

## **Expansion of the diversity of dispersin scaffolds**

MALES, Alexandra, MOROZ, Olga V <<http://orcid.org/0000-0002-0354-6119>>, BLAGOVA, Elena <<http://orcid.org/0000-0002-2041-3043>>, MUNCH, Astrid, HANSEN, Gustav H, JOHANSEN, Annette H, ØSTERGAARD, Lars H, SEGURA, Dorotea R, EDDENDEN, Alexander, DUE, Anne V, GUDMAND, Martin, SALOMON, Jesper <<http://orcid.org/0000-0003-2445-6582>>, SØRENSEN, Sebastian R, FRANCO CAIRO, João Paulo L, NITZ, Mark, PACHE, Roland A <<http://orcid.org/0000-0002-4723-6729>>, VEJBORG, Rebecca M, BHOSALE, Sandeep, VOCADLO, David J, DAVIES, Gideon J <<http://orcid.org/0000-0002-7343-776X>> and WILSON, Keith S <<http://orcid.org/0000-0002-3581-2194>>

Available from Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/36722/>

---

This document is the Published Version [VoR]

**Citation:**

MALES, Alexandra, MOROZ, Olga V, BLAGOVA, Elena, MUNCH, Astrid, HANSEN, Gustav H, JOHANSEN, Annette H, ØSTERGAARD, Lars H, SEGURA, Dorotea R, EDDENDEN, Alexander, DUE, Anne V, GUDMAND, Martin, SALOMON, Jesper, SØRENSEN, Sebastian R, FRANCO CAIRO, João Paulo L, NITZ, Mark, PACHE, Roland A, VEJBORG, Rebecca M, BHOSALE, Sandeep, VOCADLO, David J, DAVIES, Gideon J and WILSON, Keith S (2025). Expansion of the diversity of dispersin scaffolds. *Acta Crystallographica Section D Structural Biology*, 81 (3), 130-146. [Article]

---

**Copyright and re-use policy**

See <http://shura.shu.ac.uk/information.html>

Received 15 July 2024

Accepted 6 February 2025

Edited by R. J. Read, University of Cambridge,  
United Kingdom

‡ Current address: Sheffield Hallam University,  
City Campus, Howard Street, Sheffield S1 1WB,  
United Kingdom.

§ Joint first authors.

**Keywords:** enzyme catalysis; glycoside hydrolases; poly-*N*-acetylglucosamine; protein crystallography; GH20.

**PDB references:** DispTs3, 8qak; DispTs2, 8qb6; DispLp, 8qce; DispTs2, complex with di-NAG-thiazoline, 9hta

**Supporting information:** this article has supporting information at journals.iucr.org/d

# Expansion of the diversity of dispersin scaffolds

**Alexandra Males,<sup>a</sup>‡§ Olga V. Moroz,<sup>a</sup>§ Elena Blagova,<sup>a</sup> Astrid Munch,<sup>b</sup> Gustav H. Hansen,<sup>b</sup> Annette H. Johansen,<sup>b</sup> Lars H. Østergaard,<sup>b</sup> Dorotea R. Segura,<sup>b</sup> Alexander Eddenden,<sup>c</sup> Anne V. Due,<sup>d</sup> Martin Gudmand,<sup>b</sup> Jesper Salomon,<sup>b</sup> Sebastian R. Sørensen,<sup>b</sup> João Paulo L. Franco Cairo,<sup>a</sup> Mark Nitz,<sup>c</sup> Roland A. Pache,<sup>b</sup> Rebecca M. Vejborg,<sup>d</sup> Sandeep Bhosale,<sup>e</sup> David J. Vocadlo,<sup>e</sup> Gideon J. Davies<sup>a\*</sup> and Keith S. Wilson<sup>a\*</sup>**

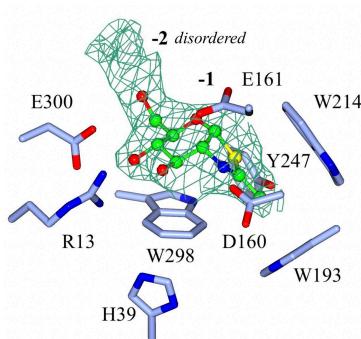
<sup>a</sup>York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, United Kingdom,

<sup>b</sup>Novonesis A/S, Biologiens Vej 2, 2800 Kongens Lyngby, Denmark, <sup>c</sup>Department of Chemistry, University of Toronto, Toronto, Ontario, Canada, <sup>d</sup>Novonesis A/S, Krogshøjvej 36, 2880 Bagsværd, Denmark, and <sup>e</sup>Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. \*Correspondence e-mail: gideon.davies@york.ac.uk, keith.wilson@york.ac.uk

Microorganisms are known to secrete copious amounts of extracellular polymeric substances (EPS) that form complex matrices around the cells to shield them against external stresses, to maintain structural integrity and to influence their environment. Many microorganisms also secrete enzymes that are capable of remodelling or degrading EPS in response to various environmental cues. One key enzyme class is the poly- $\beta$ -1,6-linked *N*-acetyl- $\alpha$ -D-glucosamine (PNAG)-degrading glycoside hydrolases, of which the canonical member is dispersin B (DspB) from CAZy family GH20. We sought to test the hypothesis that PNAG-degrading enzymes would be present across family GH20, resulting in expansion of the sequence and structural space and thus the availability of PNAGases. Phylogenetic analysis revealed that several microorganisms contain potential DspB-like enzymes. Six of these were expressed and characterized, and four crystal structures were determined (two of which were in complex with the established GH20 inhibitor 6-acetamido-6-deoxy-castanospermine and one with a bespoke disaccharide  $\beta$ -1,6-linked thiazoline inhibitor). One enzyme expressed rather poorly, which restricted crystal screening and did not allow activity measurements. Using synthetic PNAG oligomers and MALDI-TOF analysis, two of the five enzymes tested showed preferential endo hydrolytic activity. Their sequences, having only 26% identity to the pioneer enzyme DspB, highlight the considerable array of previously unconsidered dispersins in nature, greatly expanding the range of potential dispersin backbones available for societal application and engineering

## 1. Introduction

In nature, microorganisms actively react to and influence the environment in which they live by producing and secreting a wide range of biological molecules and chemical compounds. This allows them to protect themselves, to compete in and to exploit their immediate surroundings. Many microbial species are known to produce copious amounts of extracellular polymeric substances (EPS), which can protect them against multiple environmental stresses (Yin *et al.*, 2019). While the exact composition of the extracellular matrix depends largely on the microbial species and the environmental cues governing expression, it usually consists of a complex mixture of molecules; this includes proteins, nucleic acids and exopolysaccharides. Poly-*N*-acetylglucosamine (PNAG), also known as polysaccharide intracellular adhesin (PIA), is a homopolymer of linear chains of partially de-*N*-acetylated  $\beta$ -1,6-linked *N*-acetyl- $\alpha$ -D-glucosamine ( $\beta$ -1,6-GlcNAc). This key exopolysaccharide is produced by a wide range of microbial



OPEN ACCESS

Published under a CC BY 4.0 licence

species, including both Gram-negative and Gram-positive bacteria, such as *Escherichia coli*, *Staphylococcus aureus* (Cramton *et al.*, 1999), *Actinobacillus actinomycetemcomitans* (Kaplan *et al.*, 2004) and *Bacillus subtilis* (Roux *et al.*, 2015), as well as several protozoan and fungal species (Cywes-Bentley *et al.*, 2013); these species cause over 50% of nosocomial infections (Jamal *et al.*, 2018). A collection of four proteins are involved in the biosynthesis, modification and excretion of PNAG and are encoded either by the *pgaABCD* genes identified in the Gram-negative bacteria *A. actinomycetemcomitans* (Kaplan *et al.*, 2004), *E. coli* (Wang *et al.*, 2004) and *Yersinia pestis* (Lillard *et al.*, 1997) or the *icaABCD* genes of the intracellular adhesion (*ica*) locus present in Gram-positive bacteria such as *Staphylococcus* spp. (Cramton *et al.*, 1999). Most of the *PgaABCD* proteins share sequence and, where known, structural homology with their *ica* operon counterparts. Some species of bacteria, fungi and protozoa have been shown to produce PNAG although they lack a known genetic locus for its synthesis (Cywes-Bentley *et al.*, 2013), indicating a convergent evolutionary mechanism for the acquisition of PNAG synthesis with potentially considerable significance for microbial biology. Due to the ubiquity of PNAG, its production has been proposed to rival that of other common polysaccharides such as chitin and cellulose (Cywes-Bentley *et al.*, 2013).

The reservoir of cells within a sessile community, whilst beneficial to pathogen survival, can be detrimental to the external environment. The colonization of bacteria on abiotic manmade and natural surfaces causes significant problems in the food sector, medicine (for example drug resistance, mammalian infections, adhesion to implants and valves) and industry (for example machine fouling) (Donlan, 2002). Application of antibiotics directly onto sessile cells requires 10–1000 times the minimum inhibitory concentration of antibiotics needed to inhibit the planktonic form (Ceri *et al.*, 1999). New strategies to reduce the impact of antibacterial resistance while treating infections are urgently required.

Several microorganisms secrete enzymes that are capable of directly degrading EPS, enabling them to control the composition of the matrix in which they are situated and facilitate dispersion under unfavourable conditions. PNAG-hydrolyzing enzymes have been classed into two separate glycoside hydrolase CAZy families, GH20 and GH153 (<http://www.cazy.org>; Drula *et al.*, 2022). The division was based on limited sequence similarity, the presence or absence of PNAG-biosynthesis genes located within the same operon, the presence of a second deacetylase domain and the preference for deacetylated GlcNAc in different active-site subsites.

Translated from the *pgaABCD* operon, *PgaB* is a dual-functioning enzyme which is a fusion of CAZy family GH153 and CE4 enzymes (Lombard *et al.*, 2014). The N-terminal domain acts as a deacetylase for PNAG, resulting in positively charged PNAG, allowing it to interact more readily with the negatively charged cell membrane surface. The C-terminal domain functions as a glycoside hydrolase and cleaves the  $\beta$ -1,6-glycosidic bonds of the PNAG polysaccharide, recognizing a GlcN–GlcNAc–GlcNAc motif in the  $-3$ ,  $-2$ ,  $-1$  subsites (Little *et al.*, 2018). Structures of *PgaB* from *Bordetella bronchiseptica* and *E. coli* have provided vital insight into the catalytic mechanisms (Little *et al.*, 2018).

*Dispersin B* (*DspB*), isolated from *A. actinomycetemcomitans*, was one of the first identified examples of an enzyme that is capable of degrading PNAG (Kaplan, Meyenhofer *et al.*, 2003). The enzyme is a  $\beta$ -hexosaminidase belonging to CAZy glycoside hydrolase family GH20 and its crystal structure has been reported (Lombard *et al.*, 2014; Ramasubbu *et al.*, 2005). *DspB* is not merely the canonical dispersin; it is currently quite unique in the academic literature. Although the CAZy GH20 family contains over 10 000 members ( $>129$  proteins have been characterized with 27 crystal structures, the first being the chitobiase from *Serratia marcescens*; Tews *et al.*, 1996), only two GH20 enzymes have been classified as a dispersin: *DspB* from *A. actinomycetemcomitans*, mentioned above, and *DspB* from *A. pleuropneumoniae* (Kaplan *et al.*, 2004). This inspired us to search the genomic resource for new dispersins.

*DspB* is best considered to be an endo enzyme with a kinetic preference for deacetylated substrates (that is, glucosamine, GlcN) in both the  $-2$  and  $+2$  subsites. This is reflected in a faster hydrolysis of partially deacetylated substrates (Wang *et al.*, 2019). This preference has been attributed to the charge–charge interactions between the cationic glucosamine and anionic aspartates: Asp147 in the  $+2$  subsite and Asp245 in the  $+1$  subsite (Breslawec *et al.*, 2021). However, when *DspB* is provided with an unfavourable substrate, it can also act in an ‘exo’ manner, bypassing deacetylated units to ensure that the *N*-acetyl group of GlcNAc lies in the  $-1$  subsite (Wang *et al.*, 2019). The only published crystal structure of *DspB* (PDB entry 1yht; Ramasubbu *et al.*, 2005) contains glycerol in the  $-1$  subsite and this structure has been used to model ligands in the active site. There are no crystal structures of dispersins in complex with a more informative ligand.

In order both to expand the sequence and structural diversity of known dispersins and to provide structural insight into ligand binding, here we report the cloning, expression and characterization, through inhibition studies and their potential to hydrolyse colourimetric and fluorometric glycoside substrates and bacterially derived PNAG, of five new bacterial GH20 dispersins, henceforth termed *DispTs*, *DispTs2* and *DispTs3* (from *Terribacillus saccharophilus*), *DispLp* (from *Lactiplantibacillus paraplantarum*, formerly known as *Lactobacillus paraplantarum*) and *DispSf* (from *Mammaliicoccus fleurettii*, formerly known as *Staphylococcus fleurettii*). A sixth sample initially included in the study, *DispCo* (from *Curtobacterium oceanosedimentum*, chosen to further increase the dispersin diversity; we aimed to have enzymes from separate clades), could only be produced in very limited quantities and was easily degraded, which did not allow activity experiments. To summarize, the new dispersins are phylogenetically distinct from *DspB*; separated by different phyla, the new dispersins are present in Actinobacteria and Firmicutes compared with Proteobacteria. The dispersins were exposed to both synthesized (fully acetylated) and crude (partially deacetylated) PNAG; hydrolysis of biologically produced PNAG was confirmed through an indirect enzyme-linked immunosorbent assay (ELISA) and hydrolysis of synthetic PNAG was

**Table 1**

Enzyme designations, corresponding bacterial strain sources and DNA accession numbers.

Enzyme	Donor organism	GenBank ID
DispTs	<i>Terribacillus saccharophilus</i>	OM214561
DispTs2	<i>Terribacillus saccharophilus</i>	OM214562
DispTs3	<i>Terribacillus saccharophilus</i>	QQ858607
DispLp	<i>Lactiplantibacillus paraplanitarum</i>	60A0A6M5IIS2
DispSf	<i>Mammaliicoccus fleurettii</i>	94A0A3A0I2R9
DispCo	<i>Curtobacterium oceanosedimentum</i>	OM214560
DspB	<i>Aggregatibacter actinomycetemcomitans</i>	AAP31025.1
PgaB	<i>Bordetella bronchiseptica</i>	CAE32265.1

observed to varying degrees of efficiency via matrix-assisted laser desorption/ionization coupled to time-of-flight mass spectrometry (MALDI-TOF-MS). Four crystal structures were determined for three of these enzymes: one structure without a ligand, one in complex with a bespoke disaccharide  $\beta$ -1,6-linked thiazoline inhibitor, and two in complex with the inhibitor 6-acetamido-6-deoxy-castanospermine (6-Ac-Cas) in the  $-1$  subsite. Ongoing research into the medical and industrial applications of DspB highlights the need to enhance the stability, large-scale production and activity of DspB (Yakandawala *et al.*, 2009; Seijsing *et al.*, 2020; Tan *et al.*, 2015). With the sequences of these new dispersins varying in identity from 26% to 37%, this greatly expands the array of potential dispersin templates available for societal application and engineering.

## 2. Materials and methods

### 2.1. Phylogenetic analysis, structure predictions and comparisons

The dispersin phylogenetic tree was constructed by aligning the sequences taken from a *BLAST* search against DspB using *ClustalW* (Thompson *et al.*, 1994). The tree was constructed and visualized using *MEGA X* (Kumar *et al.*, 2018). For structure predictions (Varadi *et al.*, 2024; Jumper *et al.*, 2021) *alphaFold2\_multimer\_v3* was used, creating relaxed models using all five different *AlphaFold2* network variations, and the best-ranked model was picked (ranking based on the predicted local distance difference test; pLDDT). Structure comparisons were carried out using *SSM* (Krissinel & Henrick, 2004), as incorporated in *Coot* (Emsley *et al.*, 2010).

### 2.2. Genetic cloning of dispersins

The genes of interest chosen for cloning and characterization are listed in Table 1 with their corresponding donor organisms and nucleotide-sequence accession numbers. The genes encoding DispTs, DispTs2, DispTs3, DispLp, DispSf and DispCo were purchased as codon-optimized synthetic genes for *B. subtilis* expression from ThermoFisher Scientific and GeneArt. The gene encoding DspB was purchased from GenScript Biotech and codon-optimized for *E. coli* expression.

The synthetic dispersin genes were inserted into a *Bacillus* expression plasmid as described previously (Moroz *et al.*, 2017). The DNA encoding the mature polypeptide, predicted by *SignalP* (Bendtsen *et al.*, 2004), was cloned with the In-Fusion

HD EcoDry Cloning Kit in frame with the *B. clausii* secretion signal peptide, replacing the native secretion signal sequence, followed by a polyhistidine tag. The residue numbering of the dispersins in the sequence alignment and within the PDB files starts from the beginning of the mature peptide.

Recombinant *B. subtilis* clones containing the individual integrated expression constructs were selected and cultivated on a rotary shaking table in 500 ml baffled Erlenmeyer flasks each containing 100 ml LB medium supplemented with 34 mg l<sup>-1</sup> chloramphenicol. The culture was cultivated for three days at 30°C. The enzyme-containing supernatants were harvested by centrifuging the culture broth for 30 min at 15 000g and the enzymes were purified as described below.

Residues 21–381 of DspB were cloned into the NdeI and KpnI restriction-enzyme cleavage sites of the pET-29b plasmid, which contains a C-terminal hexahistidine tag.

### 2.3. Fermentation, gene expression and protein purification

The culture supernatants were filtered through a Nalgene 0.2 µm filtration unit to remove the rest of the *B. subtilis* host cells. The 0.2 µm filtrates were transferred to 20 mM MES-NaOH pH 6.0 on a G25 Sephadex column (GE Healthcare). The transferred solutions were applied onto a Source Q column (GE Healthcare) equilibrated in 20 mM MES-NaOH pH 6.0. After washing the column extensively with equilibration buffer, the proteins were eluted with a linear NaCl gradient (0–1.0 M NaCl) over five column volumes. Fractions were collected during elution and analysed by SDS-PAGE. Fractions for which only one band was seen after Coomassie staining were pooled and used for further experiments.

The protocol used for the gene expression and protein purification of DspB is described in Ramasubbu *et al.* (2005).

### 2.4. Enzymatic assays using 4-nitrophenyl- $\beta$ -N-acetyl- $\alpha$ -D-glucosaminide (pNP-GlcNAc)

Due to the problems with expression and purification resulting in a limited quality and quantity of the DispCo sample, it was excluded from all activity experiments. A sample of each of the other five dispersins was taken after cell growth, purified and stored in 50 mM HEPES, 100 mM NaCl pH 7. The purified enzyme was subsequently normalized to 25 µM in MQ/0.01% Triton X-100 and further diluted in buffer (100 mM acetic acid, 100 mM MES, 100 mM HEPES, 100 mM glycine pH 5) to a final assay concentration of 1500, 300 or 60 nM. The dispersin was reacted with 6 mM pNP-GlcNAc (CAS No. 459-18-5) for 30 min under gentle shaking (300 rev min<sup>-1</sup>). The total volume of the reaction solution was 100 µl. Reaction was stopped by the addition of 100 µl 0.6 M Na<sub>2</sub>HCO<sub>3</sub> pH 10.3. After allowing the pH to equilibrate for 20 min under gentle shaking (150 rev min<sup>-1</sup>), the endpoint absorbance was measured at 405 nm. All data points were blank-corrected using a sample with 0 nM dispersin.

### 2.5. Enzymatic assays using 4-methylumbelliferyl- $\beta$ -N-acetyl- $\alpha$ -D-glucosaminide (4-MU-GlcNAc)

A sample of each dispersin was taken after cell growth, purified and stored in 50 mM HEPES, 100 mM NaCl pH 7 or

similar. The purified enzyme was subsequently normalized to 20 nM in MQ/0.01% Triton X-100 and further diluted in buffer (20 mM acetic acid, 20 mM MES, 20 mM HEPES, 20 mM glycine pH 5) to a final assay concentration of 20, 4 or 0.8 nM. The dispersin was reacted with 5 mM 4-MU-GlcNAc (CAS No. 37067-30-4) for 20 min under gentle shaking (150 rev min<sup>-1</sup>). The total volume of the reaction solution was 100 µl. Reaction was stopped by the addition of 100 µl 0.6 M Na<sub>2</sub>HCO<sub>3</sub> pH 10.3. After allowing the pH to equilibrate for 5 min under gentle shaking (150 rev min<sup>-1</sup>), the endpoint fluorescence was measured using excitation at 368 nm and emission at 448 nm. All data points were blank-corrected using a sample with 0 nM dispersin.

### 2.6. Enzyme hydrolysis of synthetic PNAG

Synthetic PNAG was produced by an acid-reversion reaction with HF-pyridine as described previously (Leung *et al.*, 2009). The mixed oligomers were fractionated on a BioGel P4 column in dH<sub>2</sub>O and a fraction with lengths of between 6 and 10 GlcNAc units was used in the assays.

The enzymes and PNAG were diluted into 20 mM ammonium acetate pH 6.0 to final concentrations of 10 µM and 1 mg ml<sup>-1</sup>, respectively. In a shaking block, 10 µl reactions were conducted in triplicate at 37°C for 20 h. A 1 µl sample was loaded onto a MALDI 384 ground-steel target plate TF (Bruker Daltonics) and mixed on-plate with 10 mg ml<sup>-1</sup> 2,5-dihydroxybenzoic acid (DHB) dissolved in 50% acetonitrile, 0.1% trifluoroacetic acid followed by air-drying to allow crystallization. The data were collected using an ultraflexXtreme (MALDI-TOF/TOF, Bruker), with the smartbeam-II laser set to 2 kHz in positive-ion mode. The laser power was set to 60% and the ions were acquired in reflector mode (mass range 0–3000 Da) for MS analysis. The data were processed in the Bruker *flexAnalysis* software using a red phosphorus standard as a calibrant. A control reaction with the substrate and BSA was conducted, and no hydrolysis products were observed. Furthermore, the five enzymes were incubated with a chitin heptasaccharide and this substrate was partially degraded.

### 2.7. Extraction of PNAG from *Pseudomonas fluorescens*

A crude PNAG extract was prepared from *P. fluorescens* as follows. The strain was grown in M63 [15 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 100 mM KH<sub>2</sub>PO<sub>4</sub>, 1.8 µM FeSO<sub>4</sub>, 1 mM MgSO<sub>4</sub>·7H<sub>2</sub>O, 0.4% (w/v) glycerol, 0.2% (w/v) casamino acids, 0.0001% (w/v) thiamine] in Corning CellBIND 225 cm<sup>2</sup> angled neck cell-culture flasks with a vent cap (400 ml per flask) at 20°C for three days under static conditions. After cultivation, the culture was pelleted by centrifugation (10 min, 8000g, 25°C), and resuspended in 3 M NaCl to extract the surface-associated EPS much as described previously (Chiba *et al.*, 2015). The PNAG-containing supernatant obtained after a subsequent centrifugation step (10 min, 5000g, 25°C) was stored at -20°C until use.

### 2.8. Quantification of PNAG by indirect ELISA

The crude PNAG extract was diluted 1:10 in 1× PBS and subjected to enzymatic treatment for 1 h at 37°C prior to the ELISA. Nontreated samples and samples treated with heat-inactivated dispersin B were used as controls (1 h, 100°C). For quantification of the residual PNAG, ELISA plates (Nunc MultiSorp) were coated with the PNAG samples for 1 h at room temperature, rinsed and blocked overnight at 4°C (in PBST + 1% BSA). The primary antibody solution [human Anti-PNAG antibody (TAB-799CL), Creative Biolabs, diluted 1:5000 in PBST + 1% BSA] was added and the plates were incubated for 1 h at room temperature. The plates were then rinsed in PBST and treated with HRP-conjugated anti-human secondary antibodies (goat anti-human, Sigma-Aldrich, diluted 1:5000 in PBST) for 1 h, followed by development using the TMB plus2 ready-to-use 3,3',5,5'-tetramethylbenzidine-based chromogenic solution according to the manufacturer's instructions (KemEnTec Dianostics). The absorbance was measured at 450 nm using a spectrophotometer.

### 2.9. Dissociation-constant measurement by isothermal titration calorimetry (ITC)

ITC was performed for three of the enzymes. 200 µM DispTs2 or DispLp and 2000 µM GlcNAc-castanospermine were buffer-matched into 20 mM HEPES pH 7.5 or pH 7.0 and 50 mM NaCl. 50 µM DispSf and 500 µM GlcNAc-castanospermine were buffer-matched into 50 mM HEPES pH 7.5. ITC was performed using a MicroCal ITC200 calorimeter, where GlcNAc-castanospermine was added by syringe with 20 injections to the protein solution in the calorimeter cell at 25°C. A control used GlcNAc-castanospermine injected into buffer in the cell. The dissociation constant ( $K_d$ ), number of sites ( $N$ ) and enthalpy change ( $\Delta H$ ) were calculated using one-site fitting within the *MicroCal PEAQ-ITC Analysis* software (Malvern Panalytical) after subtraction of the control.

### 2.10. Crystallization of the dispersins

Initial crystallization screening was carried out for all six enzymes, including DispCo, using sitting-drop vapour diffusion with drops set up using a Mosquito Crystal liquid-handling robot (STP LabTech) with 150 nl protein solution plus 150 nl reservoir solution in 96-well format plates (MRC 2-well crystallization microplates, SWISSCI) equilibrated against 54 µl reservoir solution. The initial experiments were carried out at room temperature with a variety of commercial screens. We obtained crystals for three of the samples during screening. DispTs was dropped because of its high sequence similarity to DispTs3, DispCo had an additional N-terminal domain, most probably connected to the catalytic domain by a flexible linker, and was prone to rapid degradation, with both factors negatively influencing crystallization, and DispSf did not lead to diffraction-quality crystals.

To crystallize DispTs3 (13.3 mg ml<sup>-1</sup>), an initial seeding stock was made from crystals from JCSG condition H7: 0.2 M

**Table 2**

Data-collection statistics and structure-solution and refinement statistics.

Values in parentheses are for the outer shell.

	DispTs3	DispTs2 + 6-Ac-Cas	DispTs2 + di-NAG-thiazoline	DispLp + 6-Ac-Cas
PDB code	8qak	8qb6	9hta	8qce
Beamline	I03, DLS	I04, DLS	I03, DLS	I03, DLS
Wavelength (Å)	0.976	0.9795	0.976	0.976
Temperature (K)	100	100	100	100
Space group	<i>P</i> 2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	<i>P</i> 3 <sub>2</sub> 1	<i>P</i> 3 <sub>2</sub> 1	<i>P</i> 12 <sub>1</sub> 1
<i>a</i> , <i>b</i> , <i>c</i> (Å)	50.9, 109.4, 131.1	89.7, 89.7, 97.7	90.1, 90.1, 98.1	46.9, 82.8, 80.6
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 90	90, 90, 120	90, 90, 120	90, 98.1, 90
Completeness (%)	100 (100)	100 (100)	93.1 (76.4)	90.5 (40.3)
Multiplicity	4.9 (4.8)	10 (10)	11.6 (10.3)	3.7 (2.2)
$R_{\text{p.i.m.}}^{\dagger}$	8.2 (81.5)	1.0 (62.8)	4.0 (215.1)	5.3 (93.5)
$\langle I/\sigma(I) \rangle$	7.3 (1.2)	17.7 (1.2)	13.3 (0.5)	6.6 (0.7)
Resolution range (Å)	83.97–1.95 (2.0–1.95)	30.03–1.51 (1.54–1.51)	40.93–2.17	38.06–1.05 (1.07–1.05)
CC <sub>1/2</sub> <sup>‡</sup>	0.99 (0.34)	0.99 (0.51)	0.99 (0.53)	0.99 (0.42)
Final $R_{\text{cryst}}/R_{\text{free}}$	0.19/0.23	0.15/0.18	0.21/0.25	0.14/0.17
No. of non-H atoms (chain <i>A/B</i> )				
Protein	2602/2610	2672	2631	2710/2856
Ligand	—	16	14	16/16
Water	333	193	43	772
Solute	8 (ACT)	16 (MPD)	—	7 (PEG), 12 (EDO)
R.m.s. deviations				
Bond lengths (Å)	0.012	0.010	0.0055	0.011
Angles (°)	1.91	1.59	1.4820	1.73
Average <i>B</i> factors (Å <sup>2</sup> )				
Protein	36/31	28.7	60.1	13/17
Ligand	N/A	33.1	60.7	14/16
Water	32	36.0	54.5	32
Ramachandran plot				
Most favoured (%)	97	97.5	96.0	98.7
Allowed (%)	2.8	2.5	4.0	1.3
Outliers (%)	0.02	0	0	0

<sup>†</sup>  $R_{\text{p.i.m.}} = \sum_{hkl} \{1/[N(hkl) - 1]\}^{1/2} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$ . <sup>‡</sup> CC<sub>1/2</sub> is defined in Karplus & Diederichs (2012).

ammonium sulfate, bis-Tris pH 5.5, 25% PEG 3350. The final crystals were obtained after several rounds of microseed matrix screening (MMS; D'Arcy *et al.*, 2014; Shaw Stewart *et al.*, 2011; Shah *et al.*, 2005) using an Oryx8 robot (Douglas Instruments) into Morpheus condition B3: 0.09 *M* Halogens Mix, 30% Glycerol/PEG 4000 Mix.

Crystals of DispTs2 (36 mg ml<sup>-1</sup>) were obtained using MMS (with DispTs3 seeding stocks and then stocks from the new hits for DispTs2) into MPD Screen condition E7: 0.1 *M* citric acid pH 4.0, 20% MPD. The crystals were co-crystallized with 5 mM 6-Ac-Cas (PDB ligand code GC2).

A seeding stock made from crystals of DispLp (19 mg ml<sup>-1</sup>) in Hampton Research Crystal Screen condition D10 (0.2 *M* calcium acetate, 0.1 *M* sodium cacodylate pH 6.5, 18% PEG 8000) was added by MMS into the PACT Screen (Molecular Dimensions). Crystals were obtained in PACT condition A10: 0.2 *M* calcium chloride, 0.1 *M* sodium acetate pH 5.0, 20% PEG 6000). The crystals were co-crystallized with 5 mM 6-Ac-Cas (PDB ligand code GC2) and cryoprotected using ~30% ethylene glycol.

### 2.11. Data collection, structure solution and refinement

All computation was carried out using programs from the CCP4 suite (Agirre *et al.*, 2023). Data were collected at Diamond Light Source (DLS) and processed with *xia2* (Winter, 2010). The data-collection and processing statistics are given in Table 2. The structures of DispTs3 and DispLp

were solved by molecular replacement using *MOLREP* (Vagin & Teplyakov, 2010) with PDB entry 1yht (dispersin B from *A. actinomycetemcomitans*) as the search model, which was selected using *MrBUMP-CCP4mg* to provide a sculpted model (Ramasubbu *et al.*, 2005; Keegan & Winn, 2007). The structure of DispTs2 was solved by molecular replacement using *MOLREP* with DispTs3 as the model (Vagin & Teplyakov, 2010). The chains in all of the protein structures were traced using *Buccaneer* and the structures were refined with *REFMAC* iterated with manual model correction using *Coot* (Murshudov *et al.*, 2011; Emsley *et al.*, 2010; Cowtan, 2006). The quality of the final models was validated using *MolProbity* as part of the *Phenix* package (Adams *et al.*, 2011; Chen *et al.*, 2010).

## 3. Results

### 3.1. Phylogenetic analysis unveils new dispersin scaffolds

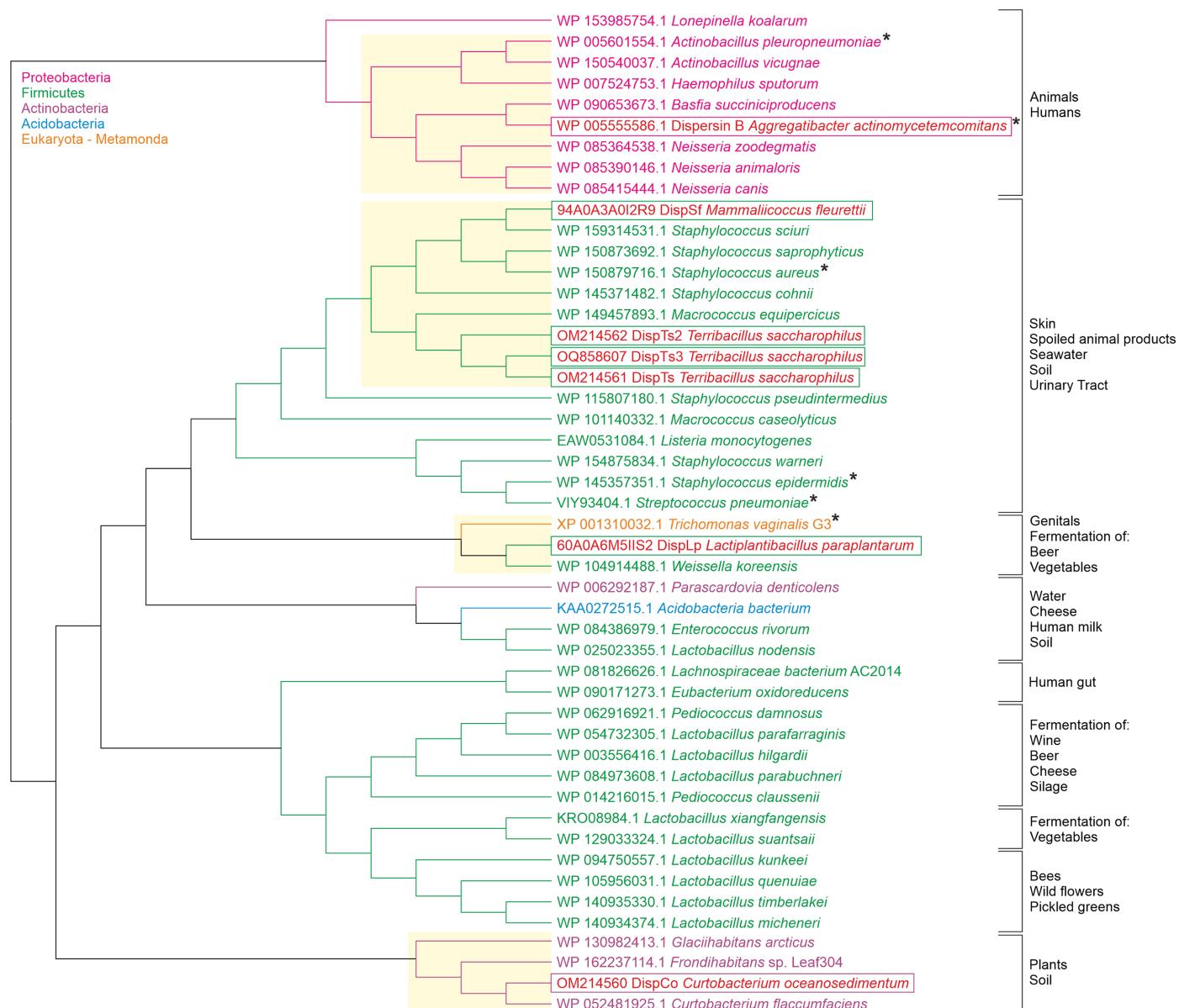
To date, more than 300 bacterial species have been shown to produce PNAG/PIA (Cywes-Bentley *et al.*, 2013) along two distinct evolutionary trajectories involving either the *pga* or *ica* machinery (Bandalovic-Torma *et al.*, 2020). Notably, the mechanisms for cell detachment by sloughing through external forces, proteases or nucleases and detergents, for example phenol-soluble modulins, are dependent on the composition of the extracellular matrix (Guilhen *et al.*, 2017). DspB is to date the sole characterized GH20 dispersin

subfamily member, suggesting that there is probably a plethora of unidentified dispersin or dispersin-like enzymes. This led us to search for a correlation between bacterial species that contain genes encoding putative dispersin enzymes and the presence of a PNAG operon. A *BLAST* search using the sequence for DspB retrieved over 1000 related sequence results with an *E*-value below  $10^{-20}$ .

Sequence alignment of 50 ‘dispersin’ enzymes from different bacterial species allowed the construction of a cladogram using maximum-likelihood methods (Fig. 1). Three distinct clades could be distinguished in the phyla Proteobacteria, Actinobacteria and Firmicutes. The Firmicute clade also contained members of the Acidobacteria and Actino-

bacteria phyla and *Trichomonas vaginalis* from the Eukaryota Metamonada phylum. Similarities between enzyme sequence, clade and the location of bacterial isolation were evident, for example the distinct clade of Actinobacteria can be isolated from plants and soil. No evolutionary information can be deduced from the cladogram since the outgroup, containing the phylum that evolved first, could not be resolved.

From these 50 sequences, six enzymes, representing three different clades, were selected for further analysis. Those chosen are crucial for obtaining dispersin diversity phylogenetically distinct from the clade containing the well characterized dispersin B from *A. actinomycetemcomitans* (Fig. 1). A clade containing DispLp, which has 30% sequence identity



**Figure 1**

Distinct phylogenetic clades separate the dispersins. A cladogram describing the distribution of predicted GH20 dispersin members across different phyla of Bacteria and Eukaryota. The dispersins discussed in the paper are identified in the boxes coloured according to their phylum. Asterisks indicate species for which there is evidence of PNAG expression (Cramton *et al.*, 1999; Kaplan, Ragunath *et al.*, 2003; Cywes-Bentley *et al.*, 2013; Izano *et al.*, 2007; Mack *et al.*, 1996). Clades containing the new dispersin enzymes are highlighted with a yellow background.

to DspB along with potential dispersins from two other species, is separate from members of a second clade containing DispSf (26% sequence identity to DspB), DispTs3 (26% sequence identity to DspB) and DispTs and DispTs2 (26% and 28% sequence identity to DspB, respectively). A third clade was identified containing the Actinobacterium phylum and DispCo (36% identity to DspB).

### 3.2. The putative dispersins were active against aryl glycosides

To determine the activity of the putative dispersins, we first performed activity-concentration kinetics with two different aryl glycosides: *p*NP-GlcNAc and 4-MU-GlcNAc. The activity-concentration profiles of the two assays showed a linear relationship for all dispersin enzymes measured under the chosen conditions (Figs. 2a and 2b). DispLp showed very low activity against both substrates, while DispTs2 showed the highest activity of the tested dispersins. Although DispTs2 and

DispTs3 have the highest sequence similarity, DispTs2 was more active on both aryl glycosides. There were slight differences in substrate specificity when directly comparing the two substrates, most notably for DspB (which showed a preference for *p*NP-GlcNAc) but also DispSf (preference for 4-MU-GlcNAc) and DispTs (preference for 4-MU-GlcNAc). DispLp has a high sequence identity of 30% to DspB, but in marked contrast showed very little activity on these substrates.

Since the highest measured activity was the hydrolysis of 4-MU-GlcNAc by DispTs2, determination of the Michaelis-Menten kinetic parameters was attempted. Although saturating concentrations of 4-MU-GlcNAc were not possible, an observed  $K_m$  of  $2.5 \pm 0.1 \text{ mM}$  and  $V_{\max}$  of  $0.59 \pm 0.02 \mu\text{M s}^{-1}$  were estimated (Fig. 2c). Although the enzymes are active on these substrates, it is likely that neither 4-MU nor *p*NP were well tolerated in the active site of the enzymes, or the substrates need to have longer oligosaccharide moieties consistent with extended subsites and an endo activity of the dispersins (Wang *et al.*, 2020).

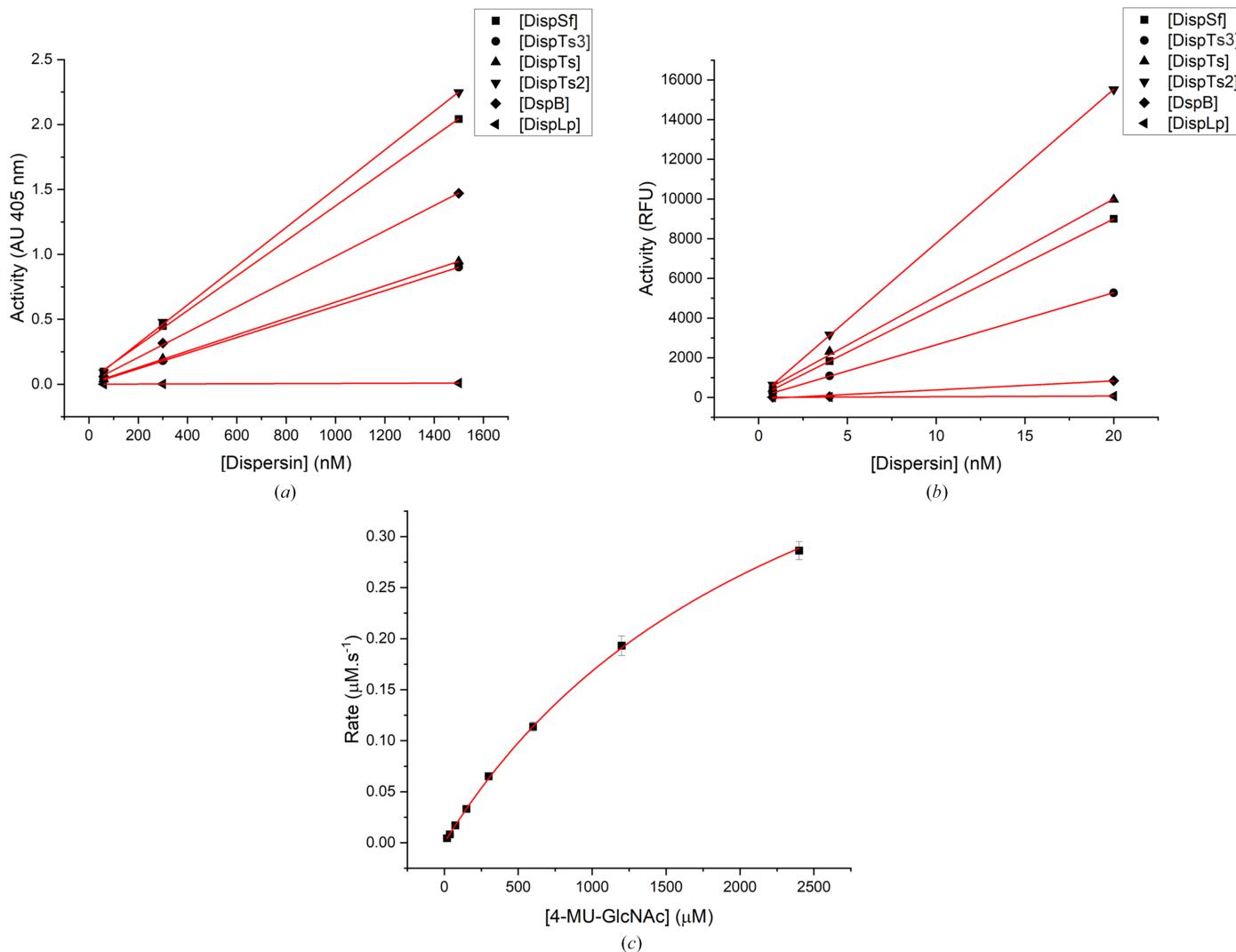


Figure 2

Dispersin activity-concentration profiles for two different substrates. All profiles shown in (a) and (b) are linear under the given conditions. (a) Activity-concentration profiles of all tested dispersins against *p*NP-GlcNAc substrate. (b) As (a) but against 4-MU-GlcNAc substrate. (c) Michaelis-Menten kinetics of DispTs2 using 4-MU-GlcNAc as the substrate.

### 3.3. The dispersins show activity against fully acetylated and partially deacetylated PNAG

After observing the activity of the putative dispersins against aryl glycosides, five enzymes (excluding DispCo) were tested for activity on PNAG. The hydrolysis was monitored using two techniques: MALDI-TOF-MS on fully synthetic (and fully acetylated) PNAG and an indirect ELISA on PNAG isolated from *P. fluorescens*.

Firstly, fully acetylated PNAG of varying lengths, mainly 6–10 GlcNAc units, was chemically synthesized using an acid-reversion reaction as previously reported and the high-molecular-weight fraction was isolated from size-exclusion chromatography (Leung *et al.*, 2009). The enzymes were incubated with the oligomer mixture overnight, the products were observed by MALDI-TOF-MS and the peak areas for each oligomer were compared (Fig. 3a). A clear trend towards a larger percentage of shorter PNAG saccharides (1–5 GlcNAc units) was seen upon incubation of the enzymes with PNAG. DispLp showed the lowest activity compared with the other dispersin enzymes, consistent with the activity data on *p*NP-GlcNAc/4MU-GlcNAc. This could be due to the lack of deacetylation that could be required for optimized binding of the substrate in preferred subsites of the active site. An intense monosaccharide peak ( $\geq 5\%$ ), perhaps inferring a preferentially exo-acting enzyme, is seen for DispTs3, DispTs, DispLp and DispSf. In contrast, less than 1.3% of the sugars hydrolysed by DispTs2 and DspB were monosaccharides, indicating a preference for acting in an endo manner.

Secondly, to verify that the enzymes showed activity on a natural, partially deacetylated, substrate, microbially derived PNAG was purified from *P. fluorescens* and exposed to enzymatic digestion. Activity was measured by indirect ELISA using an anti-PNAG primary antibody. As seen in Fig. 3(b), the novel dispersin enzymes and DspB showed activity on the natural substrate. In comparison to the synthetic PNAG substrate, DispLp showed increased activity on the partially deacetylated natural substrate, suggesting a greater preference for partially deacetylated substrates.

### 3.4. Crystal structures of the new dispersins

To understand the sequence conservation amongst the dispersins, and to better understand the key  $-1$  subsite of these enzymes, which has so far evaded structural dissection, crystal structures of DispTs3, DispTs2 and DispLp were obtained (using seeding methods as discussed in Section 3) at resolutions from 2.0 to 1.05 Å (Supplementary Fig. S1 and Table 2). DispTs3, DispTs2 and DispLp all consist of a single domain with the expected  $(\beta/\alpha)_8$  (TIM)-barrel fold for GH20 catalytic domains (Banner *et al.*, 1975; Tews *et al.*, 1996). The  $\beta$ -strands in the centre form a tunnel atop of which the active site is located in a groove, presumably to allow long chains of PNAG to bind. Superposition of the three dispersin structures and DspB revealed five areas which differed in secondary structure. Several  $\alpha$ -helices within the outer ring of DspB are present as unstructured loops in the new dispersins; in

contrast, loops within DspB have secondary structure in the dispersin variants (Supplementary Fig. S1 and Table S1). The  $\beta$ -strands  $\beta 3$  (Val75–Gly77),  $\beta 4$  (Gly80–Asn84) and  $\beta 5$  (Gly88–Pro90) are absent in DispTs3, DispTs2 and DispLp. On the opposite side of the protein to  $\beta 3$ ,  $\beta 4$  and  $\beta 5$  is an  $\alpha$ -helix,  $\alpha 7$  (Lys246–Met255), in DspB and DispLp; however, this region is present as two short  $\beta$ -strands in DispTs3 and DispTs2. The extra helices and loops present in DspB, and not in the new dispersins, reduce the length of the active-site groove, suggesting that it could be active on shorter PNAG substrates while the other dispersins may be active on longer substrates.

Since DispTs3 and DispTs2 have the highest sequence identity, as expected they have a small r.m.s.d. of 0.63 Å. In comparison, the low sequence identity between DispLp and both DispTs3 and DispTs2 resulted in larger differences; DispTs2 and DispLp have an r.m.s.d. of 1.67 Å and DispTs3 and DispLp have an r.m.s.d. of 2.30 Å.

No X-ray structures were obtained for DispTs, DispSf and DispCo, but *AlphaFold2* predictions (Jumper *et al.*, 2021) resulted in structures similar to the dispersins discussed above, with the most significant differences for DispCo, which has an additional N-terminal domain, with the closest structures being fibronectin III type (FN3) domains, as identified by *GESAMT* (Krissinel, 2012; Supplementary Fig. S2). The relative orientation of the domains is likely to be correct based on the predicted aligned error (PAE) plot, where the blue colour of the regions corresponding to connection between residues from the catalytic and N-terminal domains (adjacent to the upper right and lower left corners of the plot) indicates high confidence of the relative positions of the domains (Supplementary Fig. S2b; Varadi *et al.*, 2024).

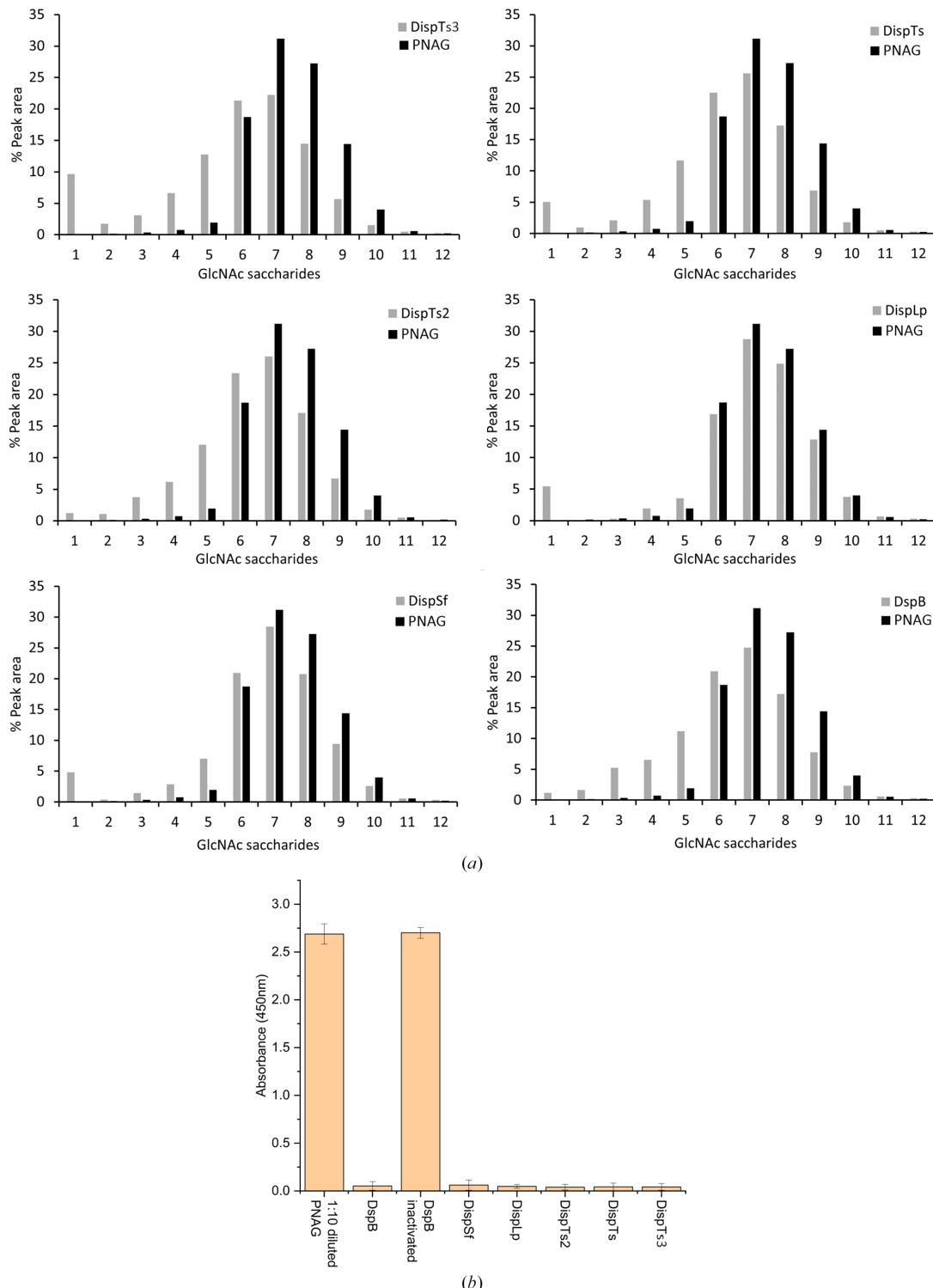
### 3.5. Complexes with 6-Ac-Cas provide insight into the active centre of dispersins

GH20 enzymes use a substrate-assisted catalytic mechanism, also referred to as neighbouring-group participation (NGP), in which the reaction proceeds via the formation and subsequent breakdown of a neutral oxazoline intermediate (Tews *et al.*, 1996; Drouillard *et al.*, 1997; Mark *et al.*, 2001). The acetamido group of the substrate acts as the nucleophile and a glutamate residue acts as the general acid/base (in this example Glu184 in DspB; Fig. 4a). The first insights into substrate distortion and catalysis were provided by studies of the *S. marcescens* chitobiase in complex with chitobiose (a disaccharide of  $\beta$ -1,4-linked GlcNAc; Tews *et al.*, 1996; Drouillard *et al.*, 1997). Whilst the catalytic mechanism is conserved for GH20 dispersins, there has been no information on the mode of ligand binding, with only a glycerol present in the published 3D structure. In order to gain insight into the dispersin active site, we first sought an inhibitor that would be amenable to structural analysis.

The use of iminosugars, which contain a substituted nitrogen in place of the ring oxygen, has provided important mechanistic insights into glycoside hydrolases. 6-Acetamido-6-deoxy-castanospermine (6-Ac-Cas), a derivative of castanospermine

which has a fused 5,6-indolizine ring system, is specific towards enzymes that use neighbouring-group participation and features an acetamido group introduced at the C2 position of the glucopyranose ring (Fig. 4b; Liu *et al.*, 1991). Three members of the GH20 family, the  $\beta$ -N-acetylhexosaminidase

HexA from *Streptomyces coelicolor* A3(2) (ScHexA; PDB entry 4c7f; Thi *et al.*, 2014), the  $\beta$ -hexosaminidase Hex1T from *Paenibacillus* sp. TS12 (PDB entry 3suw; Sumida *et al.*, 2012) and a lacto-*N*-biosidase from *Bifidobacterium bifidum* (PDB entry 5bx; Hattie *et al.*, 2015), as well as a similar



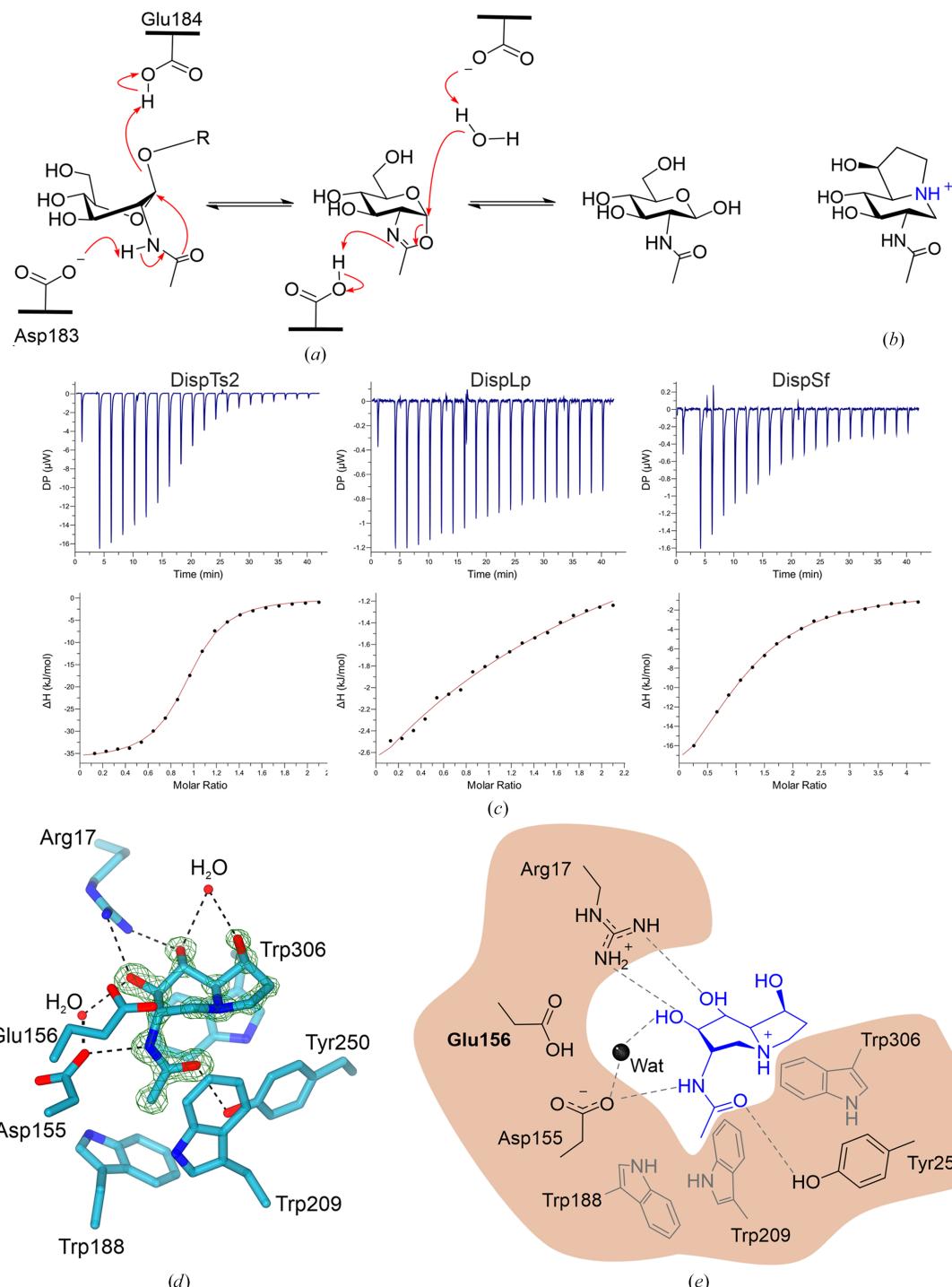
**Figure 3**

Activity of the putative dispersin enzymes on PNAG. (a) The enzymes were exposed to synthesized PNAG. After MALDI-TOF-MS data collection, the peak areas for each GlcNAc saccharide (including different adducts) were calculated. They were then totalled and converted to a percentage. (b) Indirect ELISA assay using crude PNAG from *P. fluorescens*.

neighbouring-group participating family GH84 enzyme from *Bacteroides thetaiotaomicron* (PDB entry 2xj7; Macauley *et al.*, 2010), have previously been crystallized in complex with 6-Ac-Cas in the  $-1$  subsite. Therefore, we sought to determine

whether 6-Ac-Cas was a suitable inhibitor for investigating the mechanism of the dispersin subfamily.

Binding constants for 6-Ac-Cas against a selection of dispersin enzymes were determined by isothermal titration



**Figure 4**

Catalytic mechanism and structure of 6-acetamido-6-deoxy-castanospermine. (a) Neighbouring-group participation catalytic mechanism of DspB. Here, we show the intermediate as a charged oxazolinium ion as predicted from calculations on related systems (Calvelo *et al.*, 2023). (b) Structure of 6-Ac-Cas. (c) Dispersin inhibition by 6-Ac-Cas. Thermodynamics of binding: the raw data are shown in the baseline-adjusted injection profile (top) and the titration curve with one-site fitting in red (bottom). Left: DispTs2,  $0.93 \pm 0.004$  sites,  $-36.7 \pm 0.2 \text{ kJ mol}^{-1}$ . Middle: DispLp,  $0.90 \pm 1.36$  sites,  $-19.1 \pm 0.35 \text{ kJ mol}^{-1}$ . Right: DispSf,  $1.1 \pm 0.02$  sites,  $26.4 \pm 1.11 \text{ kJ mol}^{-1}$ . (d) Active-site residues of DispLp and water molecules with 6-Ac-Cas in complex. Hydrogen bonds are represented by dashed black lines and the maximum-likelihood/ $\sigma_A$ -weighted  $2F_{\text{obs}} - F_{\text{calc}}$  map is shown in green contoured at  $0.90 \text{ e } \text{\AA}^{-3}$ . (e) Scheme of DispLp active-site residue interactions with 6-Ac-Cas.

calorimetry (ITC; Fig. 4c). 6-Ac-Cas has micromolar affinity towards DispTs2 and DispLp, with a  $K_d$  of 6 and 15  $\mu\text{M}$ , respectively. This is similar to literature values for other GH20 enzymes; 6-Ac-Cas with a GH20 exo- $\beta$ -N-acetylhexosaminidase from *Vibrio harveyi* had a  $K_d$  of 12.9  $\mu\text{M}$  (Meekrathok *et al.*, 2018). In marked contrast, the  $K_d$  of DispLp for 6-Ac-Cas was 1.12 mM.

Enzyme-inhibitor complexes were obtained with DispTs2 at a resolution of 1.51  $\text{\AA}$  and DispLp at a resolution of 1.05  $\text{\AA}$  after soaking the crystals in a solution containing 6-Ac-Cas, which we showed to be a potent inhibitor (Table 2). In the active site of both enzymes 6-Ac-Cas was distorted into a  $^1\text{S}_3$  conformation, which is consistent with the proposed catalytic pathway of GH20 enzymes based upon the  $^1\text{S}_3/{}^4\text{E}$  (Michaelis complex/product) conformation for GH20 enzymes that was first observed for the *S. marcescens* chitobiase (Tews *et al.*, 1996).

6-Ac-Cas was bound into the highly negatively charged  $-1$  subsite notably via aspartate, glutamate and tyrosine residues (Figs. 4d and 4e and Supplementary Fig. S3c). The *N*-acetyl group is positioned in a hydrophobic pocket within the  $\beta$ -barrel. The acetamido carbonyl oxygen of 6-Ac-Cas is within hydrogen-bonding distance of the amine moiety of the indolizine ring at 2.56 and 2.61  $\text{\AA}$  in the active sites of DispTs2 and DispLp, respectively (Figs. 4d and 4e). Two key residues are involved in the NGP mechanism: a glutamate residue acts as the general acid/base (Glu184 in DspB, for example) and an aspartate residue deprotonates the *N*-acetamido group (Asp183 in DspB) (Fig. 4a). The catalytic glutamate residues, Glu161 in DispTs2 and Glu156 in DispLp, are 3.4 and 3.6  $\text{\AA}$  away from the anomeric carbon of 6-Ac-Cas, respectively, consistent with closer positioning to the glycosidic oxygen during catalysis. The position is stabilized by an interaction with His93 (DispTs2) and His94 (DispLp). A water molecule is poised for attack of the anomeric carbon at hydrogen-bonding distance to Glu156 (DispLp; Supplementary Fig. S4). Consistent with the key role of this glutamate, the E184Q variant of DspB lost its functionality (Manuel *et al.*, 2007). The catalytic aspartate, Asp160 in DispTs2 and Asp155 in DispLp, interacts with the NH group of the *N*-acetamido moiety, as required for the mechanism (Figs. 4d and 4e). A second water molecule in the active site is coordinated to the catalytic aspartate and O3 of 6-Ac-Cas. Mutation of the aspartate to an alanine in a GH20  $\beta$ -hexosaminidase from *Streptomyces plicatus* resulted in the observation of the 2-acetamido group in two conformations, with only one of these being viable for catalysis (Williams *et al.*, 2002) and a 13 333-fold reduction in the catalytic efficiency of *p*NP-GlcNAc hydrolysis (Manuel *et al.*, 2007). Glu161 and Asp160 of DispTs2 are  $\sim$ 5.2  $\text{\AA}$  apart, confirming that Glu161 is the general ‘glycosidic’ acid/base in catalysis.

A further three residues form hydrogen bonds to the ligand 6-Ac-Cas to facilitate ligand conformational changes, specifically to stabilize the transition-state conformation (Figs. 4d and 4e). A tyrosine, Tyr247 (DispTs2) and Tyr250 (DispLp), hydrogen-bonds to the oxygen of the *N*-acetyl group. The *N*-acetyl carbonyl group acts as the nucleophile during catalysis and the aspartate and tyrosine residues assist in polar-

izing and orientating the group (Williams *et al.*, 2002). Interestingly, in the structure of unliganded DispTs3 an acetic acid solute molecule was present in a similar position to the *N*-acetyl group of the GlcNAc. The acetic acid also formed hydrogen bonds to Asp160 and Tyr247 with distances of 2.55 and 2.75  $\text{\AA}$ , respectively (Supplementary Fig. S3d). Arg13 of DispTs2 (and likewise Arg17 from DispLp) forms two hydrogen bonds to the C3 and C4 hydroxyls of 6-Ac-Cas with distances of approximately 2.8  $\text{\AA}$ . Previous analysis of the importance of the arginine (Arg27) from DspB in ligand stabilization was analysed by mutating the residue to either alanine or lysine, which reduced the catalytic efficiency of *p*NP-GlcNAc cleavage by 1714-fold and 2400-fold, respectively, compared with the WT DspB enzyme when analysed by absorbance at 405 nm (Manuel *et al.*, 2007). Glu300 of DispTs2 forms two hydrogen bonds to the C4 hydroxyl, at a distance of 2.7  $\text{\AA}$ , and to the C6 hydroxyl on the pyrrole ring, at a distance of 2.75  $\text{\AA}$ . Mutation of the equivalent Glu332 of DspB to glutamine reduced the catalytic efficiency of *p*NP-GlcNAc hydrolysis by 2000-fold compared with the WT (Manuel *et al.*, 2007). Unusually, DispLp has an alanine instead of a glutamate at this position. This substitution could explain the 184-fold reduction in the dissociation constant of DispLp for 6-Ac-Cas compared with DispTs2 and 6-Ac-Cas. Therefore, distortion of 6-Ac-Cas in the active site of DispLp must rely on the interactions with Arg17 and Trp306. Interestingly, there are two waters in the DispLp structure that superpose well with the OE1 and OE2 of glutamate (Glu332 in DspB and Glu300 in DispTs2), one of which is coordinated by Gln252 (corresponding to Leu or Val in the other two dispersins); these waters might compensate for the Glu/Ala substitution (Supplementary Fig. S4).

Aromatic residues in the active site are involved in positioning the ligand correctly in the active site. A tryptophan residue, Trp298 (DispTs2) or Trp306 (DispLp), at the base of the active site provides important  $\pi$ - $\pi$  stacking interactions through alignment of the indolizine rings of the tryptophan and the ligand. A second tryptophan, Trp193 in DispTs3 (Trp193 in DispTs2 and Trp188 in DispLp) is present at the base of the *N*-acetyl group. A third tryptophan, Trp214 in DispTs3 (Trp214 in DispTs2 and Trp209 in DispLp) forms the side hydrophobic pocket in which the *N*-acetyl group is situated; mutation of the corresponding DspB residue, W237A, completely abolished all detectable activity on *p*NP-GlcNAc, suggesting that the hydrophobic pocket is essential to capture the substrate (Manuel *et al.*, 2007).

The active-site pocket of all three dispersin enzymes and DspB is not as deep or enclosed as that of exo-acting GH20 enzymes. For example, the hexosaminidase from *S. plicatus* (*SpHex*), which has only exoglycosidase activity, has two unstructured loops, Thr272–Phe278 and Asp401–Tyr411, that lie on opposite sides of the active site, and which confine the top of the active-site pocket, restricting the enzyme to exo activity only (Mark *et al.*, 2001; Little *et al.*, 2012). In comparison, the cleft in which PNAG would bind to the three dispersins is shallow and could allow both endo and exo activity.

### 3.6. Complex of DispTs2 with GlcNAc- $\beta$ (1,6)-GlcNAc-thiazoline

In order to trap a longer oligosaccharide complex, and building on the known neighbouring-group reaction mechanism, a novel disaccharide was synthesized (initial attempts with GlcNAc-thiazoline alone had not yielded high diffraction-quality crystals). This compound includes an additional  $\beta$ -1,6-linked GlcNAc to the well known GlcNAc-thiazoline, a potent transition-state/intermediate mimic and inhibitor of GH20 and related enzymes (Mark *et al.*, 2001; Macauley *et al.*, 2005; Knapp *et al.*, 1996). However, the design and synthesis, detailed in the supporting information, was performed before we had in-depth knowledge of the  $-2$  subsite requirements of these enzymes.

We therefore conducted soaking experiments with the bespoke  $\beta$ -GlcNAc- $\beta$ (1,6)-GlcNAc-thiazoline (di-NAG-thiazoline; see the supporting information for synthesis details). While these soaks produced crystals with poorer diffraction compared with 6-Ac-Cas, this compound held particular interest due to the additional sugar unit. The electron density for the first unit of di-NAG-thiazoline (corresponding to GlcNAc-thiazoline, NGT in the PDB dictionary) was well defined, reflecting its mimicry of the reaction intermediate, but only disordered density was observed for the  $-2$  subsite. As subsequently discovered, the  $-2$  subsite preferentially accommodates GlcN rather than GlcNAc, which likely contributed to the observed disorder of the GlcNAc moiety in this complex (Fig. 5).

### 3.7. The dispersins have signature conserved regions despite low sequence identities

Having demonstrated that these enzymes were all hexosaminidases active on PNAG and obtained the crystal structures, we next sought to analyse any sequence features that were conserved amongst the dispersins and to map them onto the 3D structure to aid future dispersin categorization. All of the putative dispersins were not previously members of CAZy family GH20 (Lombard *et al.*, 2014); therefore, a sequence alignment with DispB, a single-domain  $\beta$ -1,4 N-acetyl-glucosaminidase (StrH) from *S. pneumoniae* TIGR4 and the representative multi-domain ScHexA was performed (Fig. 6).

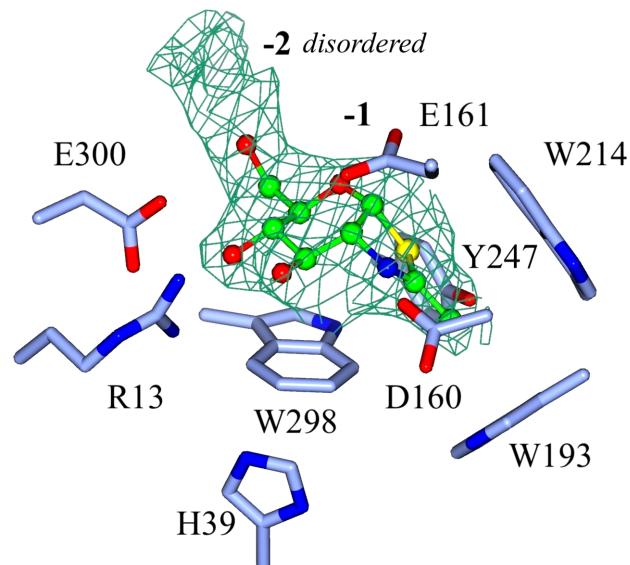
The regions of high sequence conservation are situated within the active site and on the top face of the enzyme, and 14 residues (Fig. 7a) that are conserved across both GH20 single-domain and multi-domain enzymes are located facing inwards towards the centre of the barrel (Fig. 7b). Of these 14, eight residues are conserved across all GH20 enzymes analysed and six residues are only conserved across the dispersin subfamily (His53, Asp116, Trp216, Asp218, Trp330 and Gly331 of DispB; Fig. 7c). The *N*-acetyl group of GlcNAc in the  $-1$  subsite is surrounded by three tryptophan residues that form a compact hydrophobic pocket. Tyr237, which is located at the side of the active-site pocket against the *N*-acetyl group, is conserved across all GH20 enzymes; however, Trp330 and Trp216 are specifically conserved in all dispersin enzymes. His53 is located

at the base of the active site perpendicular to Trp330. Asp116 is at hydrogen-bonding distance from the catalytic residue Asp184, Asp218 is at hydrogen-bonding distance from Tyr237, and Gly331 is found between Trp330 and Glu332, which make important ligand interactions. These conserved residues in the dispersin subfamily are important for positioning and stabilizing key catalytic residues.

In the structure of DispTs2, Glu300 (Glu332 in DispB) forms important ligand interactions with the C4 and C6 hydroxyls of the pyrrole ring to stabilize the ligand conformational changes during catalysis. This residue is conserved in all dispersins apart from DispLp (Ala308) and the equivalent residue is a glycine in StrH. Therefore, as well as implications for its catalytic efficiency attributed to a loss in hydrogen-bonding capacity, DispLp might be able to accept a  $\beta$ -1,4-linked substrate since there would not be any steric clashes from the 4-position, with the groove now 3 Å larger.

A further 11 residues are conserved throughout the dispersin subfamily. Gly20 is located in the central  $\beta$ -barrel ( $\beta$ 1); Ser64 (loop between  $\beta$ 2 and  $\beta$ 3) and Glu166 ( $\alpha$ 5) form a hydrogen bond; Ala102 ( $\alpha$ 2) is a surface residue; Phe171 ( $\alpha$ 5) forms stacking interactions against Pro113 ( $\beta$ 6), which is conserved across the GH20 family; Asn217 (the loop after  $\beta$ 8) is located between Tyr216 and Asp218 that play important roles in the active site and Tyr236 ( $\beta$ 9) is located next to Asn217 in the structure; Asn271, Asn273 and Tyr275 (the loop between  $\beta$ 10 and  $\alpha$ 9) stabilize the loop region between  $\alpha$ 10 and  $\beta$ 12; and Asp290 ( $\alpha$ 10) is a surface residue.

The catalytic motif for GH20 enzymes, required for their NGP catalytic mechanism, consists of a catalytic aspartate and



**Figure 5**  
Structure of the active site of the complex of DispTs2 with di-NAG-thiazoline. The second GlcNAc (not shown) is not well defined, most probably because the  $-2$  subsite preferentially accommodates GlcN rather than GlcNAc. The  $-1$  and  $-2$  subsites are shown in bold. The maximum-likelihood/ $\sigma_A$ -weighted  $2F_{\text{obs}} - F_{\text{calc}}$  map is shown in green contoured at  $0.16 \text{ e } \text{\AA}^{-3}$ . We did not create a new ligand library for this case because of poor density fit of the second unit; NAG-thiazoline only (NGT in the PDB ligand library) was modelled into the structure, shown in green.

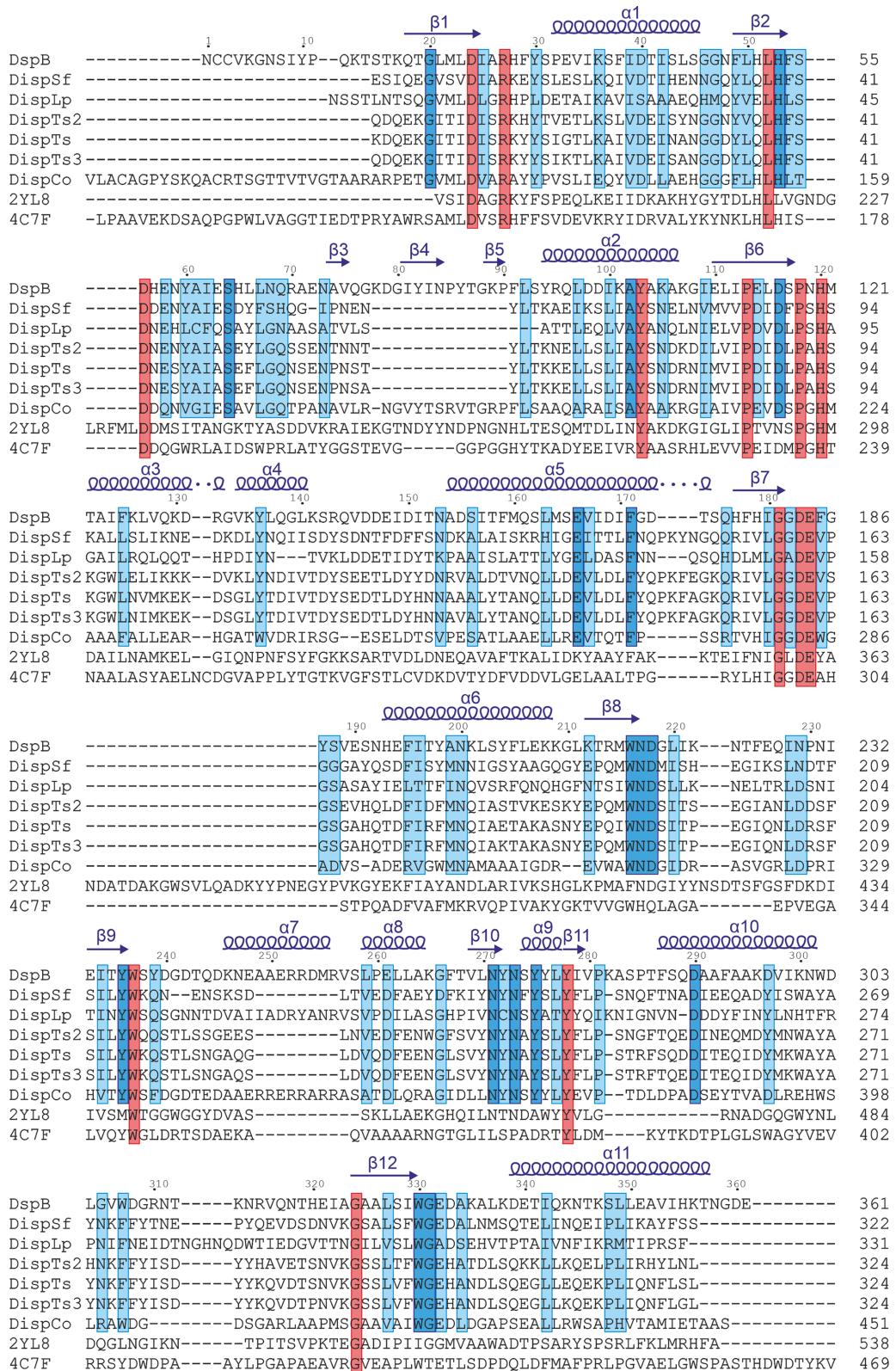
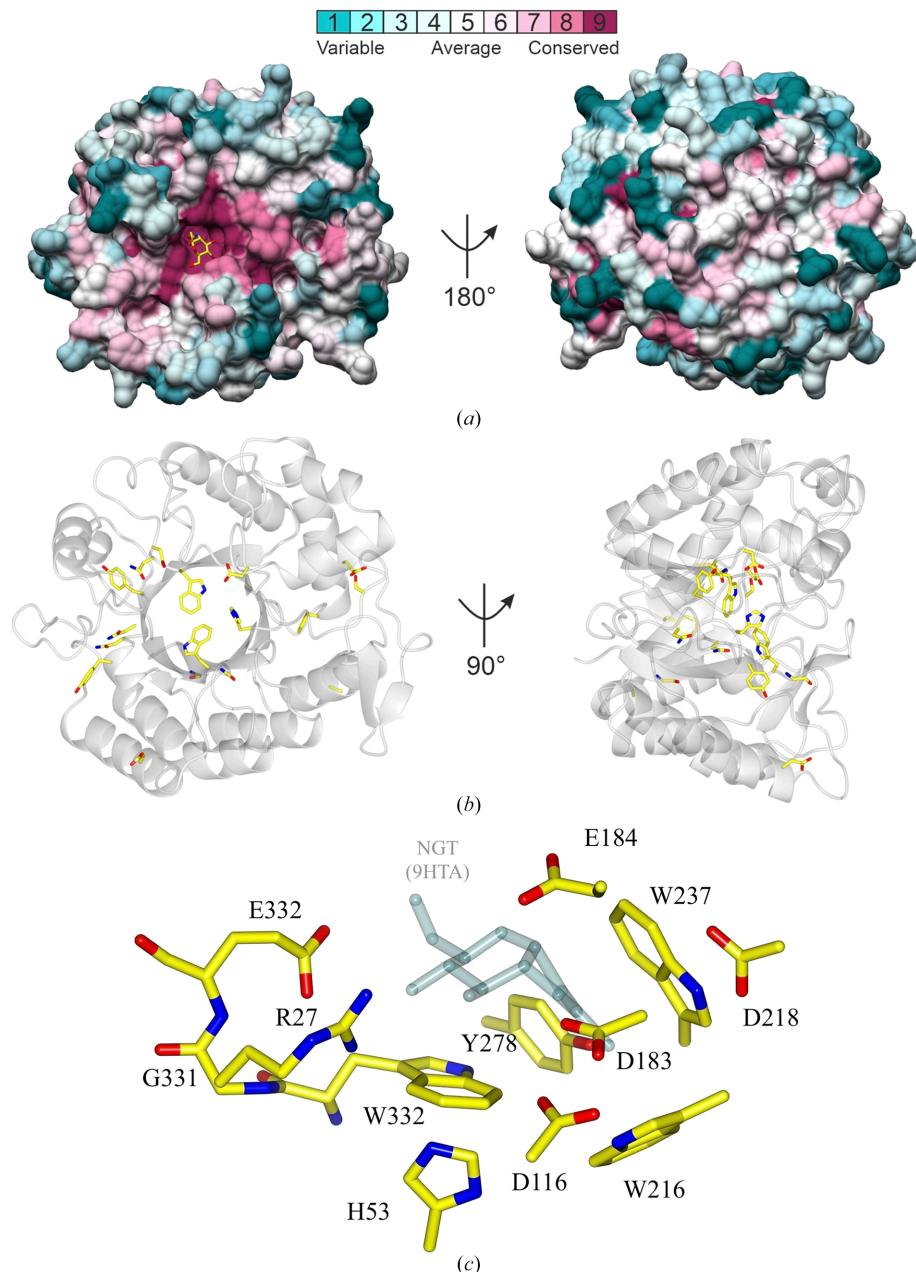


Figure 6

Sequence alignment of the PNAG catalytic domain of GH20 family members: DspB, the newly characterized dispersins, StrH (PDB entry 2yl8) and ScHexA (PDB entry 4c7f). Residues with a red background are conserved across all GH20 proteins. Residues with a dark blue background are conserved across all dispersins. Residues with 70% conservation across dispersins are highlighted in a light blue box. The numbering and the secondary-structure elements across the top of the alignment correspond to the sequence and fold of DspB:  $\alpha$ ,  $\alpha$ -helix;  $\beta$ ,  $\beta$ -strand. For DispCo, the N-terminal domain was omitted; the first residue included was Val102. In PDB entry 2yl8, the catalytic residue Glu361 of StrH is mutated to a glutamine; only the GH20 domain from residues 190 to 538 was used in the alignment. For ScHexA (PDB entry 4c7f), the GH20 domain between residues 153 and 535 (the C-terminus) was used in the alignment. Domain boundaries were predetermined (Val-Cid *et al.* 2015).

**Figure 7**

Conserved residues in GH20 enzymes are clustered in the active site and on the top face. (a) Surface representation of DspB (PDB entry 1yht) in white with residues coloured according to the degree of sequence conservation using the aligned proteins from Fig. 1. GlcNAc, in yellow, was modelled into the active site by superimposing DspB (PDB entry 1yht) with a  $\beta$ -N-acetylhexosaminidase from *Akkermansia muciniphila* (PDB entry 7cbo; Xu *et al.*, 2020). This figure was produced using the *ConSurf* server (Landau *et al.*, 2005; Ashkenazy *et al.*, 2016). (b) Ribbon representation of DspB (PDB entry 1yht) with conserved residues specific to the dispersin subfamily members (not conserved in other GH20 enzymes) highlighted in yellow. (c) A close-up view of the dispersin active site in the example of DspB (PDB entry 1yht), with key residues numbered for DspB. NAG-thiazoline (NGT) from the DispTs2-di-NAG-thiazoline complex structure (PDB entry 9hta), in semi-transparent grey, is shown to indicate the ligand-binding site.

glutamate. In DspB, DispCo and ScHexA, the DE motif is preceded by HXGG(DE), whereas StrH contains the sequence NIGLDE. DispSf, DispTs2, DispTs and DispTs3 have the sequence VLGGDE and DispLp has the sequence MLGADE (Fig. 6). Hence, there is no consistent sequence motif requirement for dispersin catalytic sites. The two main regions of sequence conservation between dispersins only are Trp216–Trp218 and Asn271–Tyr275, which are important in catalysis and loop stabilization, respectively.

#### 4. Discussion

Under certain stresses and signals, microorganisms use different mechanisms to break the extracellular matrix for cell dispersion. Depending on the composition of the EPS, proteases, DNases and PNAGases are responsible for the release. Several new PNAG-cleaving dispersins have been identified and characterized in this study. The five novel dispersins examined were identified amongst taxonomically

well separated bacterial genera, which could point to a common ancestral source. However, it cannot be excluded that these enzymes evolved independently on multiple occasions. Interestingly, there was no clear link to a specific ecological niche as the bacterial hosts originated from diverse sources. This could suggest that the activity is linked to a more fundamental microbial characteristic, such as dispersal (Penesyan *et al.*, 2021).

Around 50 additional enzymes were identified through a *BLAST* search based on sequence similarity to DspB. Further analysis of these enzymes, using the assays described in this paper, could expand the dispersin subfamily. The location of the DNA sequences compared with the sequences involved in forming the PNAG biosynthetic machinery could be important for further verification. Genes within the *pga/ica* operons could be identified in *T. saccharophilus* (*pgaC* and *pgaD*), *M. fleurettii* (*icaB*, *icaD* and *pgaC*) and *L. paraplanatum* (*pgaC*); however, neither the *pga* nor *ica* gene operons could be identified in *C. oceanosedimentum*. It cannot be excluded that this species, and the other species with only select genes in the cluster assigned, carry unassigned operons related to PNAG production. These enzymes could also be targeting other types of extracellular polysaccharides produced by the host cells. Another intriguing explanation could be that these enzymes impose a competitive or cooperative advantage in a polymicrobial environment. Bacterial species are known to interact actively in sessile communities (Burmølle *et al.*, 2014), forming predatory or symbiotic relationships depending on the species composition. It could be speculated that these PNAG-degrading enzymes, found in microorganisms without any apparent genes coding for PNAG production, could in fact serve as dispersal agents in multispecies communities where other PNAG producers are present, enabling the dispersin-producing microorganisms to compete or corroborate with their neighbours. It would have been of interest to carry out further studies on DispCo with the aim of obtaining a sample including the N-terminal domain allowing crystallization and kinetic measurements, but this was beyond the scope of this project. *AlphaFold2* modelling of DispCo suggested that its N-terminal domain was fibronectin-like.

Obtaining the structure of new dispersins in complex with a ligand confirmed the substrate-assisted catalysis mechanism and revealed important residues involved in catalysis through positioning the substrate for catalysis and stabilizing the conformational changes along the reaction coordinate of the enzyme. The inhibitor 6-Ac-Cas is supported by a hydrogen-bonding network between residues in the active site of DispTs2 and DispLp and the dispersin-specific tryptophan residues which form the base of the active site.

The GH20 family contains enzymes that cleave a variety of different substrates, although these enzymes have high sequence identity. A small number of enzymes are chitobiases, which cleave the  $\beta$ -1,4-GlcNAc linkage of chitin, and lacto-*N*-biosidases, which cleave the  $\beta$ -1,3-linkage between GlcNAc and galactose (Tews *et al.*, 1996). The predominant type of enzymes in this family are  $\beta$ -hexosaminidases, acting on both *N*-acetylglucosamine and *N*-acetylgalactosamine. Primarily,

these enzymes are not known to cleave PNAG, while a broader range of these enzymes cleave *p*NP-GlcNAc. All five of the dispersins tested showed measurable activity on both small-molecule substrates, *p*NP-GlcNAc and 4-MU-GlcNAc, with DispTs2 as the most active on both substrates and DispLp as the least active on both substrates. Most notably, DspB is the dispersin that shows the largest difference in specificity towards the two substrates. This may be attributed to the small differences between the two substrates, where *p*NP-GlcNAc is the smaller substrate and is able to display a small partial charge from the resonance structure of the  $-NO_2$  group. Conversely, 4-MU-GlcNAc is slightly larger and displays a larger fused aromatic structure which is more polarizable. Recently, a new fluorogenic substrate has been developed which includes a carbamate linker between the GlcNAc and the fluorophore, 7-amino-4-methylcoumarin, to increase the distance and allow efficient hydrolysis (Wang *et al.*, 2022).

In conclusion, the activity of DispTs2 is comparable to that of DspB. Supplementary evidence for the location of the various subsites, the preference of these subsites for GlcNAc or GlcN from the surrounding residues, for example Asp147, Asp245 and Glu248 in DspB, and the ability of the dispersins to be both exo and endo acting require a crystal structure with a complex of a dispersin and a polysaccharide ligand. Moreover, the differences in the secondary structure that we observed between the different dispersins might have an impact on their exo/endo-acting propensities, which could be an interesting topic for future studies. Further research is also needed into the organization of the PNAG biosynthetic machinery in the cell envelope, in reference to the PgaABCD and IcaABCD enzymes, and the association of PNAG with itself and other components on the cell surface. Recently, a general acid/base, GFP-tagged mutant of DspB was used as a probe to detect PNAG oligomers in high-density and isolated regions, PNAG islands, on the periphery of the cell during the early log phase and extending between bacteria as the point of contact (Eddenden *et al.*, 2020).

It is hoped that the enzymes discussed in this paper will act as an alternative for using DspB to elucidate these questions. Whether they have superior function under alternate conditions or alternate immunoreactivity will need to be established. This work paves the way for the unearthing of additional dispersin enzymes.

## 5. Related literature

The following references are cited in the supporting information for this article: Fulmer *et al.* (2010), Jiang *et al.* (2004), Maiti *et al.* (2007) and Reynolds & Evans (1942).

## Acknowledgements

We thank Diamond Light Source for access to beamlines I03 and I04 (proposal No. mx-13587), which contributed to the results presented here. We also thank Johan and Sam from the YSBL for collecting the data. We thank the Technology

Facility for their assistance in collecting the mass-spectrometry data of the synthetic PNAG hydrolysis products.

### Conflict of interest

Conflict of interest statement: Novonesis is a commercial enzyme supplier who sell enzymes for polysaccharide degradation.

### Funding information

GJD is funded by the Royal Society Ken Murray Research Professorship. AM was funded in York by BBSRC grant BB/T004819/1. MN, AE and DJV thank the Natural Sciences and Engineering Research Council of Canada (NSERC), grant Nos RGPIN-06334 and RGPIN-06466, for support.

### References

Adams, P. D., Afonine, P. V., Bunkóczki, 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). *Methods*, **55**, 94–106.

Agirre, J., Atanasova, M., Bagdonas, H., Ballard, C. B., Baslé, A., Beilsten-Edmands, J., Borges, R. J., Brown, D. G., Burgos-Mármol, J. J., Berrisford, J. M., Bond, P. S., Caballero, I., Catapano, L., Chojnowski, G., Cook, A. G., Cowtan, K. D., Croll, T. I., Debreczeni, J. É., Devenish, N. E., Dodson, E. J., Drevon, T. R., Emsley, P., Evans, G., Evans, P. R., Fando, M., Foadi, J., Fuentes-Montero, L., Garman, E. F., Gerstel, M., Gildea, R. J., Hatti, K., Hekkelman, M. L., Heuser, P., Hoh, S. W., Hough, M. A., Jenkins, H. T., Jiménez, E., Joosten, R. P., Keegan, R. M., Keep, N., Krissinel, E. B., Kolenko, P., Kovalevskiy, O., Lamzin, V. S., Lawson, D. M., Lebedev, A. A., Leslie, A. G. W., Lohkamp, B., Long, F., Malý, M., McCoy, A. J., McNicholas, S. J., Medina, A., Millán, C., Murray, J. W., Murshudov, G. N., Nicholls, R. A., Noble, M. E. M., Oeffner, R., Pannu, N. S., Parkhurst, J. M., Pearce, N., Pereira, J., Perrakis, A., Powell, H. R., Read, R. J., Rigden, D. J., Rochira, W., Sammito, M., Sánchez Rodríguez, F., Sheldrick, G. M., Shelley, K. L., Simkovic, F., Simpkin, A. J., Skubak, P., Sobolev, E., Steiner, R. A., Stevenson, K., Tews, I., Thomas, J. M. H., Thorn, A., Valls, J. T., Uski, V., Usón, I., Vagin, A., Velankar, S., Vollmar, M., Walden, H., Waterman, D., Wilson, K. S., Winn, M. D., Winter, G., Wojdyr, M. & Yamashita, K. (2023). *Acta Cryst. D79*, 449–461.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. & Ben-Tal, N. (2016). *Nucleic Acids Res.* **44**, W344–W350.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). *Nature*, **255**, 609–614.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). *J. Mol. Biol.* **340**, 783–795.

Breslawec, A. P., Wang, S., Li, C. & Poulin, M. B. (2021). *J. Biol. Chem.* **296**, 100203.

Bundalovic-Torma, C., Whitfield, G. B., Marmont, L. S., Howell, P. L. & Parkinson, J. (2020). *PLoS Comput. Biol.* **16**, e1007721.

Burmølle, M., Ren, D., Bjarnsholt, T. & Sørensen, S. J. (2014). *Trends Microbiol.* **22**, 84–91.

Calvelo, M., Males, A., Alteen, M. G., Willems, L. I., Vocadlo, D. J., Davies, G. J. & Rovira, C. (2023). *ACS Catal.* **13**, 13672–13678.

Ceri, H., Olson, M. E., Stremick, C., Read, R. R., Morck, D. & Buret, A. (1999). *J. Clin. Microbiol.* **37**, 1771–1776.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst. D66*, 12–21.

Chiba, A., Sugimoto, S., Sato, F., Hori, S. & Mizunoe, Y. (2015). *Microb. Biotechnol.* **8**, 392–403.

Cowtan, K. (2006). *Acta Cryst. D62*, 1002–1011.

Cramton, S. E., Gerke, C., Schnell, N. F., Nichols, W. W. & Götz, F. (1999). *Infect. Immun.* **67**, 5427–5433.

Cywes-Bentley, C., Skurnik, D., Zaidi, T., Roux, D., DeOliveira, R. B., Garrett, W. S., Lu, X., O'Malley, J., Kinzel, K., Zaidi, T., Rey, A., Perrin, C., Fichorova, R. N., Katayani, A. K., Maira-Litrán, T., Gening, M. L., Tsvetkov, Y. E., Nifantiev, N. E., Bakaletz, L. O., Pelton, S. I., Golenbock, D. T. & Pier, G. B. (2013). *Proc. Natl Acad. Sci. USA*, **110**, e2209.

D'Arcy, A., Bergfors, T., Cowan-Jacob, S. W. & Marsh, M. (2014). *Acta Cryst. F70*, 1117–1126.

Donlan, R. M. (2002). *Emerg. Infect. Dis.* **8**, 881–890.

Drouillard, S., Armand, S., Davies, G. J., Vorgias, C. E. & Henrissat, B. (1997). *Biochem. J.* **328**, 945–949.

Drula, E., Garron, M. L., Dogan, S., Lombard, V., Henrissat, B. & Terrapon, N. (2022). *Nucleic Acids Res.* **50**, D571–D577.

Eddenden, A., Kitova, E. N., Klassen, J. S. & Nitz, M. (2020). *ACS Chem. Biol.* **15**, 1204–1211.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst. D66*, 486–501.

Fulmer, G. R., Miller, A. J. M., Sherden, N. H., Gottlieb, H. E., Nudelman, A., Stoltz, B. M., Bercaw, J. E. & Goldberg, K. I. (2010). *Organometallics*, **29**, 2176–2179.

Guilhen, C., Forestier, C. & Balestrino, D. (2017). *Mol. Microbiol.* **105**, 188–210.

Hattie, M., Ito, T., Debowski, A. W., Arakawa, T., Katayama, T., Yamamoto, K., Fushinobu, S. & Stubbs, K. A. (2015). *Chem. Commun.* **51**, 15008–15011.

Izano, E. A., Sadovskaya, I., Vinogradov, E., Mulks, M. H., Velliyagounder, K., Ragunath, C., Kher, W. B., Ramasubbu, N., Jabbouri, S., Perry, M. B. & Kaplan, J. B. (2007). *Microb. Pathog.* **43**, 1–9.

Jamal, M., Ahmad, W., Andleeb, S., Jalil, F., Imran, M., Nawaz, M. A., Hussain, T., Ali, M., Rafiq, M. & Kamil, M. A. (2018). *J. Chin. Med. Assoc.* **81**, 7–11.

Jiang, Z.-H., Gandhi, S. & Koganty, R. R. (2004). World Patent WO2004028475A2.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikолов, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.

Kaplan, J. B., Meyenhofer, M. F. & Fine, D. H. (2003). *J. Bacteriol.* **185**, 1399–1404.

Kaplan, J. B., Ragunath, C., Ramasubbu, N. & Fine, D. H. (2003). *J. Bacteriol.* **185**, 4693–4698.

Kaplan, J. B., Velliyagounder, K., Ragunath, C., Rohde, H., Mack, D., Knobloch, J. K. & Ramasubbu, N. (2004). *J. Bacteriol.* **186**, 8213–8220.

Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.

Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst. D63*, 447–457.

Knapp, S., Vocadlo, D., Gao, Z. N., Kirk, B., Lou, J. P. & Withers, S. G. (1996). *J. Am. Chem. Soc.* **118**, 6804–6805.

Krissinel, E. (2012). *J. Mol. Biochem.* **1**, 76–85.

Krissinel, E. & Henrick, K. (2004). *Acta Cryst. D60*, 2256–2268.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. (2018). *Mol. Biol. Evol.* **35**, 1547–1549.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005). *Nucleic Acids Res.* **33**, W299–W302.

Leung, C., Chibba, A., Gómez-Biagi, R. F. & Nitz, M. (2009). *Carbohydr. Res.* **344**, 570–575.

Lillard, J. W. Jr, Fetherston, J. D., Pedersen, L., Pendrak, M. L. & Perry, R. D. (1997). *Gene*, **193**, 13–21.

Little, D. J., Pföh, R., Le Mauff, F., Bamford, N. C., Notte, C., Baker, P., Guragain, M., Robinson, H., Pier, G. B., Nitz, M., Deora, R., Sheppard, D. C. & Howell, P. L. (2018). *PLoS Pathog.* **14**, e1006998.

Little, D. J., Poloczek, J., Whitney, J. C., Robinson, H., Nitz, M. & Howell, P. L. (2012). *J. Biol. Chem.* **287**, 31126–31137.

Liu, P. S., Kang, M. S. & Sunkara, P. S. (1991). *Tetrahedron Lett.* **32**, 719–720.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). *Nucleic Acids Res.* **42**, D490–D495.

Macauley, M. S., He, Y., Gloster, T. M., Stubbs, K. A., Davies, G. J. & Vocadlo, D. J. (2010). *Chem. Biol.* **17**, 937–948.

Macauley, M. S., Whitworth, G. E., Debowski, A. W., Chin, D. & Vocadlo, D. J. (2005). *J. Biol. Chem.* **280**, 25313–25322.

Mack, D., Fischer, W., Krokotsch, A., Leopold, K., Hartmann, R., Egge, H. & Laufs, R. (1996). *J. Bacteriol.* **178**, 175–183.

Maiti, K. K., Lee, W. S., Takeuchi, T., Watkins, C., Fretz, M., Kim, D. C., Futaki, S., Jones, A., Kim, K. T. & Chung, S. K. (2007). *Angew. Chem. Int. Ed.* **46**, 5880–5884.

Manuel, S. G., Ragunath, C., Sait, H. B., Izano, E. A., Kaplan, J. B. & Ramasubbu, N. (2007). *FEBS J.* **274**, 5987–5999.

Mark, B. L., Vocadlo, D. J., Knapp, S., Triggs-Raine, B. L., Withers, S. G. & James, M. N. G. (2001). *J. Biol. Chem.* **276**, 10330–10337.

Meekrathok, P., Stubbs, K. A. & Suginta, W. (2018). *Int. J. Biol. Macromol.* **115**, 1165–1173.

Moroz, O. V., Blagova, E., Lebedev, A. A., Nørgaard, A., Segura, D. R., Blicher, T. H., Brask, J. & Wilson, K. S. (2017). *Acta Cryst. D* **73**, 32–44.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst. D* **67**, 355–367.

Penesyan, A., Paulsen, I. T., Kjelleberg, S. & Gillings, M. R. (2021). *npj Biofilms Microbiomes* **7**, 80.

Ramasubbu, N., Thomas, L. M., Ragunath, C. & Kaplan, J. B. (2005). *J. Mol. Biol.* **349**, 475–486.

Reynolds, D. D. & Evans, W. L. (1942). *Org. Synth.* **22**, 56–58.

Roux, D., Cywes-Bentley, C., Zhang, Y. F., Pons, S., Konkol, M., Kearns, D. B., Little, D. J., Howell, P. L., Skurnik, D. & Pier, G. B. (2015). *J. Biol. Chem.* **290**, 19261–19272.

Seijsing, F., Nilebäck, L., Öhman, O., Pasupuleti, R., Ståhl, C., Seijsing, J. & Hedhammar, M. (2020). *MicrobiologyOpen* **9**, e993.

Shah, A. K., Liu, Z.-J., Stewart, P. D., Schubot, F. D., Rose, J. P., Newton, M. G. & Wang, B.-C. (2005). *Acta Cryst. D* **61**, 123–129.

Shaw Stewart, P. D., Kolek, S. A., Briggs, R. A., Chayen, N. E. & Baldock, P. F. M. (2011). *Cryst. Growth Des.* **11**, 3432–3441.

Sumida, T., Stubbs, K. A., Ito, M. & Yokoyama, S. (2012). *Org. Biomol. Chem.* **10**, 2607–2612.

Tan, Y., Ma, S., Liu, C., Yu, W. & Han, F. (2015). *Microbiol. Res.* **178**, 35–41.

Tews, I., Perrakis, A., Oppenheim, A., Dauter, Z., Wilson, K. S. & Vorgias, C. E. (1996). *Nat. Struct. Mol. Biol.* **3**, 638–648.

Thi, N. N., Offen, W. A., Shareck, F., Davies, G. J. & Doucet, N. (2014). *Biochemistry*, **53**, 1789–1800.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst. D* **66**, 22–25.

Val-Cid, C., Biarnés, X., Fajíes, M. & Planas, A. (2015). *PLoS One*, **10**, e0128075.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D. & Velankar, S. (2024). *Nucleic Acids Res.* **52**, D368–D375.

Wang, S., Breslawec, A. P., Alvarez, E., Tyrlik, M., Li, C. & Poulin, M. B. (2019). *ACS Chem. Biol.* **14**, 1998–2005.

Wang, S., Breslawec, A. P., Li, C. & Poulin, M. B. (2020). *Chemistry*, **26**, 10719–10723.

Wang, S., Breslawec, A. P. & Poulin, M. B. (2022). *Bioorg. Chem.* **119**, 105532.

Wang, X., Preston, J. F. & Romeo, T. (2004). *J. Bacteriol.* **186**, 2724–2734.

Williams, S. J., Mark, B. L., Vocadlo, D. J., James, M. N. & Withers, S. G. (2002). *J. Biol. Chem.* **277**, 40055–40065.

Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.

Xu, W., Yang, W., Wang, Y., Wang, M. & Zhang, M. (2020). *Biochem. Biophys. Res. Commun.* **529**, 876–881.

Yakandawala, N., Gawande, P. V., LoVetri, K., Romeo, T., Kaplan, J. B. & Madhyastha, S. (2009). *J. Ind. Microbiol. Biotechnol.* **36**, 1297–1305.

Yin, W., Wang, Y., Liu, L. & He, J. (2019). *Int. J. Mol. Sci.* **20**, 3423.