate binding module, protein X-ray crystallography, molecular bioinformatics

**Introduction**

Shiga toxins (Stx) are the main virulence factors of a group of *Escherichia coli* strains that cause severe foodborne human diseases, such as haemorrhagic colitis and haemolytic-uraemic syndrome. Shiga toxin-producing *E. coli* (STEC) strains acquire the Stx genes through infection by Shiga toxin-encoding, lambdoid bacteriophages (Krüger and Lucchesi, 2015). The first outbreak of an enterohaemorrhagic *E. coli* occurred in 1982. Since then, it has become established that Stx-encoding phages are responsible for driving the dissemination of Stx genes and the emergence and virulence of STEC strains (Allison, 2007).

The genomes of Stx viruses have been found to be quite heterogeneous, including the sequences of the toxin genes they carry (Allison, 2007; Smith et al., 2007; Smith et al., 2012). The most highly conserved gene across a selection of sequenced Stx phages (Φ24Β, GenBank: HM208303.1; VT2-Sa, NCBI: NC\_000902.1; Min27, NCBI: NC\_010237.1; phage 1717, NCBI: NC\_011357.1; phage 86, NCBI: NC\_008464.1; BP-4795, NCBI: NC\_004813.1) has been identified using SEED (Overbeek et al., 2005). The gene, annotated in the genome of Φ24Β as vb\_24B\_21, is located immediately downstream from the *stx* operon encoding the Shiga toxins and immediately upstream of the genes mediating bacterial lysis: *S*; *R*; *Rz* and *Rz1*. The latter encode a set of proteins that perforate the bacterial membrane and break down the peptidoglycan cell wall, enabling phage release from the bacterial host cell. The expression of vb\_24B\_21 is linked to the expression of those lysis genes (Veses-Garcia et al., 2015).

Protein vb\_24B\_21 in Φ24Β is named 933Wp42 in phage 933W (Nübling et al., 2014) and NanS-p in its *E. coli* prophage form (Saile et al., 2016). The protein is 645 amino acids long and it has been shown *in vitro* to deacetylate triacetin, 4-methylumbelliferyl-acetate (4-MUF-Ac), 5-N-acetyl-9-O-acetyl neuraminic acid (Neu5,9Ac2; the most abundant neuraminic acid derivative in humans), mucin (an intestinal, filamentous glycoprotein rich in neuraminic acid derivatives) and various synthetic mono-, di-, and tri-O-acetylated derivatives of Neu5Ac and N-glycolylneuraminic acid (Nübling et al., 2014; Saile et al., 2016; Feuerbaum et al., 2018). These data proved that vb\_24B\_21 can process both free and glycosidically bound sialic acid. The acquisition of this viral gene may add to the endogenous sialic acid esterase activity of *E. coli*, the protein NanS, which deacetylates Neu5,9Ac2, and may be involved in the hydrolysis of intestinal mucin and the uptake of neuraminic acid by the bacteria that uses it as a carbon source (Steenbergen et al., 2009).

Sequence-based predictions reveal that the deacetylase activity of vb\_24B\_21 maps to a carbohydrate esterase domain of the SASA family (Pfam 03629) spanning residues 72-395 (Rangel et al., 2016; Saile et al., 2016). This domain is a homolog of the endogenous NanS esterase of *E. coli*. In vb\_24B\_21 (but not NanS), the esterase domain is flanked by N- and C-terminal sequences of unknown function. The N-terminal sequence, DUF1737, is variable across phage sialic acid esterases, but the C-terminal sequence is conserved, which suggests that the latter holds an important functional role in these enzymes. To assist the characterization of this conserved phage protein, we have studied the three-dimensional structure of its shared esterase and C-terminal domains using homology modelling, X-ray crystallography and structural bioinformatics.

**Methods**

**Cloning**

Gene *vb\_24B\_21* was amplified from the genome of phage vB\_EcoP\_24B using primers containing a *NcoI* (Forward: ccatggcatttaaacactatga) and *SalI (*Reverse: cagctgcggtacgaaatggatattc) restriction sites and cloned into the pETM-11 vector (EMBL) that adds a His6-tag and a Tobacco Etch Virus (TEV) cleavage site N-terminally to the inserted gene.

**Protein production**

The full-length protein vb\_24B\_21 (UniprotKb: G3CFL3; residues 1-645) was expressed in tagged formed (as described above) in *E. coli* SoluBL21 cells (Genlantis) in Luria Bertani medium supplemented with 50 µg/ml kanamycin (Sigma). Cultures were grown at 37 ºC to an OD600 of 0.5. Protein expression was induced adding 0.1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) and growth continued for further 3 h. Cells were harvested by centrifugation. The pellet was resuspended in lysis buffer (50 mM NaH2PO4, 300 mM NaCl) supplemented with 10 mM imidazole and 2 mM phenylmethylsulfonyl fluoride and lysed by sonication. After centrifugation the supernatant was applied to a Ni2+-NTA column (Qiagen) equilibrated in lysis buffer. The column was washed with lysis buffer with increasing imidazole concentrations (20-50 mM) and the protein eluted with lysis buffer supplemented with 250 mM imidazole. Next, the sample was buffer exchanged into 50 mM NaH2PO4, 50 mM NaCl using a PD-10 desalting column (GE Healthcare) and further purified by gel filtration on a Superdex S75 16/60 column (GE Healthcare). Protein concentration was determined by A280. Samples were stored at 4ºC until further use.

**Crystallization**

Crystals were grown at 21°C in 2 µL drops obtained by mixing 1:1 protein stock at a concentration of 32 mg/mL and mother liquor. In growth condition (A), crystals were obtained in 48-well VDX plates (Hampton Research) in hanging drops from solutions containing 0.9 M LiSO4, 10% [v/v] glycerol, 0.1 M Hepes pH 7.0. In condition (B), crystals grew in 96-well MRC crystallization plates (Molecular Dimensions) in sitting drops from 200 mM NaCl, 193 mM NaH2PO4, 400 mM K2HPO4, 100 mM Imidazole pH 7. For X-ray data collection, crystals grown in both (A) and (B) settings were vitrified by flash freezing in liquid N2 under dry mounting conditions.

**Crystal structure elucidation**

X-ray diffraction data were collected at the Diamond Light Source (Didcot, UK) and processed using XDS (Kabsch, 2010). Native data were collected from a single crystal grown in condition (A), but independently in two different beamlines. The two sets (extending to resolutions of 0.97 Å and 4.5 Å, respectively) were merged together in XSCALE (Kabsch, 2010). This approach was applied to increase the completeness of the lower resolution shells in the atomic resolution set. The resulting merged data were used for structural determination. Due to the low sequence identity of the C-terminal domain of vb\_24B\_21 to known structures and the uncertainty of its 3D-fold, phasing attempts first used Wide Search Molecular Replacement (WSMR) (Stokes-Rees and Sliz, 2010) but were unsuccessful. Phases were then obtained by SAD phasing using a single crystal obtained under growth condition (B) and exploiting the anomalous signal from three native sulfur atoms in the protein (2x Met; 1x Cys). A substructure solution (CCall=21.0%, CCweak=17.1%) was obtained in SHELXD (Sheldrick, 2010) using anomalous data to 2.1 Å resolution. A mainchain trace consisting of 196 residues (out of 218 in the protein fragment) was then built in SHELXE (Sheldrick, 2010) and yielded a correlation of 44.90% against native data. Further model building and refinement against native data (A) used PHENIX (Adams et al., 2011) and COOT (Emsley et al., 2011). Model statistics were calculated in MOLPROBITY (Williams et al., 2018). X-ray data and model refinement statistics are given in **Table 1**.

**Homology modelling**

The sequence of vb\_24B\_21 was used to search the Protein Data Bank ([www.rcsb.org](http://www.pdb.org) uses the MMseqs2 algorithm; [Steinegger and Soeding, 2017)](https://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.3988.html). Matches were obtained for the esterase domain (residues 73-392), but not for the DUF1737 (residues 1-72) or C-terminal (residues 393-645) sequences. For the esterase domain, the closest homolog was NanS (PDB code 3PT5). Similarity extended to loop regions that are of similar length. A homology model of the vb\_24B\_21 esterase domain was then calculated using MODELLER (Webb and Sali., 2016) with NanS as template. Model parameters were validated with MOLPROBITY (Williams et al, 2018).

**Results**

**Vb\_24B\_21 is an atypical carbohydrate esterase**

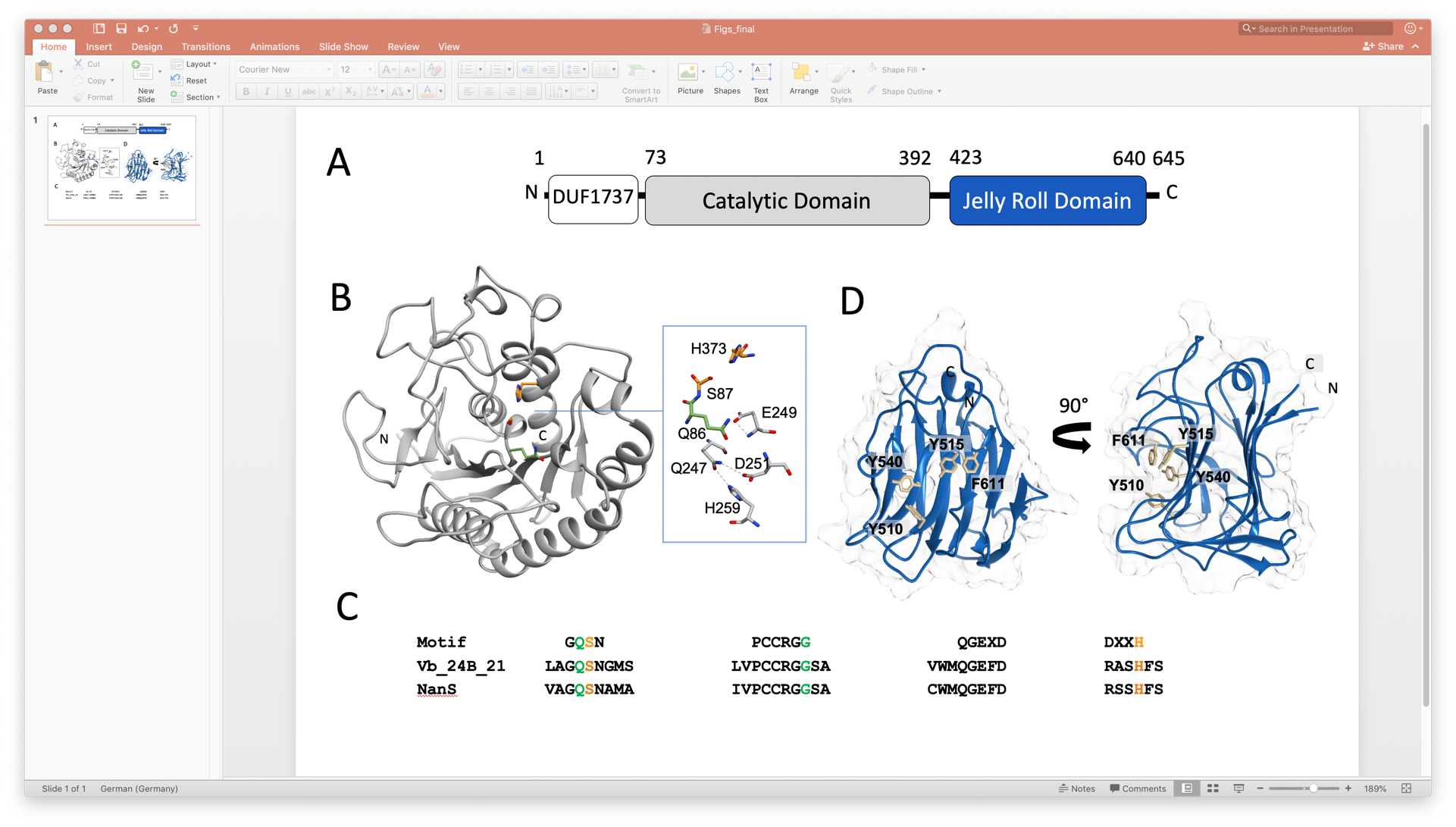
Vb\_24B\_21A is a multi-domain protein consisting of the N-terminal DUF1737 sequence, an esterase domain and a C-terminal uncharacterized region (**Fig 1A**). To assist characterizing this protein, we aimed to identify homologs of known 3D structure by performing a search of the Protein Data Bank ([www.rcsb.org](http://www.pdb.org)[)](https://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.3988.html). This revealed homologous structures only for the esterase domain, with the closest homolog being NanS (57% seq. id., 68% conservation; **Fig S1**). The high conservation allowed us to calculate a homology model for the esterase domain of vb\_24B\_21, which allowed us to confirm the compatibility of the vb\_24B\_21 sequence with the structural features of NanS (**Fig 1B**).

Carbohydrate esterase (CE) enzymes are classified in 16 defined families in the Carbohydrate-Active Enzyme (CAZy) database (<http://www.cazy.org>). NanS currently lies in a non-classified section of the database. CE families encompass a variety of 3D-architectures, but several adopt the SGNH fold (Nakamura et al., 2017). SGNH CE enzymes owe their name to their catalytic Ser and His residues and to Gly and Asn residues that form the oxyanion hole (Molgaard et al, 2000). Notably, NanS shares the overall architecture of SGNH CEs, but does not contain the typical catalytic site (Rangarajan et al, 2011). Based on the features of NanS, Rangarajan et al. distinguished between a canonical group I and an atypical group II of esterases. Group I contains four typical motifs (Lo et al, 2003; Nakamura et al, 2017): (1) a GD**S** motif that hosts the catalytic Ser acting as nucleophile as well as proton donor to the oxyanion hole; (2) a conserved Gly that contributes to the oxyanion hole; (3) a GX**N** in which the asparagine also forms the oxyanion hole; and (4) **D**XX**H** containing the two remaining catalytic residues of the triad, aspartic acid and histidine. Aspartate is not always present, resulting then in a simple Ser-His catalytic dyad. In group II, defined by NanS, (1) is a larger G**QS**N motif, carrying the catalytic serine and an adjacent glutamine that contributes to the oxyanion hole. The latter functionally replaces the asparagine in the canonical GXN motif 3. In turn, the latter becomes a longer motif QGEXD, now involved in stabilizing the orientation of the glutamine in the GQSN motif 1. In NanS, the Asp in the DXXH motif is not present and only the histidine residue of the motif remains, therefore leading to a Ser-His catalytic dyad. Both the catalytic dyad Ser-His and the atypical composition of the oxyanion hole in NanS are strictly conserved in vb\_24B\_21 (**Fig 1B,C**). Thus, we conclude that vb\_24B\_21 can be classed as an atypical deacetyl esterase of type II.

**The C-terminal domain of vb\_24B\_21 adopts a jelly-roll fold**

Next, we aimed to elucidate the 3D-structure of vb\_24B\_21 using X-ray crystallography. Crystallization trials were set with the full-length protein. However, structure elucidation revealed that degradation of the sample had taken place and that crystals only contained the C-terminal domain of vb\_24B\_21 (residues 423-640). The atomic structure of this domain was then elucidated to 0.97 Å resolution (**Table 1**).

The structure revealed that the domain adopts the extended concanavalin-A-like jelly-roll fold typical of lectins (Loris et al, 1998; Brinda et al, 2005) (**Fig 1D**). Specifically, the jelly-roll domain (JRD) of vb\_24B\_21 folds into 12 anti-parallel -strands arranged in two connected greek key motifs that pack against each other to form a -sandwich. One of the -sheets is composed of seven short -strands connected by prominent, long loops. The loops cluster on the surface of the fold forming a groove on the concave side of the domain. The opposite side of the -sandwich is formed by a flatter -sheet of five long -strands joined by short connectors.



**Fig 1: Domain structure of vb\_24B\_21 esterase**

**A.** Domain composition of vb\_24B\_21 esterase. The boundaries of the catalytic domain are predicted from sequence conservation with NanS. The boundaries of the JRD correspond to the crystal structure; **B.** Homology model of the vb\_24B\_21 esterase domain. Inset shows vb\_24B\_21 residues that in NanS form the catalytic dyad (orange) and the atypical oxyanion hole (green); **C.** Functional motifs are strictly conserved in vb\_24B\_21 and NanS; **D.** Crystal structure of vb\_24B\_21 JRD. A pronounced surface groove is observed on the concave side of the -sandwich. Aromatic residues located within the groove are displayed.

**Table 1: X-ray diffraction data and model refinement statistics**

**Data set Native1 S-SAD**

Crystal growth condition A B

Space Group P212121 P212121

Cell dimensions

a, b, c, (Å) 54.45, 54.64, 72.25 54.47, 54.97, 72.74

*X-ray data collection*

X-ray Source Diamond I04 / I04-11 Diamond I02

Detector ADSC Q315 / Pilatus 2M1 Pilatus 6M-F

Wavelength (Å) 0.7514 / 0.91731 1.600

Resolution (Å) 25.0 - 0.97 / 25.0 - 4.51 40-1.7

(Outer Resolution; Å) (0.98-0.97) (1.75-1.70)

Unique reflections 128671 (3878) 45102 (3115)

Multiplicity 7.1 (4.7) 25.8 (17.2)

Completeness (%) 99.5 (99.6) 97.2 (80.5)

Rsym(I) (%) 7.9 (155.2) 6.1 (13.3)

I/(I) 14.64 (1.00) 45.6 (16.4)

CC1/2 (%) 99.8 (40.0) 100.0 (99.5)

*Model Refinement*

No. of Reflections: working / test set 122213 / 6441

Rwork / Rfree (%) 15.01 / 17.19

A.u. content

Protein residues 225

Solvent atoms 431

Ligands 1 x glycerol

Bond r.m.s.d.

Lengths (Å) / Angles (o)2 0.007 / 0.983

Ramachandran plot (%)2

Favoured / Outliers 97.22 / 0

1Native data were obtained by merging two sets collected independently from a same crystal. The statistics reported are those corresponding to the merged data, as used for structure determination.

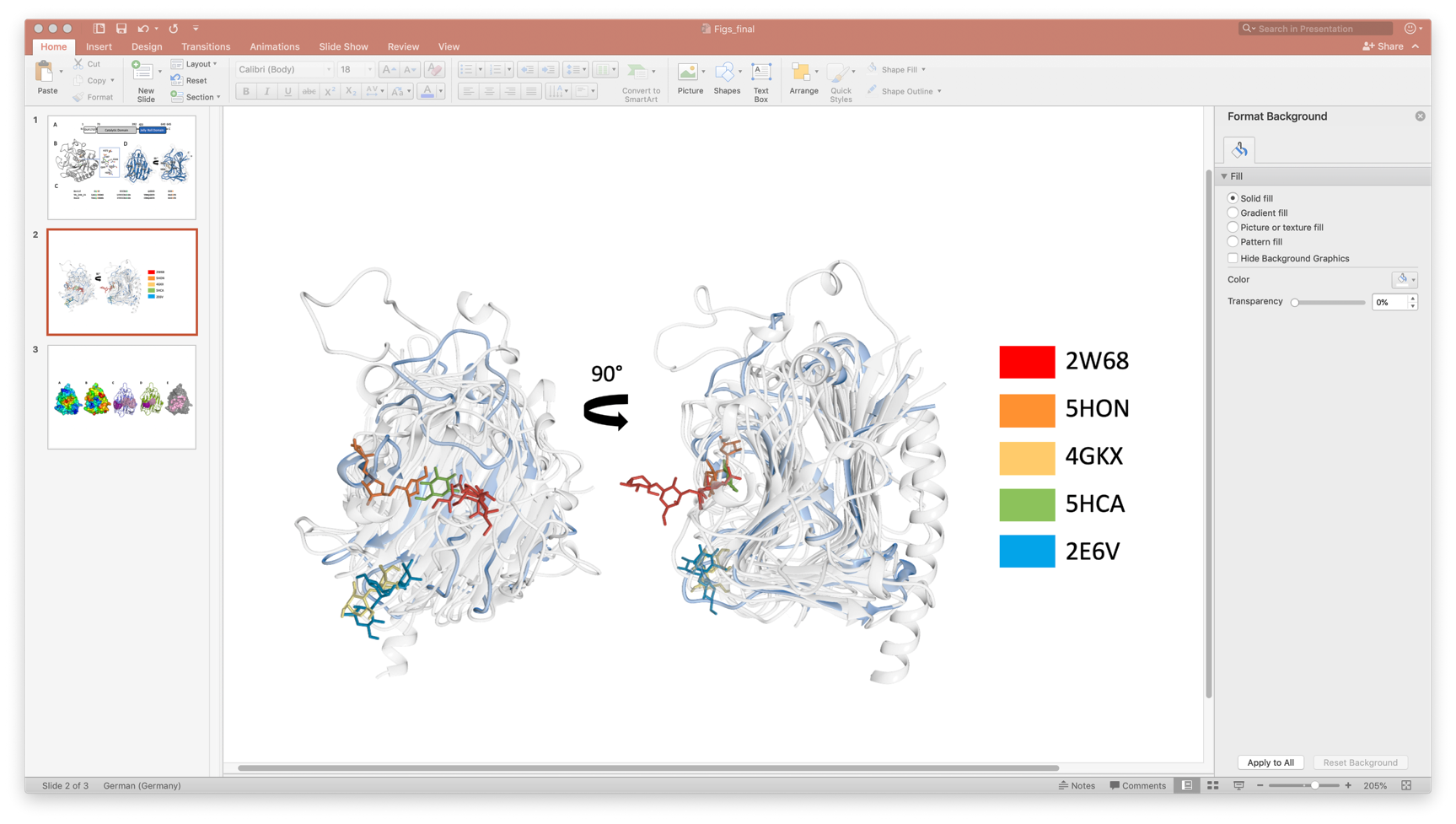
2 Calculated using MOLPROBITY (Williams et al, 2018).

**Structural neighbours of vb\_24B\_21 JRD are carbohydrate binding domains**

The jelly-roll fold of vb\_24B\_21’s C-terminal domain places it within the large superfamily b.29.1, ‘Concanavalin A-like lectins/glucanases’ in the SCOP hierarchical database of protein domain structures (<http://scop.mrc-lmb.cam.ac.uk)> and clan CL0004 ‘Concanavalin’ in the Pfam sequence database of families and domains ([https://pfam.xfam.org](https://pfam.xfam.org/)). Concanavalin, the representative of these protein groups, is a carbohydrate-binding lectin protein. Carbohydrate-binding is the predominant function among proteins encompassed by the categories in both databases. Domains with a concanavalin-like fold can have either catalytic function (as those in the glycoside hydrolase family 7) or non-catalytic carbohydrate binding roles. A minority of related families possess divergent activities, such as peptidase A4 or the calcium-binding C-terminal domains of thrombospondin. Thus, we aimed to reveal whether the JRD from vb\_24B\_21 is capable of performing catalysis by searching databases of catalytic site structural templates, including those in the Catalytic Site Atlas (Furnham et al, 2014), using the GASS (Moraes et al, 2017), ASSAM (Nadzirin et al, 2012), and ProFunc (Laskowski et al, 2005) servers. Catalytic residues were not be identified by any of the searches, suggesting that this might primarily be a carbohydrate binding module (CBM), as are most members of the lectin family.

We further explored the function of vb\_24B\_21A JRD by identifying structural neighbours through comparing its 3D-structure with that of existing protein structures in the Protein Data Bank using DALI (Holm, 2019). This revealed a set of closest structural relatives sharing sequence identities of 16-10%. In order to extract a functional hypothesis for vb\_24B\_21A JRD, we focused here on nearest neighbours with bound ligands: the N-terminal lectin domain of *Vibrio cholerae* sialidase (PDB code 2W68; DALI Z-score 18.4), a mammalian cargo receptor for the export of glycoproteins from the endoplasmic reticulum (4GKX; Z-score 13.5), *Entamoeba histolytica* calreticulin (5HCA; Z-score 13.4), a canine transporter lectin for the secretion of mannose-rich glycoproteins (2E6V; Z-score 13.2); and *Geobacillus stearothermophilus* arabinanase(5HON; Z-score 13.0) (**Fig 2**). In all cases, the ligand was a carbohydrate and the binding occurred on the concave side of the jelly-roll fold. At this location, vb\_24B\_21 JRD presents its pronounced groove (**Fig 1D**), making it likely to also correspond to a carbohydrate binding site in this protein. However, vb\_24B\_21 JRD does not share sequence similarity with its structural neighbours and carbohydrate binding residues are not conserved in vb\_24B\_21. Furthermore, the topography of the carbohydrate binding site is different across these proteins. This lack of conservation did not allow us to propose potential carbohydrate binding residues in vb\_24B\_21.

Of the ligand-bound structures above (Fig 2), only 2W68 is currently classified in CAZy, being in the carbohydrate binding module category CBM40. To better assess the similarity of vb\_24B\_21 JRD with catalogued CBMs, we compared the closest structural neighbours of vb\_24B\_21 JRD identified by DALI (in a 90% non-redundant version of the PDB) against 3D-structures of CBMs in CAZy. This revealed that vb\_24B\_21 JRD does not share close similarity to other CBM families adopting a jelly-roll fold, with the CBM40 domain from 1KIT being the most structurally similar (Z-score 17.6; 16% sequence identity). The next family is CBM32 represented by 5XNR (Z-score 10.7; 8% seq. id.) followed by 4AZZ in CBM66 (10.6; 7%). Structures in CBM16 (2ZEW; 10.2; 9%), CBM61 (2XOM; 9.5; 17%), CBM4 (3P6B; 9.2; 15%), CBM22 (4XUP; 9.1; 12%), CBM42 (6SXT, 9.0, 7%), CBM22 (1H6X, 8.8; 11%), CBM70 (4D0Q, 8.7; 18%), CBM30 (1WMX, 8.3; 9%) and CBM11 (6R3M, 8.2; 14%) are presumably also distant homologues of vb\_24B\_21 JRD.



**Figure 2. Comparison of vb\_24B\_21 JRD and its structural neighbours**

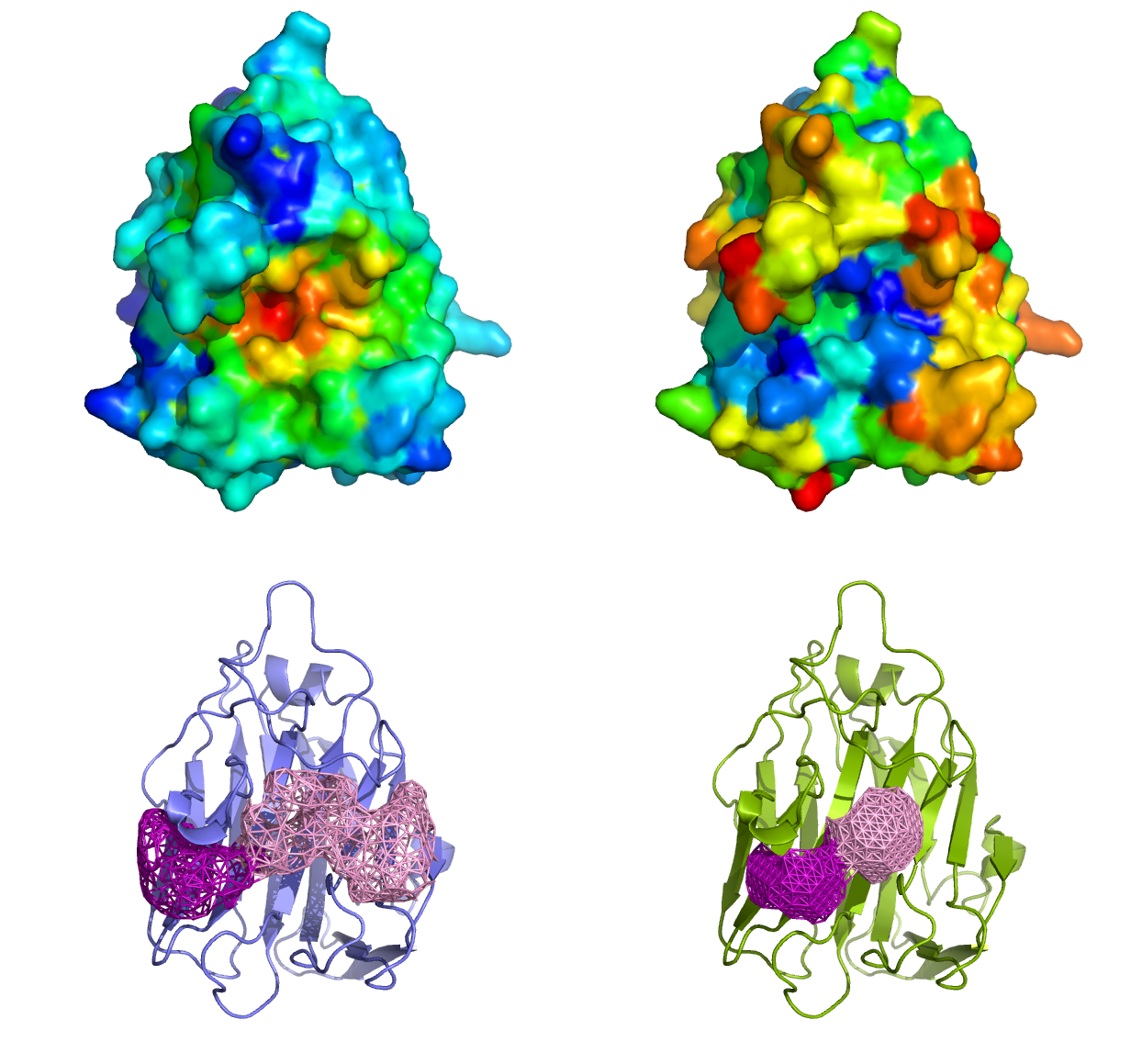
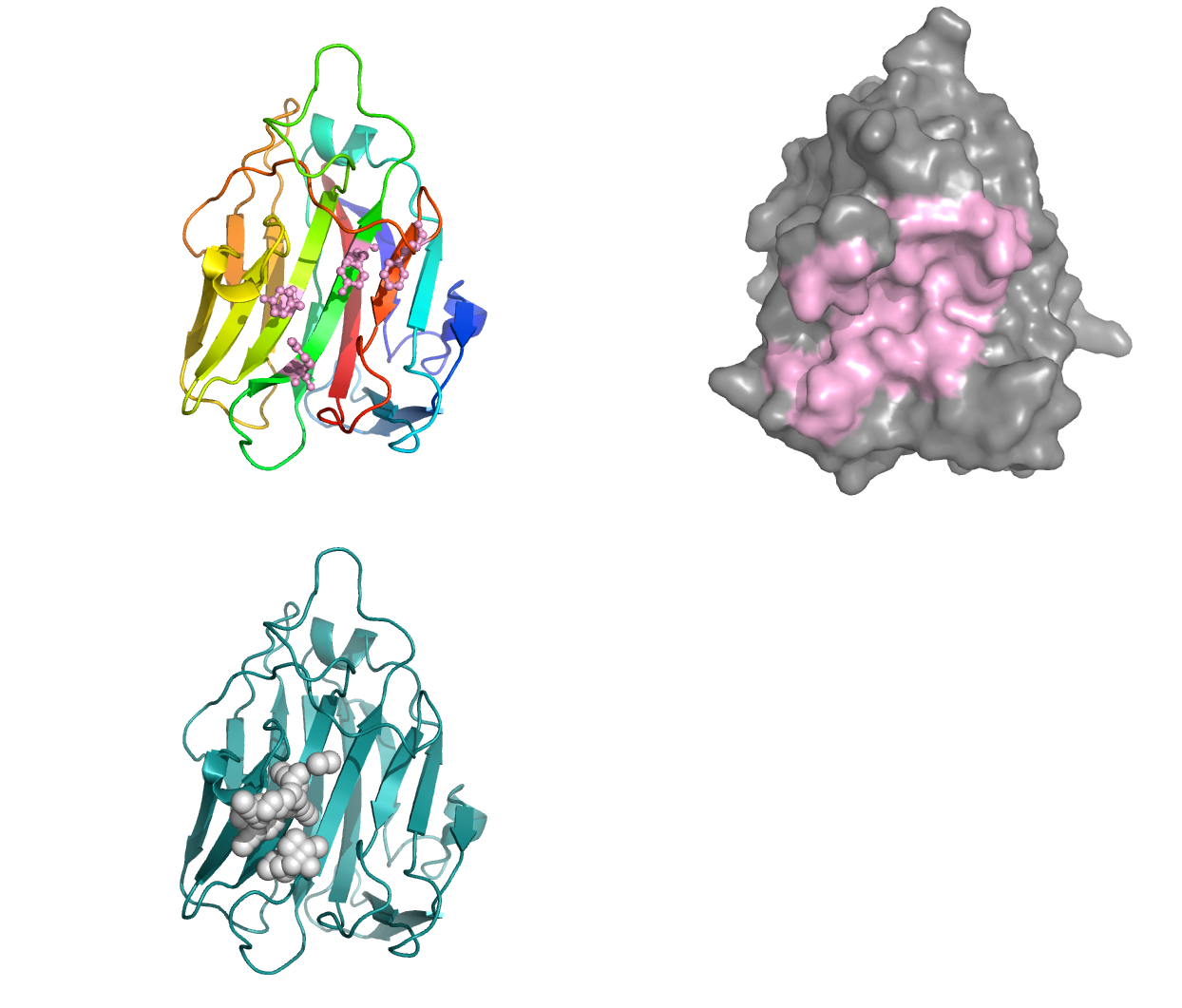
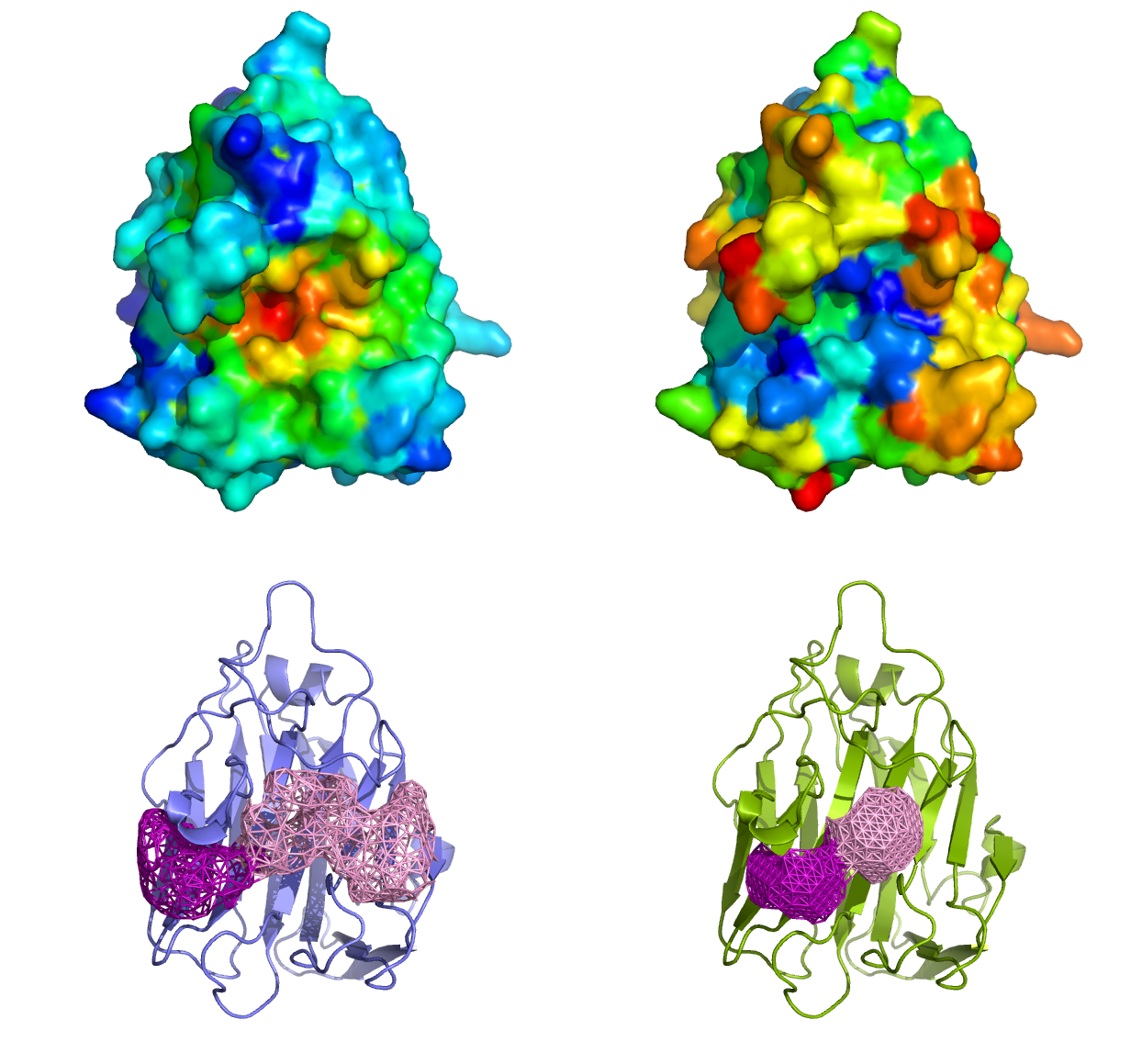
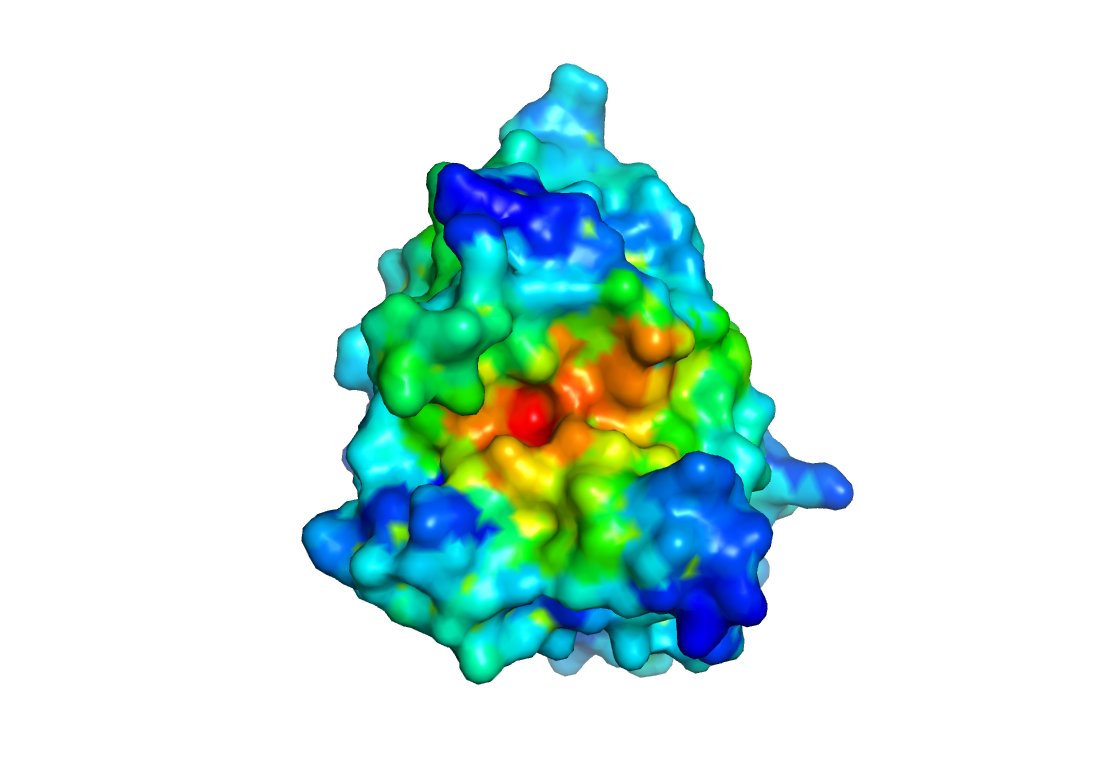
Superimposition of vb\_24B\_21 JRD with its closest ligand-bound structural neighbours identified using DALI (Holm, 2019). Structures share a close agreement in their secondary structure topology, but diverge in the length and conformation of their many loops. Carbohydrate ligands are displayed. Vb\_24B\_21 JRD is shown in blue.

**Bioinformatics analysis of vb\_24B\_21 JRD binding groove**

The jelly-roll fold has two distinct locations for binding sites: (1) a site within the variable loop cluster that interconnects the β-strands at one end of the β-sandwich, and (2) a groove or pocket in the concave face of the β-sandwich. In a given JRD, both or either site can be operational (Abbott and van Bueren, 2014). The structural neighbours of vb\_24B\_21 reveal bound carbohydrate ligands in either site 1 or 2 (**Fig 2**). As a wide variety of structure-based bioinformatics methods are available for predicting the location of binding sites in proteins (Rigden, 2017), we aimed to reveal the ligand binding sites in the JRD from vb\_24B\_21 by studying its surface properties. For this, we used the crystal structure to analyze (*i*) its surface topology in Profunc (Laskowski et al, 2005) and (*ii*) its surface atom type propensities using STP (Mehio et al, 2010) and LISE (Xie et al, 2013) (**Fig 3A-C**). The results consistently confirmed the observed groove as a ligand binding site, with no secondary candidate sites. In all cases, the reverse side of the protein showed no strong features consistent with ligand binding.

Next, we explored the local characteristics of the binding groove to search for evidence that it could accommodate carbohydrate molecules. A notable feature common among carbohydrate-binding sites is the presence of solvent-exposed aromatic residues which can favourably interact with the planar hydrophobic faces of ring-forming saccharide moieties (Malik and Ahmad, 2007; Gilbert et al., 2013). The vb\_24B\_21 JRD contains four such residues in its proposed binding groove: Y510, Y515, Y540, F611 (**Fig 1D**). An analysis of sequence conservation using Consurf (Ashkenazy et al., 2010) revealed that the three Tyr residues are moderately to strongly conserved (Y510 is often conservatively replaced by phenylalanine) across homologs pointing to a possible functional role (**Fig 3D**). Further, we analysed the groove with the ISMBLab server in carbohydrate mode. This uses probability density distributions to identify interacting atoms on protein surfaces (Tsai et al., 2012). It predicted the full-length of the groove to be engaged in carbohydrate binding (**Fig 3E**). In brief, there is unanimity among the predictions in pointing consistently towards carbohydrate binding in the β-sandwich groove, also associated with binding across the superfamily.

**A B C D E**



**Fig 3. The groove of vb\_24B\_21 JRD is a predicted carbohydrate binding site**

**A.** Identification of triplets of adjacent surface atomic groups that can support protein-ligand interactions calculated in STP (Mehio et al, 2010). The surface of vb\_24B\_21 JRD is coloured in a red (highest ligand binding probability) to blue (lowest probability). Triplets with highest probability in this calculation are centered around residues Y515 and Y540; **B.** ConSurf (Ashkenazy et al, 2010) sequence conservation mapping from white (not conserved) to black (strictly conserved); **C.** Cavities calculated using ProFunc (Laskowski et al, 2005). The contiguous top and second ranked cavities (pink and purple, respectively) are shown as a wire mesh; **D.** Detection of triplets of protein surface atoms that are statistically enriched at ligand-binding sites as calculated by LISE (Xie et al, 2013). The contiguous top and third ranked cavities (pink and purple, respectively) are shown as a wire mesh; **E.** The ISMBLab server (Tsai et al, 2012) strongly identifies the cleft as carbohydrate binding. All residues in the top-ranking patch, using the Support Vector Machine methodology, are shown in pink.

**Conclusion**

Carbohydrate esterases are hydrolytic enzymes that catalyze the removal of ester-based chemical modifications in carbohydrates, commonly performing de-O or de-N-acylation. They are currently classified in 16 families in the CAZy database. Esterase vb\_24B\_21 is a close homolog of NanS (currently in an unclassified section of CAZy), with which it shares all active site atypical features that classify it as a group II SGNH esterase.

Phage vb\_24B\_21 (but not bacterial NanS) contains a DUF domain in N-terminal position and a jelly-roll at its C-terminus. Our preliminary modelling indicates that DUF1737 might adopt an α+β fold, but do not allow us to propose a hypothesis with regard to its function. In contrast, the crystal structure elucidation of the jelly-roll domain in C-terminal position reveals that it predictably acts as a non-catalytic, ancillary CBM. CBMs potentiate the activity of their associated catalytic domains against carbohydrate substrates. They target the parent enzyme to the substrate, increasing its local concentration and enhancing the rate and efficiency of catalysis (Boraston et al, 2004; Gilbert et al, 2013). In carbohydrate processing enzymes – including vb\_24B\_21 – catalytic modules and CBMs often occur as discrete structural domains in a single polypeptide chain. CBMs are grouped into 86 sequence-based CAZy families, which comprise 7 different fold-families (Boraston et al, 2004). The largest, fold-family 1, encompasses CBMs with a β-sandwich fold that include lectins. The CBM from vb\_24B\_21 can be classed in this group. Multiple lines of evidence point to the existence of a single carbohydrate binding site in vb\_24B\_21, located in a surface groove on the concave face of the domain (site 2). CBM fold families and location of the binding site in the fold, however, are not indicative of substrate specificity. Moreover, the substrate specificities of the catalytic and CBM domains are often decoupled. Thus, a prediction of ligand specificity for the JRD from vb\_24B\_21 was not possible in this study. Interestingly, the elongated geometry of the groove and the lack of an “aromatic cradle” signature pocket (Abbott and van Bueren, 2014) points to vb\_24B\_21 binding linear carbohydrate chains, as opposed to end units as its nearest neighbour, *Vibrio cholerae* sialidase.Notably, vb\_24B\_21 in the Φ24Β genome and orthologues in other temperate phages are always found upstream of the genes directing phage release and host cell lysis. In Φ24Β, vb\_24B\_21 is only expressed with genes in the lysis cassette and not when the phage is being maintained as a prophage (Veses-Garcia et al., 2015). The fact that it is conserved across a vast swathe of temperate phages and that its expression is linked to prophage induction, suggests that vb\_24B\_21 might play a yet unidentified role in lytic phage replication. Future studies will be required to establish the carbohydrate specificity of vb\_24B\_21 and its role in supporting the activity of vb\_24B\_21 in lysis.

**Data availability**

Structure coordinates and experimental data have been deposited with the PDB (entry 6YP6). X-ray diffraction images for native and S-SAD data have been deposited with Zenodo (http://doi.org/10.5281/zenodo.3754586; http://doi.org/10.5281/zenodo.3754662; http://doi.org/10.5281/zenodo.3754702)

**Acknowledgments**

We gratefully acknowledge the financial support of the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/I013431/1) and the AFF funds of the University of Konstanz.

**Author contribution**

HA, DJR, OM conceived the study; MV performed cloning, protein production, and crystallization; BF, KD performed crystallographic structural elucidation; DJR, BF performed bioinformatics analysis; OM, DJR, BF, HA wrote the manuscript and all authors made revisions.

**References**

Abbott, D.W., van Bueren, A.L., 2014. Using structure to inform carbohydrate binding module function. Curr. Opin. Struct. Biol. 28, 32–40.

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Echols, N., Headd, J.J., Hung, L.W., Jain, S., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R.D., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C., Zwart, P.H., 2011. The Phenix software for automated determination of macromolecular structures. Methods 55, 94-106.

Allison, H.E., 2007. Stx-phages: Drivers and mediators of the evolution of STEC and STEC-like pathogens. Future Microbiol. 2, 165–174.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N., 2010. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res*.* 38, W529-33.

Boraston, A.B., Bolam, D.N., Gilbert, H.J., Davies, G.J., 2004 Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. Biochem. J. 382, 769–781.

Brinda, K. V., Surolia, A., Vishveshwara, S., 2005. Insights into the quaternary association of proteins through structure graphs: A case study of lectins. Biochem. J. 391, 1-15.

Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., 2010. Features and development of Coot. Acta Crystallogr. Sect. D Biol. Crystallogr. 66, 486–501.

Feuerbaum, S., Saile, N., Pohlentz, G., Müthing, J., Schmidt, H., 2018. De-O-Acetylation of mucin-derived sialic acids by recombinant NanS-p esterases of Escherichia coli O157:H7 strain EDL933. Int. J. Med. Microbiol. 308, 1113–1120.

Furnham, N., Holliday, G.L., De Beer, T.A.P., Jacobsen, J.O.B., Pearson, W.R., Thornton, J.M., 2014. The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res. 42, D485-9

Gilbert, H.J., Knox, J.P., Boraston, A.B., 2013. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. Curr. Opin. Struct. Biol. 23, 669–677.

Holm, L., 2019. Benchmarking fold detection by DaliLite v.5. Bioinformatics 35, 5326–5327.

Kabsch, W., 2010. XDS. Acta Crystallogr. D. Biol. Crystallogr. 66, 125–32.

Tsai, K.C., Jian, J.W., Yang, E.W., Hsu, P.C., Peng, H.P., Chen, C.T., Chen, J.B., Chang, J.Y., Hsu, W.L., Yang, A.S., 2012. Prediction of Carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. PLoS One 7, e40846.

Krüger, A., Lucchesi, P.M.A., 2015. Shiga toxins and stx phages: Highly diverse entities. Microbiol. 161, 1–12.

Laskowski, R.A., Watson, J.D., Thornton, J.M., 2005. ProFunc: A server for predicting protein function from 3D structure. Nucleic Acids Res. 33, W89-93.

Lo, Y.C., Lin, S.C., Shaw, J.F., Liaw, Y.C., 2003. Crystal structure of Escherichia coli thioesterase I/protease I/lysophospholipase L1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases through a conserved hydrogen bond network. J. Mol. Biol. 330, 539–551.

Loris, R., Hamelryck, T., Bouckaert, J., Wyns, L. 1998. Legume lectin structure. Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol. 1383, 9–36.

Malik, A., Ahmad, S., 2007. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct. Biol*.* 7, 1.

Mehio, W., Kemp, G.J.L., Taylor, P., Walkinshaw, M.D., 2010. Identification of protein binding surfaces using surface triplet propensities. Bioinformatics 26, 2549–2555.

Mølgaard, A., Kauppinen, S., Larsen, S., 2000. Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. Structure 8, 373–383.

Moraes, J.P.A., Pappa, G.L., Pires, D.E.V., Izidoro, S.C., 2017. GASS-WEB: A web server for identifying enzyme active sites based on genetic algorithms. Nucleic Acids Res. 45, W315–W319.

Nadzirin, N., Gardiner, E.J., Willett, P., Artymiuk, P.J., Firdaus-Raih, M., 2012. SPRITE and ASSAM: Web servers for side chain 3D-motif searching in protein structures. Nucleic Acids Res. 40, W380-6.

Nakamura, A.M., Nascimento, A.S., Polikarpov, I., 2017. Structural diversity of carbohydrate esterases. Biotechnol. Res. Innov. 1, 35–51.

Nübling, S., Eisele, T., Stöber, H., Funk, J., Polzin, S., Fischer, L., Schmidt, H., 2014. Bacteriophage 933W encodes a functional esterase downstream of the Shiga toxin 2a operon. Int. J. Med. Microbiol. 304, 269–274.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., et al., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702.

Rangarajan, E.S., Ruane, K.M., Proteau, A., Schrag, J.D., Valladares, R., Gonzalez, C.F., Gilbert, M., Yakunin, A.F., Cygler, M., 2011. Structural and enzymatic characterization of NanS (YjhS), a 9-O-acetyl N-acetylneuraminic acid esterase from Escherichia coli O157:H7. Protein Sci. 20, 1208–1219.

Rangel, A., Steenbergen, S.M., Vimr, E.R., 2016. Unexpected diversity of Escherichia coli sialate O-acetyl esterase NanS. J. Bacteriol. 198, 2803–2809.

Rigden, D.J., 2017. Function prediction using patches, pockets and other surface properties. In *From Protein Structure to Function with Bioinformatics: Second Edition* pp 327–360. Springer Netherlands

Saile, N., Voigt, A., Kessler, S., Stressler, T., Klumpp, J., Fischer, L., Schmidt, H., 2016. Escherichia coli O157:H7 strain EDL933 harbors multiple functional prophage-associated genes necessary for the utilization of 5-N-acetyl-9-O-acetyl neuraminic acid as a growth substrate. Appl. Environ. Microbiol. 82, 5940–5950.

Sheldrick, G.M., 2010. Experimental phasing with SHELXC/D/E: Combining chain tracing with density modification. Acta Crystallogr. Sect. D Biol. Crystallogr. 66, 479–485.

Smith, D.L., Rooks, D.J., Fogg, P.C.M., Darby, A.C., Thomson, N.R., McCarthy, A.J., Allison, H.E., 2012. Comparative genomics of Shiga toxin encoding bacteriophages. BMC Genomics 13, 311.

Smith, D.L., Wareing, B.M., Fogg, P.C.M., Riley, L.M., Spencer, M., Cox, M.J., Saunders, J.R., McCarthy, A.J., Allison, H.E., 2007. Multilocus characterization scheme for shiga toxin-encoding bacteriophages. Appl. Environ. Microbiol*.* 73, 8032–8040.

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol*.* 35, 1026–1028.

Stokes-Rees, I., Sliz, P., 2010. Protein structure determination by exhaustive search of protein data bank derived databases. Proc. Natl. Acad. Sci. U.S.A*.* 107, 21476–21481.

Steenbergen, S.M., Jirik, J.L., Vimr, E.R., 2009. YjhS (NanS) is required for Escherichia coli to grow on 9-O-acetylated N-acetylneuraminic acid. J Bacteriol. 191, 7134-9.

Tsai, K.C., Jian, J.W., Yang, E.W., Hsu, P.C., Peng, H.P., Chen, C.T., Chen, J.B., Chang, J.Y., Hsu, W.L., Yang, A.S., 2012. Prediction of Carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. PLoS One 7, e40846.

Veses-Garcia, M., Liu, X., Rigden, D.J., Kenny, J.G., McCarthy, A.J., Allison, H.E., 2015. Transcriptomic analysis of shiga-toxigenic bacteriophage carriage reveals a profound regulatory effect on acid resistance in Escherichia coli. Appl. Environ. Microbiol. 81, 8118–8125

Webb, B., Sali, A., 2016. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinforma. 2016, 5.6.1-5.6.37

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., Jain, S., Lewis, S.M., Arendall, W.B., Snoeyink, J., Adams, P.D., Lovell, S.C., Richardson, J.S., Richardson, D.C. 2018. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 27, 293–315.

Xie, Z.R., Liu, C.K., Hsiao, F.C., Yao, A., Hwang, M.J., 2013. LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. Nucleic Acids Res. 41, W292-6.

**SUPPLEMENTARY MATERIALS**

**Structural annotation of the conserved carbohydrate esterase vb\_24B\_21 from Shiga toxin-encoding bacteriophage Φ24B**

Franke et al, 2020

**Fig S1: Sequence alignment of the esterase domains of vb\_24B\_21 and NanS**

Sequence alignment used T-coffee (Notredame et al, 2000).

****

**References**

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302, 205-17.