Molecular warfare

A structural biology view on pathogen weapons — GbpA from *Vibrio cholerae* — and host defenses — Vg from the honey bee

Mateu Montserrat-Canals

Thesis for the degree of *Philosophiae Doctor*

Norwegian Centre for Molecular Medicine / Department of Chemistry
Faculty of Mathematics and Natural Sciences
University of Oslo

Spring 2020 – Summer 2023

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2660

ISSN 1501-7710

Cover: UiO.
Print production: Graphic center, University of Oslo.

Acknowledgements

The research work of this thesis was carried out between February 2020 and June 2023, the work on GbpA started from September 2021. The work was performed at the University of Oslo (UiO), initially at the Norwegian Center for Molecular Medicine (NCMM) with a subsequent change in supervision and location to the Department of Chemistry in September 2021. This PhD was originally funded by the PROTON ITN as part the Marie Skołodowska-Curie actions with subsequent core funding from NCMM. I am grateful to NCMM for its support.

I am extremely grateful to my supervisor, professor Ute Krengel, for her dedication and commitment to me as a student, not only in the academic but also in the personal sphere. Your support has been key to the completion of this PhD. Thanks as well to professor Hartmut Luecke for giving me the opportunity to move to Oslo for my PhD. Thanks to the rest of my co-supervisors: professor Reidar Lund and professor Michelle Cascella, their expertise has not been needed given the direction the project has taken but their support is appreciated.

I would like to thank researchers at facilities who have assisted me in performing experiments and providing their help and expertise which have contributed so much to this thesis. From the ESRF special thanks to Philippe Carpentier as well as Gordon Leonard. From UiO, I also received help from Norbert Roos from the EM facilities at IBV as well as Alejandro Barrantes Bautista from the department of Clinical Dentistry. I am also thankful to Esko Oksanen for his visit to Oslo and his experienced input into the crystal optimization for neutron studies. Finally, I want to thank Zöe Fisher at the DEMAX facilities (Lund, Sweden) for her work on producing deuterated protein, going the extra mile to ensure sample quality.

Key to enduring the most stressful and difficult parts of my PhD has been having a partner to go along with during the whole process. Many thanks to Flore for her help, support and friendship. Special thanks to Gabriele Cordara, for his uncountable contributions to my research and being an inexhaustible source of knowledge about anything remotely related to structural biology. Thanks to Abelone and, in particular, Henrik for setting the basis of the work on GbpA and introducing me to it. I am thankful to the rest of the members of the Krengel group and associated people who have been around during my time here, thank you Tamjidmaa, Natalia, Eirik, Ayla, Mari, Lin and Aysu for creating such a nice environment. Thanks to Tamjidmaa for her help with remote data collection. It has been a pleasure to work with Eirik and Ayla and contribute to their master projects. Thanks to Ayla for the abstract translation.

Thanks as well to the members of the old Luecke group. To Eva for her continued support with the vitellogenin project. Thanks to Joel for his initial co-supervision work. To Javi and Marta for their persistent humor and “optimism” and to Rasma for her kindness and willingness to help.

I am grateful to my collaborators on the vitellogenin project. Thanks to Vilde for her enthusiasm and support as well as to professor Gro Amdam and professor Øyvind Halskau. Many thanks to Kilian for refloating the vitellogenin project and for kindly hosting me during my visit to Osnabrück together with professor Arne Moeller.

Thanks to Biocat PhD school for the countless opportunities that it has provided through its funding to allow me to develop as a researcher. Thanks as well to the IBA PhD school.

Finally, I would like to express massive amounts of love and gratitude to all my friends and family that make life such a beautiful thing. Thanks to Hannah, for reading over my thesis and providing some much needed corrections and to Matylda for her input in graphic design.

Table of Contents

Abstract

This thesis focuses on the molecular mechanisms of two protein targets involved in pathogenesis and immunity, providing a glimpse into the infection mechanisms of pathogenic bacteria as well as some of the defenses displayed by host organisms. *N*-acetylglucosamine (GlcNAc) binding protein A (GbpA) from the cholera-causing bacteria *Vibrio cholerae* promotes colonization of the host by binding to glycans in the intestine while increasing their production in the host. In addition, GbpA is important for the survival of the pathogen in the environment by providing both attachment to chitin as well as a nutrient source through the degradation of the semicrystalline polysaccharide thanks to its lytic polysaccharide monooxygenase (LPMO) activity. Representing the host defense, vitellogenin (Vg) is the main lipoprotein precursor in egg yolk and also displays a range of functions related to immunity. Vg can bind pathogen-associated molecular patterns (PAMPs), opsonize bacteria for phagocytosis, has bactericidal and antiviral activity and has been shown to be involved in trans-generational immune priming. Despite the abundance of functional data, both regarding GbpA but especially for Vg, relatively little is known about the molecular mechanisms underlying their functions. For GbpA not much is known about the dynamics and mechanisms involved in binding chitin and other GlcNAc saccharides. In addition, much is still unknown regarding its enzymatic mechanisms as an LPMO. LPMOs have only been discovered recently, but research on these enzymes is highly relevant given their applications in the recycling of biomass that contains crystalline polysaccharides. In the case of Vg, the structure of the full-length protein is unknown, hampering our abilities to link functional data to the structural elements responsible for it.

Here, methods for the optimized production of GbpA samples suitable for structural analysis are reported. General improvements in protein yield and purity were obtained by using the Vmax™ X2 expression system. In addition, perdeuterated protein was produced. Perdeuteration was achieved in our lab and also scaled up at international facilities, setting the stage for future neutron crystallography experiments and enabling small-angle neutron scattering (SANS) experiments to study GbpA binding to chitin. Studies on GbpA binding to chitin were complemented with negative-stain electron microscopy (EM) analysis, showing how chitin fibers are decorated with GbpA. These results form the basis of future cryogenic electron tomography (cryo-ET) studies to obtain higher-resolution information of the interaction. Of relevance for the binding of GbpA to chitin is the discovery of a new metal-binding site that is part of the carbohydrate-binding surface of GbpA. Structural analysis of homologs, together with mutagenesis and stability studies, have provided clues into the significance of this newly identified metal-binding site. Furthermore, in order to better characterize the LPMO catalytic mechanism of GbpA, X-ray crystallography studies were performed to understand the interaction of oxygen species with the active site. The preliminary results presented here provide a foundation for further experiments using both X-ray and neutron crystallography.

Structural work on Vg form the honey bee is also reported, which was purified directly from the hemolymph of the insect. An initial biophysical characterization accompanying a full-length structure prediction was performed. In addition, the first cryogenic electron microscopy (cryo-EM) structure of the full-length Vg at high resolution (3.2 Å) is reported. Previously unmapped glycosylation and metal-binding sites were identified. The structure provided information on the von Willebrand factor type D (vWFD) domain, previously uncharacterized for any member of the Vg superfamily. In addition, structural alignment based on the prediction generated using Alphafold of the only domain not observed experimentally allowed its identification as a C-terminal cystine knot (CTCK) domain.

Sammendrag

Fokuset til denne oppgaven er den molekylære mekanismen til to forskjellige proteiner, hvorav en er involvert i patogenese og den andre er involvert i immunitet. De gir et innblikk i bakterielle infeksjonsmekanismer, samt noen av forsvarsmekanismene til vertsorganismer. Det første proteinet kommer fra kolera-fremkallende bakteriet Vibrio cholerae*. N-acetylglukosamin (GlcNAc)-bindende protein A (GbpA) promoterer koloniseringen av vertsorganismer ved å binde til glykaner i tarmen, samtidig som den oppregulerer utrykket til disse glykanene hos verten. GbpA er også viktig for overlevelsen av* Vibrio cholerae *i vandige miljø, ettersom den gir bakteriene et festepunkt til kitin. I tillegg fungerer kitin som en primær næringskilde til V. cholerae, takket lytiske polysakkaridmonooksygenase (LPMO)-aktiviteten til GbpA. Det andre proteinet vitellogenin (Vg) er hovedforløperen til lipoproteiner i eggeplommer, og demonstrerer et spekter av andre funksjoner relatert til immunitet. Vg har bakteriedrepende og antivirale funksjoner, og har påvist involvering i transgenerasjonell immunpriming. Vg kan også binde til patogenassosiert molekylære mønstre (PAMPs) og opsonisere bakterier for fagocytose. Til tross for rikelige mengder funksjonell data for både GbpA og spesielt Vg, relativt lite er kjent angående de molekylære mekanismene som underligger funksjonen til begge proteinene. For GbpA er ikke mye kjent om dynamikken og mekanismene for binding til kitin og andre GlcNAc-sakkarider. Mye er også ukjent i hensyn til de enzymatiske mekanismene til GbpA som et LPMO. Oppdagelsen av LPMOer er relativt nytt, og videre forskning er relevant grunnet deres anvendelighet i resirkuleringen av krystallinske polysakkarider for biodrivstoffindustrien. Strukturen til Vg er ukjent. Dette forhindrer oss i å koble proteinets funksjonelle data til de strukturelle elementene som er ansvarlig for det.*

Her rapporterer vi metoder for optimalisert produksjon av GbpA-prøver som er egnet til strukturanalyse. Generelle forbedringer i proteinutbytte og renhet var oppnådd med et Vmax™ X2 utrykkssystem. Perdeuterasjon av GbpA var oppnådd på vårt lab og oppskalert på internasjonale fasiliteter, noe som setter cenen for fremtidige nøytronkrystallografiske og nøytron-småvinkelspredningseksperimenter på GbpA bundet til kitin. Komplimenterende elektronmikroskopiske studier med negativ farging viste hvordan GbpA dekorerte kitinfibrene den var bundet til. Studiene har ført til oppdagelsen av et nytt metalbindingssted på den karbohydrat-bindende overflaten til GbpA. Strukturanalyse av homologer sammen med mutagenese- og stabilitetsstudier har gitt hint på betydningen til denne nylig funnete metallbindingssete. Røntgenkrystallografiske studier var gjennomført for å forstå interaksjonen til oksygensubstanser med det aktive setet, noe som bedre karakteriserte LPMO-relatert katalytiske mekanismen til GbpA. De preliminære resultatene presentert i denne oppgaven legger grunnlaget for fremtidige røntgen- og nøytronkrystallografiske eksperimenter.

Vi rapporterer også strukturell arbeid på vitellogenin fra honningbie som var renset rett fra hemolymfen til insektet. En innledende, biofysisk karakterisering fulgt av en komplett strukturprediksjon var gjennomført. Det strukturelle arbeidet rapportert her representerer den første kryogenisk elektronmikroskopiske (cryo-EM) strukturen av full-lengde Vg ved høy oppløysing. Vi har identifisert en tidligere ukjent glykosyleringssete og noen metallbindingssteder. Strukturen ga informasjon om von Willebrand-faktor-type-D (vWFD) domenen, tidligere ukarakterisert for alle medlemmer av Vg superfamilien. En strukturell sammenstilling basert på en prediksjon ved hjelp av AlphaFold tillot identifikasjonen av den eneste domenen som ikke var observert i våreeksperimenter som en C-terminal cystein knute.

Introduction

Infectious diseases are among the leading causes of death worldwide [1] and their effects on health and economy have become more obvious after the COVID-19 pandemic. Furthermore, understanding pathogenesis is fundamental for animal health and food production systems, shaping our relationship with the natural world. Structural biology provides ideal tools to study pathogenicity at the most basal level. Basic science on the molecular mechanisms underlying protein function allows for the development of technologies that improve both human health and the use we make of natural resources. This thesis focuses on two unrelated but highly relevant protein targets that exemplify pathogen-host interactions from opposing points of view: the pathogen and its offensive mechanisms – GbpA from the human pathogen *V. cholerae* – and the host and its defenses – Vg as part of the immune system of the honey bee.

On the pathogen: GbpA from *V. cholerae*

Cholera and its molecular mechanisms

Cholera is an ancient severe diarrheal disease (reviewed in Kanungo *et al.* [2] and WHO [3]). If not treated immediately, cholera can lead to fatal dehydration. The ongoing seventh cholera pandemic is wreaking havoc around the globe and is responsible for over 140,000 deaths annually [4]. Transmitted primarily through the consumption of contaminated food or water, cholera affects mainly developing countries. Most outbreaks occur in warm climates in the wake of natural disasters and war. Examples are provided by the outbreak in Haiti in 2010 after the earthquake and in Yemen from 2017 as a result of the ongoing civil war. The treatment of cholera usually focuses on symptoms, mostly through rehydration. Some vaccines exist, but their long-term effectiveness is limited. In addition, climate change is expected to complicate the fight against cholera through higher temperatures and an increase in the prevalence of natural disasters [5,6].

Cholera is caused by toxigenic strains of *Vibrio cholerae*. This species is Gram-negative, comma-shaped and halophilic with a polar flagellum. To date, only the O1 and O139 of the more than 200 serogroups – classified based on the O antigen polysaccharide structure – of *V. cholerae* are known to cause epidemic cholera. O139 is now virtually extinct, whereas the serogroup O1 is responsible for current outbreaks. Further serological classification of *V. cholerae* O1 is determined by biotypes (Classical or El Tor) and serotypes (Inaba, Ogawa or Hikojima). Classical strains are believed to be responsible for the previous six cholera pandemics in modern history, while El Tor strains are responsible for the current pandemic. El Tor strains are more infectious, persist longer in the environment and are associated with a higher rate of asymptomatic and mild cases.

The main virulence factor of *V. cholerae* is the cholera toxin (CT) (reviewed by Heggelund *et al.* [7]). In the host intestine, *V. cholerae* secretes CT, which binds the GM1 ganglioside present in epithelial cells. Upon internalization and retrograde trafficking, the CT is proteolytically processed and produces the irreversible ADP ribosylation of G proteins, affecting a myriad of processes through the activation of adenylate cyclase. Importantly, an efflux of chloride generates a general loss of electrolytes and fluids.

In order to colonize the host intestine, *V. cholerae* uses a range of colonization factors [8]. Specific adhesins are likely responsible for a non-reversible attachment. Those include the hemagglutinin FrhA and likely the outer membrane protein OmpU. However, colonization is likely initiated by non-specific

adhesins with low individual affinity. These include the outer membrane adhesion factor multivalent molecule 7 (Mam7), establishing general protein-protein and protein-lipid interactions, and the *N*-acetylglucosamine binding protein A (GbpA). GbpA recognizes *N*-acetylglucosamine molecules found in glycoproteins (particularly mucins) and glycolipids in the epithelial cell surfaces. GbpA has been shown to recognise *N*-acetylglucosamine (GlcNAc) units in intestinal glycoproteins (particularly mucins) and glycolipids, helping the bacteria colonize the human gut in pathogenic *V. cholerae* strains [8–10]. Furthermore, GbpA has been found to increase mucin production by intestinal cells in a NF-κB dependent manner [11]. In turn, intestinal mucins can also upregulate the production of GbpA by *V. cholerae* in a dose-dependent manner, ultimately leading to a successful intestinal colonization of the intestine and pathogenesis by *V. cholerae* [11]. Interestingly, GbpA from *Aeromonas* was recently found to induce proliferation of intestinal cells in zebrafish [12]. It has been suggested that GbpA is a promising target for vaccine design against cholera, with rabbits immunized with GbpA providing a significant survival advantage [9]. Interestingly, while deletion strains for GbpA showed only around a 50% decrease in attachment to epithelial cells *in vitro*, the effect of the deletion is much more radical when studied in an infant mouse model system, decreasing colonization efficiency around 10-fold and pointing to the major role of GbpA in intestinal colonization. Furthermore, some adhesins such as GbpA, have a dual role by participating not only in the intestinal colonization of the host intestine but also in the environmental survival for *V. cholerae* [13,14] (**Figure 1**).

Chitin and the environmental survival of *V. cholerae*

V. cholerae is a facultative pathogen and naturally found in aquatic environments worldwide. Therefore, the marine environment provides the main reservoir of the bacteria between outbreaks. The bacteria can be found free-living or attached to both biotic and abiotic surfaces forming biofilms [15–17]. Among those surfaces, chitinous ones are of particular importance. The chitinous surfaces of planktonic crustaceans such as copepods and water fleas are likely the most relevant reservoirs of *V. cholerae* [16,18]. Chitinous surfaces provide the bacteria with nutrients – both carbon and nitrogen – [19] as well as an increased tolerance to stress and protection from predators [13,20]. Chitin and its abundance have been linked to cholera outbreaks, with its seasonal occurrence linked to algal blooms that feed zooplankton. In addition, formation of *V. cholerae* biofilms on chitin allows for the horizontal transfer of genes – the pathogenicity is encoded by genetic mobile elements – [21,22], promotes an hyper-infectious state [23] and improves resistance to stresses such as stomach acidity [24]. Therefore, chitin can be considered not only a reservoir of *V. cholerae* but also a vector, with lower doses of the bacteria required for infection if it is associated with the polysaccharide [24].

Chitin (see Landman and Harries [25] for a complete review) is the most abundant polymer in the aquatic environment [26]. It is produced in large amounts by arthropods but also mollusks, algae, fungi, fish and amphibians. It is a polymer of β-(1→4) linked GlcNAc units. The acetyl amide group increases the hydrogen-bonding capacity of the polymer when compared to cellulose, giving chitin increased strength. Chitin is found associated with structural proteins and minerals, which include mostly amorphous calcium carbonate but also calcium phosphate and magnesium carbonate. All chitins are made of crystalline nano-fibers embedded into less crystalline chitin. There are two main forms of crystalline chitin, α-chitin and β-chitin. α-chitin is the most abundant, found in arthropod cuticles as well as fungal cell walls. In α-chitin, polymers are organized in an antiparallel fashion, forming tightly knit sheets that are stacked and stabilized by hydrogen bonds. These inter-sheet hydrogen bonds are

not formed between the sheets of β-chitin, which are composed of parallel polymers. β-chitin is found mostly in squid pens and some tubes synthesized by worms as well as in diatoms. Therefore, β-chitin is less resistant to mechanical and chemical degradation and the chitin of choice to obtain separated nanofibrils in a laboratory environment [27]. In general, the biomechanical properties of chitin vary significantly depending on the degree and type of crystallization as well as the presence of mineralization and associated proteins.

Different colonization factors and adhesins are responsible for attachment of *V. cholerae* to chitin. As for intestinal colonization described above, it is believed that an initial reversible attachment is followed by a secondary firm anchorage that will eventually lead to microcolony formation [20]. The colonization factors include the chitin-regulated pili (ChiRP) [19], the mannose-sensitive hemagglutinin (MSHA) [28], the toxin-coregulated pilus (TCP) and GbpA [9]. GbpA is found and highly conserved in all *V. cholerae* strains, pointing to its relevance for environmental survival [29].

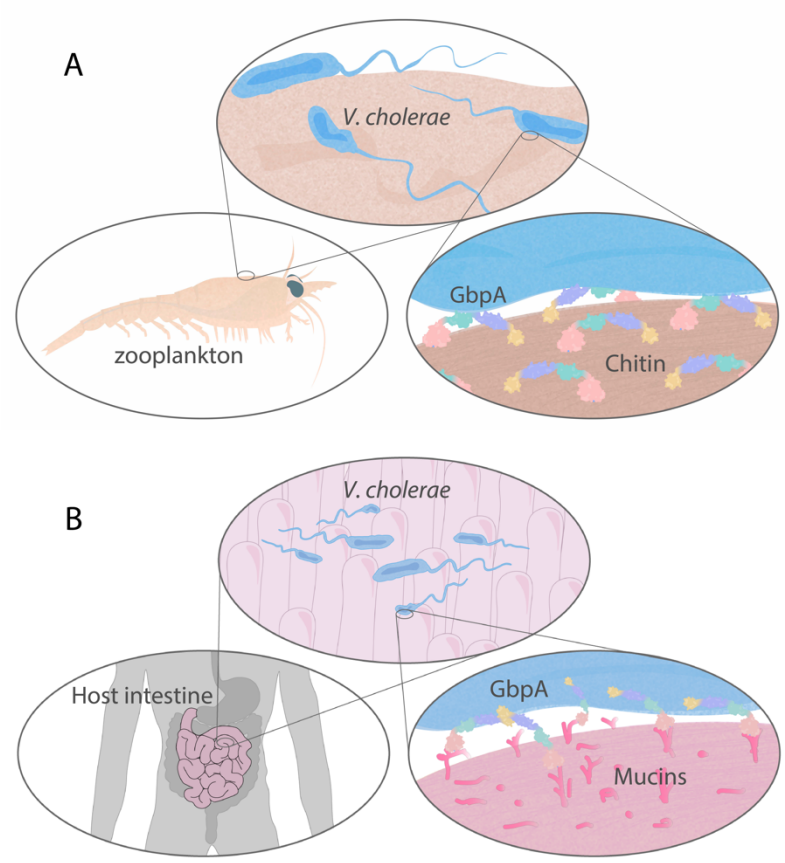


Figure 1: Dual role of GbpA of V. cholerae *in (**A**) environmental survival through attachment and degradation of the chitinous cuticles of zooplankton and (**B**) colonization of the host through binding of secreted intestinal mucins. Figure produced with Adobe Illustrator.*

LPMOs and GbpA

GbpA not only has the ability to bind to crystalline chitin, but also to oxidatively degrade it through its lytic polysaccharide monooxygenase (LPMO) activity [30]. Crystalline polysaccharides such as chitin or cellulose are by far the most abundant biopolymers on earth, but notably difficult to degrade and modify given their dense packing. LPMOs were only discovered recently [31] and thanks to their flat, surface-exposed active site are able to produce breaks in their crystalline substrates. LPMO research is therefore fundamental for the development of biofuel and nanomaterial technologies. Once the dense crystalline packing has been disrupted by LPMOs, glycosyl hydrolases (GHs) can further degrade the polysaccharides.

LPMOs are classified in different families in the Carbohydrate Active enZyme (CAZy) database based on their sequence similarity (AA9–AA11 and AA13–AA17) [32]. LPMOs are characterized by the presence of a single catalytic copper ion in the active site. The copper ion is coordinated by a histidine brace, which is formed by an N-terminal amino group of a histidine residue and its side chain (N_δ) as well as a another histidine residue side chain (N_ε). In addition, a tyrosine or phenylalanine side chain is located axially of the copper ion. LPMO domains are around 200 residues long, containing a β-sandwich and loops of varying length in an overall pyramidal shape, with the active site in the flat base [33]. Often, LPMOs contain additional domains with different functions, including carbohydrate-binding modules (CBMs) [34] as well as flexible C-terminal tails [35].

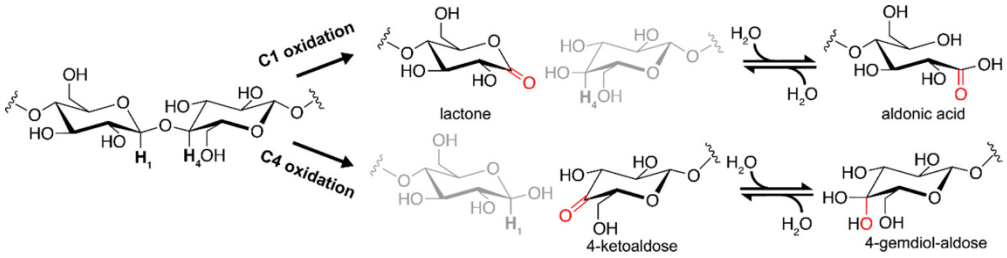


Figure 2: Oxidized products produced from a LPMO reaction on cellulose*. Reproduced from Chylenski et al.*[36] *with permission.*

GbpA is not the only LPMO involved in pathogenesis. Fungal LPMOs have been related to plant pathogenesis [37–39] and recently, a close homolog to GbpA was found to be a virulence factor for the human pathogen *Pseudomonas aeruginosa,* attenuating the terminal complement cascade in a way dependent on its LPMO activity [40].

LPMOs catalyze the oxidation of the C1 carbon or C4 of carbohydrate moiety using O_2 or H_2O_2 as a co-substrate into a lactone, acting both as monooxygenases and peroxygenases. Interestingly, in the absence of substrate, LPMOs generate H_2O_2 from O_2 (**Figure 2** and **Figure 3A**) [41,42]. However, although a plethora of structural information from LPMOs is available, the details of the reaction mechanism are still poorly understood (**Figure 3B**).

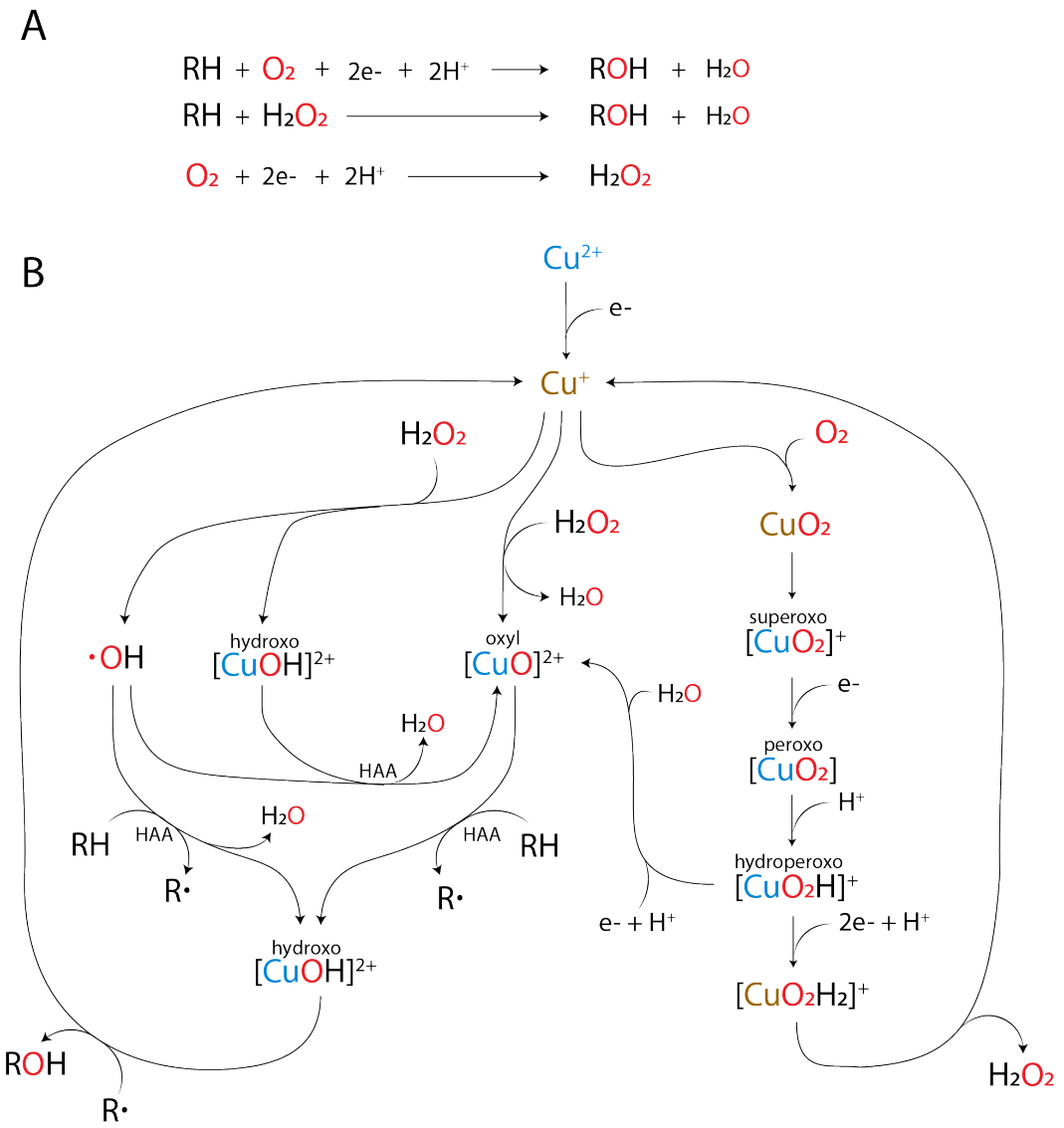


Figure 3: LPMO reaction mechanisms. A*. General reaction scheme for LPMOs with oxygen or hydrogen peroxide as co-substrates in the presence and absence of polysaccharide substrate (RH).* ***B.*** *Putative reaction mechanisms for LPMOs for the oxidation of polysaccharides, using both oxygen and hydrogen peroxide as co-substrates. A mechanism for the production of H_2O_2 as described in Caldararu et al.*[42] *is also shown. Based on Figure S14 from Bissaro et al.* [43] *and Figure 14 from Hadegård et al.* [44]*. Blue and bronze colors represent the different copper oxidation states.*

The first step of the LPMO reaction is the reduction of Cu(II) to Cu(I), either by a small molecule such as ascorbic acid and phenolic compounds or protein oxidoreductases [45]. Reduced copper in the active site can then bind O_2, generating a Cu-superoxo species $[Cu\text{-}O_2]^+$ [41]. However, this species is not sufficiently reactive for hydrogen atom abstraction (HAA) from the carbohydrate substrate, which is required for catalysis. HAA is more likely produced by a monooxygen species such as Cu-hydroxide $[Cu\text{-}OH]^{2+}$ or a Cu-oxyl $[Cu\text{-}O]^+$ [44]. Therefore, reduction and protonation of the Cu-superoxo species followed by water release is needed to proceed with the LPMO reaction starting from O_2. In the process, other dioxygen species are produced such as Cu-peroxo $[Cu\text{-}O_2]$ and Cu-hydroperoxo $[Cu\text{-}O_2H]^+$. These processes have been thoroughly studied using X-ray and, crucially, neutron

crystallography [46–49] in combination with quantum mechanical (QM) calculations. Bacik *et al*. [46] and O´Dell *et al*. [47] identified Cu-peroxo species, while Schröder *et al*. [48] cryo-trapped Cu-superoxo and Cu-hydroperoxo species and identified a neutral second shell histidine residue likely responsible for providing protons needed during the reduction of the Cu-superoxo species. Interestingly, the residue is neutral as it has been determined by careful analysis of high resolution X-ray crystal structures [50] and can become an imidazolate in the presence of a strong base such as the superoxo species [48]. The histidine residue is conserved in AA9 LPMOs, while other residues in the proximity of the active site could fulfill similar roles in other LPMO families [44]. For an AA10 LPMO from *Jonesia denitrificans*, a second shell glutamate residue has been suggested to be involved in the protonation of the dioxygen species leading to H_2O_2 production from O_2 [42].

Recently, it has been shown that H_2O_2 is the preferred co-substrate of LPMOs [43]. Crucially, using H_2O_2 as an electron donor requires only initial priming by another reducing agent and can support the reaction cycle sustainably with two electrons, while an O_2 reaction cycle requires two extra electrons for each reaction cycle. It remains unknown what the origin of those electrons is. Upon reaction with the Cu(I) active centre, H_2O_2 can undergo homolytic or base-assisted bond cleavage to produce both the Cu-oxyl and Cu-hydroxo species likely responsible for HAA on the polysaccharide substrate. Detailed crystallographic information about these reaction steps is lacking given the absence of substrate in most LPMO structures (**Figure 3B**). However, X-ray crystal structures of an AA9 LPMOs bound to small soluble polysaccharides have been reported [51,52], and a recent carbohydrate-free crystal neutron structures of the same protein [49].

An interesting twist to LPMO catalysis was recently suggested in which the stochiometric consumption of H_2O_2 is coupled only with the production of a substrate radical intermediate, whereas only the production of oxidized products is O_2 dependent [53]. With only the histidine residues in the histidine brace conserved in all LPMOs, it remains to be seen to what extent the reaction mechanism varies between different LPMOs.

GbpA at the molecular level

Most of the structural information available for GbpA comes from the X-ray crystal structure and small-angle X-ray scattering studies (SAXS) performed by Wong *et al*. [10]. GbpA possesses a modular structure of four domains with flexible linkers, showing an elongated solution structure (**Figure 4**). A construct including the first three domains was crystallized. The first domain ($GbpA_{LPMO}$) binds chitin, has LPMO activity [30] and was shown to also be responsible for intestinal colonization together with the second and third domain. The histidine brace is formed by the N-terminal histidine (H24) – after the signal peptide is cleaved off – and also H121, while the phenylalanine conserved in all AA10 LPMOs is F193 (**Figure 4**). Another conserved residue present in the active site is E67, likely to be involved in proton transfer along the reaction mechanism [42]. Residues important for chitin binding (Y61, E62, D188) have been identified based on conservation with CBP21 from *Serratia marcescens* [54,55].

The second ($GbpA_{D2}$) and third ($GbpA_{D3}$) domains of GbpA were found to be responsible for the interaction with *V. cholerae*. $GbpA_{D2}$ shows structural similarity to a domain of the flagellin protein p5 known to interact with the bacterial surface whereas $GbpA_{D3}$ shows structural similarity to a pili-binding protein [10]. However, the exact bacterial binding partners of $GbpA_{D2}$ and $GbpA_{D3}$ are still unidentified.

The fourth domain of GbpA is a carbohydrate-binding module 73 ($GbpA_{D4}$, CBM73). Although the structure of $GbpA_{D4}$ from *V. cholerae* has not been solved, the crystal structure of a full-length GbpA ($GbpA_{FL}$) including D4 was recently solved for *Vibrio campbellii* [56] . The protein is phylogenetically close with sequence identity between the two CBM73 domains of 66.7 %. In addition, the structure of a CBM73 domain from *Cellvibrio japonicus* has been solved using nuclear magnetic resonance (NMR) spectroscopy – sequence identity of 23.9 % [57]. In the same study, CBM73 was found unable to bind short oligosaccharides, requiring interaction with more than one parallel carbohydrate chain as in crystalline chitin. $GbpA_{D4}$ is non-essential for pathogenesis [10], indicating it has no relevant role in intestinal colonization and it likely does not interact with mucins.

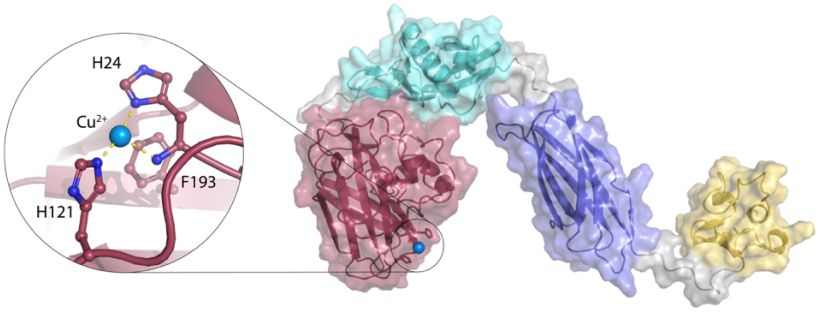


Figure 4: Molecular structure of GbpA predicted using Alphafold [58]***, zoomed in on the histidine brace****. $GbpA_{LPMO}$ is shown in red, $GbpA_{D2}$ is colored cyan, $GbpA_{D3}$ purple and $GbpA_{D4}$ is yellow. Copper (blue sphere) has been manually added to the structure prediction of full-length GbpA, while the histidine brace was obtained from 7PB7.*

As in the case of CT, GbpA is known to be recognized and secreted by the type II secretion system (T2SS) of *V. cholerae* [9].

On the host defense: Vg from the honey bee

On Vg pleiotropy

Vitellogenin (Vg) is part of the large lipid transfer protein (LLTP) superfamily. Found in almost all animal taxa, LLTPs appeared as a result for the increased need of lipid transport in metazoans [59]. LLTPs are major players in lipid transport. Members of the superfamily include insect apolipophorins, the apolipoprotein B (apoB) and the microsomal triglyceride transfer protein (MTP). LLTPs have been classified in three major groups: apoB-like LLTPs, MTP like LLTPs and Vg-like LLTPs [59].

Vg is a lipoprotein present in almost all egg-laying animals that has traditionally been studied in the context of vitellogenesis, when it is internalized by developing oocytes as a precursor for egg-yolk lipoproteins [60]. Classically characterized as female-specific, processed Vg is a major source of nutrients to the embryo during the first stages of development. During vitellogenesis, Vg expression in somatic cell lineages is boosted and the protein is secreted into circulation. Subsequent incorporation in yolk granules of the oocyte occurs via receptor-mediated endocytosis.

Interestingly, the presence of Vg is not only limited to females, but also found circulating in males and sterile castes in social insects[61,62]. In the last two decades, a range of new functions have been discovered for Vg. These new functions are mostly, but not only, related to immunity. Interestingly, along their evolutionary history, other LLTPs have also acquired immune-related roles [63,64]. LLTPs have likely been good candidates for immune neo-functionalization given the high levels in circulation and central role in lipid transport.

Vg has been found to have antibacterial activity in amphixious and bony fish. It is also relevant in different ways for the immunity of scallops, mud crabs and the malaria mosquito *Anopheles gambiae* (all reviewed in S.Zhang *et al*. [65]). In addition, Vg has been shown to recognize pathogen-associated molecular patterns (PAMPs) in fishes, where it can bind molecules both from Gram-positive and Gram-negative bacteria and also viruses, acting as an effector to neutralize pathogens and as an opsonin, promoting phagocytosis by macrophages [66–71]. These findings have recently been extended to simpler non-bilaterian organisms such as corals [72]. In honey bees, Vg has shown to recognize not only PAMPs but also endogenous danger-associated molecular patterns (DAMPs) [73]. Interestingly, trans-generational immune priming has also been observed through Vg ability to bind immune elicitors such as PAMPs and provide them to the developing eggs [74]. This suggests a link between Vg reproductive and immune functions. In addition, honey bees use Vg as a nutrient source to produce royal jelly, which is fed to queens and young larvae [75] and also carries immune elicitors [76].

In insects, Vg has been found to regulate and functionally interact with different hormones [75,77]. For social insects such as bees and ants, Vg governs social roles for sterile workers in a way dependent on their nutritional and metabolic status. It has been established that Vg functions together with juvenile hormone (JH) to regulate the social behavior and effect the metabolic status of worker bees. High Vg titers in hemolymph suppress JH and produce hive-staying and wintering bees, while high JH titers suppress Vg and produce forager bees [77–80]. Vg affects the longevity of queen and worker honey bees by keeping the levels of the pro-aging JH and, in addition, providing oxidative stress protection [73,81,82]. Vg titers can account for the drastically different lifespans of queen bees, wintering workers and foragers. Vg has been found to have social functions in other insects such as mosquitos and cockroaches [83,84], while its antioxidant properties have also been observed in organisms as distinct as the worm *C. elegans* [85] and eel [86].

As shown by the many studies cited above, the honey bee (*Apis mellifera* and *Apis cerana*) represents a particularly interesting model organism for the study of Vg. Queen honey bees represent the epitome of fertility, being able to lay their own body weight in eggs daily [87]. Exceptionally high levels of Vg are required for such a specialized task. In addition, Vg is also the main circulating protein in workers, representing up to 30-50 % of the total protein content. Furthermore, the honey bee is an ecologically and economically highly relevant species given its roles in pollination and food production. Crucially, low Vg levels have been identified as a predictive marker of Colony Collapse Disorder (CCD), which has been devastating to honey bee populations worldwide [88].

Vg at the molecular level

The functional complexity of Vg is underlined at the molecular level by a highly complex molecule [89,90]. Vgs are large multidomain proteins containing large structured cavities as well as flexible regions, with a molecular mass of more than 200 kDa. Crucially, they bind around 16% of their weight in lipids [91].

Post-translational modifications (PTMs) include phosphorylation, glycosylation and sulphation to different degrees. They are also known to coordinate metals such as Zn^{2+} in significant amounts [92,93]. Proteolytic cleavage is also an important part of Vg biology, occurring both after internalization in the oocyte and after synthesis. Vgs have been reported to form dimers, but there is no conclusive evidence regarding their oligomerization state[94]. However, Vg varies significantly in different taxa, particularly between vertebrates and invertebrates. This variation includes the presence and absence of certain domains and the distribution and presence of PTMs. This variation is further increased by the presence of more than a single Vg gene in many species and the existence of Vg-like proteins. Thus, Vgs have undergone a significant amount of neo- and sub-functionalization along their evolutionary history.

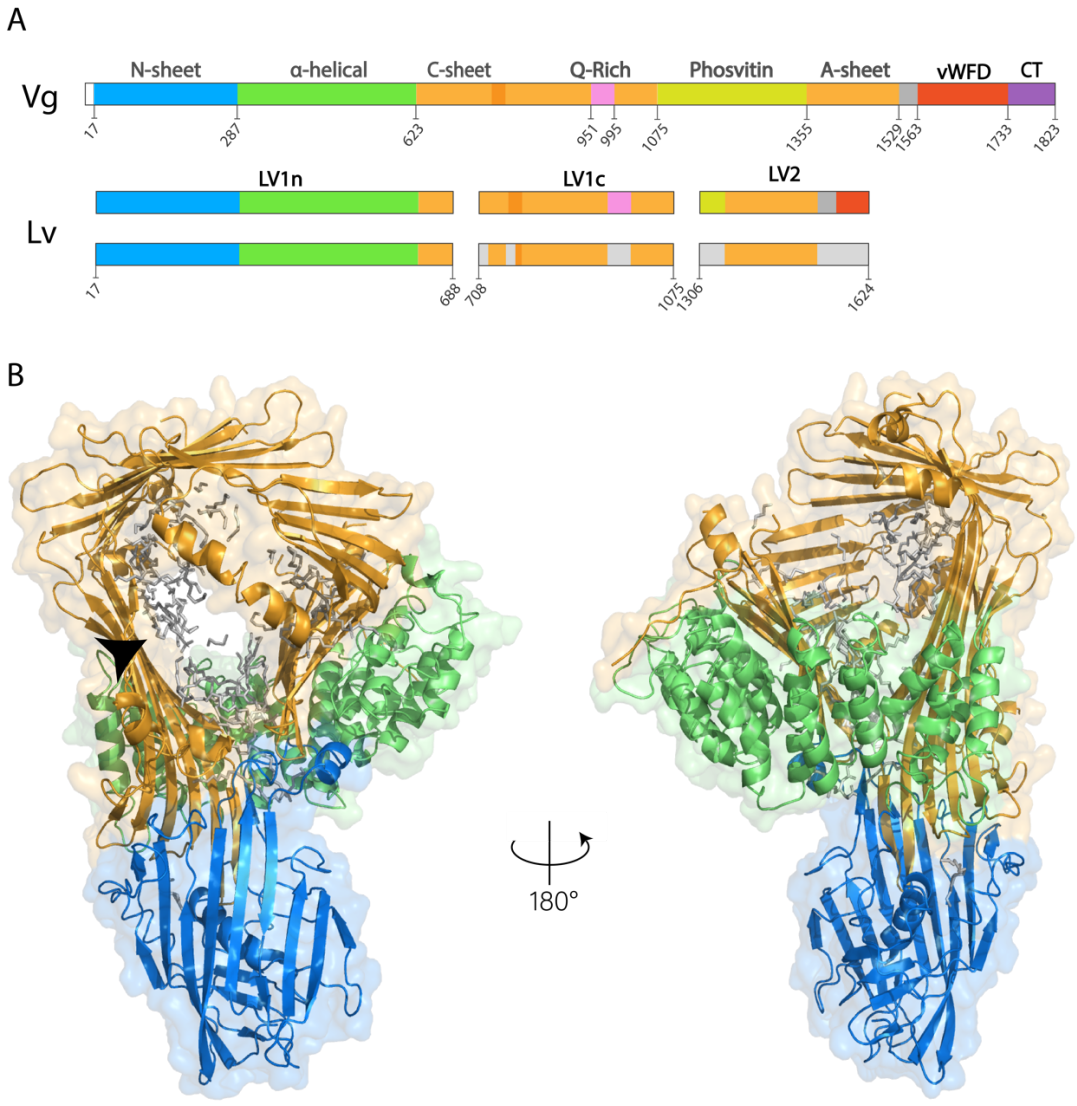


Figure 5: Lipovitellin from lamprey eggs. A. *Domain architecture of the silver lamprey Vg and its cleaved product lipovitellin (Lv) found in egg yolk. In the lower representation of Lv, only the parts for which density could be observed in the crystal structure are colored. B. Crystal structure of Lv (PDB ID:1LSH [91]). The different components of the Vg lipid-binding module can be observed. The black arrow points to the lipid-binding cavity, where fragments of lipids can be observed, including phospholipid head groups and hydrocarbon chains.*

Most of the structural information available on Vg comes from the crystal structure of lipovitellin (Lv) from the silver lamprey (*Ichthyomyzon unicuspis*)[91,94,95]. Lv is the cleaved product of Vg obtained from egg yolk and includes the lipid-binding module that is characteristic of Vgs and LLTPs in general (**Figure 5**).

The lipid-binding module is located at the N-terminus of Vg and consists of around 1400 amino acids and several subdomains: the N-sheet, the α-helical subdomain and the A and C sheets. The N-sheet is one strand short of forming a β-barrel and known to be responsible for interaction with the Vg receptor and oocyte internalization[96,97]. The α-helical subdomain consists of an unusual right-handed supercoil of α-helices. This subdomain has been suggested to be involved in DAMP recognition and oxidative stress protection in the honey bee [73]. Wrapped by the α-helical subdomain, the A and C β-sheets are believed to form most of the lipid-binding cavity. These subdomains are generally the least conserved among Vgs given that they form less protein-protein interactions. Among different LLTPs the needs for lipid-binding are different and different lipid-binding cavities are observed. This is exemplified by the crystal structure of MTP solved recently[98]. In contrast to Lv, the MTP lipid-binding cavity is believed to accommodate a single lipid.

At the C-terminus of the lipid-binding module Vgs have a von Willebrand Factor type D domain (vWFD) and an uncharacterized small domain of unknown function. The vWFD domain is present in seemingly unrelated proteins such as the von Willebrand factor and intestinal mucins, where it is involved in the formation of strong oligomers through disulphide bond formation[99,100]. In vertebrate Vgs, the vWFD domain is often referred to as β-component.

Insect Vgs often contain polyserine flexible regions between the N-sheet and α-helical subdomains in the lipid-binding module [101]. This region has been found to be a target region for proteases and is protected by phosphorylation. From studies on the honey bee, it has been hypothesized that a fragment resulting from proteolysis containing the N-sheet can translocate to the nucleus and regulate gene expression [102,103].

Vertebrate Vgs also contain highly phosphorylated regions known as phosvitin. These subdomains are found in loops in the lipid-binding cavity and they are believed to be responsible for coordinating metals to be delivered to the developing oocyte.

Although a range of molecular information is available, we are still far from understanding the molecular basis of Vg many functions, including vitellogenesis, its acquired roles in immunity and the interesting functions it has developed in social insects. Pointing to its relevance and unknown structure, Vg was recently used as an example in a news article by Nature on the new possibilities unleashed by the development of Alphafold [104].

Aims

The overall aim of this thesis was to characterize the unknown molecular mechanisms underlying the functions of both GbpA and Vg as examples of molecules involved in pathogenesis and immunity from different perspectives, shining light on the immune warfare that has shaped the evolution of life on earth.

Specifically, for GbpA, the aim was to understand how binding to chitin occurs and how it affects the structure of the protein, promoting bacterial colonization. Accordingly, it was planned to use nanometer-resolution structural biology tools such as small-angle scattering (SAS) and electron microscopy (EM). Such techniques allow the study of protein interactions with an insoluble substrate like chitin. Information about how GbpA binds to and affects chitin would allow a better understanding of the role of GbpA during *V. cholerae* colonization of its aquatic reservoirs and environmental survival in general. It would also provide clues regarding how the bacteria bind to GbpA itself, and how the protein is utilized by the bacteria. Another aim was to gain a deeper understanding of the catalytic mechanism of GbpA and LPMOs in general. Here, snapshots obtained by high-resolution methods such as X-ray and, importantly, neutron crystallography could provide crucial information. The biggest bottleneck for neutron crystallography is the sample preparation step, with the need for large crystals of, ideally, perdeuterated protein. Thus, another aim was to develop and optimize methods to obtain such samples of GbpA to ultimately use neutron crystallography.

For Vg, the main aim was to obtain an accurate and complete structural model. A high-resolution structure of the full-length protein would be a major step forward for this important, but structurally under-characterized, protein. To this end, cryogenic electron microscopy (cryo-EM) is a suitable technique, given the difficulties of producing recombinant Vg and the small amounts of pure sample that can be obtained from its native sources. Vg is also a target large enough for cryo-EM.

Summary of Manuscripts

Manuscript I (draft):

Using *Vibrio natriegens* for high-yield production of challenging expression targets

Natalia Mojica, Flore Kersten, Mateu Montserrat-Canals, Ute Krengel

The manuscript describes the production of different protein targets in the Vmax™ X2 expression system and compares them to traditional *E. coli*-based expression systems. Vmax™ X2, developed from a fast-growing member of the *Vibrio* genus, is advantageous for the production of naturally secreted proteins from *V. cholerae* since it possesses an equivalent protein secretion machinery. Vmax™ X2 produces high amounts of cholera toxin as well as GbpA and secretes it to the culture supernatant, improving not only expression yields (by 10 and 6-fold respectively) but also purity and ease of handling. The expression of other non-secreted protein targets is also reported, showing even more remarkable yield improvements when compared to *E. coli* based expression systems.

Significance: The manuscript adds to the growing body of literature supporting the use of the Vmax™ X2 expression system, highlighting some of its specific benefits for the expression of proteins from bacteria phylogenetically close to *V. natriegens*. It also encourages the use of the Vmax™ X2 expression system for challenging expression targets.

Contributions from the author: Produced $GbpA_{FL}$ in *E. coli* and *V. natriegens*. Contributed to the writing of the manuscript. Produced Figure 4.

Manuscript II (ACS Omega):

Perdeuterated GbpA enables neutron scattering experiments of a lytic polysaccharide monooxygenase

Henrik Vinther Sørensen, Mateu Montserrat-Canals, Jennifer S. M. Loose, Zoë Fisher, Martine Moulin, Matthew P. Blakeley, Gabriele Cordara, Kaare Bjerregaard-Andersen, Ute Krengel

The manuscript describes production of perdeuterated GbpA, paving the way for future neutron scattering experiments. Perdeuterated $GbpA_{FL}$ and $GbpA_{LPMO}$ have been produced in-home and scaled up at international deuteration facilities. Perdeuteration of in-home produced GbpA was assessed with mass spectrometry (MS). The activity of the deuterated protein ($GbpA_{FL}$) was analyzed and confirmed with established assays. In addition, the fold of the deuterated protein ($GbpA_{LPMO}$) was compared to the hydrogenated protein using X-ray crystallography, showing no significant differences. Preliminary small-angle neutron scattering (SANS) data are provided as a proof of concept.

Significance: The manuscript shows the first instance of perdeuteration of an LPMO. Perdeuteration (or the almost total substitution of hydrogen atoms for deuterium) of samples increases their suitability for neutron scattering experiments, removing the negative contribution to scattering from hydrogen atoms that cancels out deuterium signal and generally improving the signal-to-noise ratio. The manuscript lays the foundations for neutron scattering experiments of GbpA.

Contributions from the author: Produced the hydrogenated $GbpA_{LPMO}$ sample for X-ray crystallography. Collected and processed the corresponding crystallographic data. Produced deuterated $GbpA_{LPMO}$. Processed the crystallographic data for deuterated $GbpA_{LPMO}$. Performed the anomalous scattering experiments at the absorption edges of Cu and Zn, and analyzed the data. Contributed to the writing of the manuscript. Produced Figures 1D, 6, S2A, S3 and Tables 3, S1 and S2.

Manuscript III (draft):

Tangled up in fibres: How a lytic polysaccharide monooxygenase binds its chitin substrate

Henrik Vinther Sørensen, Mateu Montserrat-Canals, Reidar Lund, Ute Krengel

The manuscript describes how a method for studying chitin-GbpA interactions with SANS has been developed keeping chitin and the protein in suspension, allowing contrast matching. Contrast matching is possible thanks to the protocols developed in **Manuscript II** for GbpA perdeuteration. SANS is subsequently used to show how GbpA completely decorates the chitin fibers. The challenges and attempts at obtaining low-resolution models for GbpA bound to chitin are discussed. EM experiments of chitin-GbpA complexes provided complementary data on how the binding to chitin occurs, yielding extra insight complementary to the SANS findings. The data support a chitin colonization mechanism in which GbpA is secreted in large amounts by *V. cholerae,* preparing the ground for microcolony formation.

Significance: The manuscript describes the first SANS experiments on an LPMO and its substrate. New insight into how an LPMO binds chitin at the nanometer scale is provided.

Contributions from the author: Produced the hydrogenated $GbpA_{FL}$ sample for negative-stain EM. Collected and analyzed the EM data. Integrated the EM data and discussion into the manuscript. Produced Figures 4 and 5.

Manuscript IV (draft):

Calcium binding site in AA10 LPMO from *Vibrio cholerae* suggests modulating effects during environment survival and infection

Mateu Montserrat-Canals, Kaare Bjerregard-Andersen, Henrik Vinter Sørensen, Gustav Vaaje-Kolstad, Ute Krengel

The manuscript describes the identification of a new metal-binding site in the proximity of GbpA active site and chitin-binding surface. Using recombinantly produced protein variants, the effects of Ca^{2+} binding on the stability of GbpA were observed with differential scanning fluorimetry. The binding of other metals is also discussed. The effects on protein stability of different metals are studied using small-angle X-ray scattering (SAXS). Multiple sequence alignment and structural analysis of related LPMOs highlight the importance of the discovered binding site.

Significance: The manuscript represents the first description of this new metal-binding site of GbpA, showing a putative mechanism by which salts could modulate the catalytic activity and binding to chitin of GbpA, with potential effects to *V. cholerae* survival, infection and toxicity, with implications for LPMOs in general.

Contributions from the author: Produced $GbpA_{LPMO}$ and $GbpA_{FL}$ samples for differential scanning fluorimetry analysis, both for the wild-type protein and other engineered samples. Produced $GbpA_{FL}$ samples for SAXS and collected scattering data on them. Performed multiple sequence analysis of related orthologs and structurally analyzed AA10 LPMOs. Wrote the manuscript draft. Produced Figures 1, 2 and 3 as well as Table 2.

Manuscript V (*FEBS open bio*):

Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity

Vilde Leipart, Mateu Montserrat-Canals, Eva S. Cunha, Hartmut Luecke, Elias Herrero-Galan, Øyvind Halskau, Gro V. Amdam

The manuscript presents the structural model of the full-length honey bee Vg restrained by low-resolution negative-stain EM data. Classical homology modelling techniques and AlphaFold were used to produce the predicted model. Detailed modelling of the vWFD domain identifies a conserved Ca^{2+} binding site and the lack of cysteines available for intermolecular disulphide bonds. Basic biochemical characterization was performed on the protein directly purified from the bee´s hemolymph.

Significance: The model presents the first full-length prediction of Vg with experimental constraints. New insight is provided into the largely uncharacterized vWFD domain.

Contributions from the author: Performed biochemical characterization of purified Vg using blue native polyacrylamide gel electrophoresis (BN-PAGE) and size-exclusion chromatography (SEC). Contributed to the writing of the manuscript. Produced Figure 5.

Available at: https://doi.org/10.1002/2211-5463.13316

Manuscript VI (draft):

Cryo-EM high-resolution structural determination of native honey bee vitellogenin

Mateu Montserrat-Canals, Kilian Schnelle, Vilde Leipart, Øyvind Halskau, Gro Amdam, Arne Moeller, Hartmut Luecke, Eva Cunha

The near-atomic resolution structure of Vg obtained from the honey bee hemolymph is presented. The structure has been obtained using cryo-EM.

Significance: The structure represents the first experimental structure of an invertebrate Vg and the first for full-length Vg. It provides structural insight into the vWFD domain, for which the structure had not been solved in an LLTP. The C-terminal Cystine Knot (CTCK) domain was identified for the first time based on folding predictions by Alphafold. Information about post-translational modifications, metal and lipid binding as well as the identification of a new cleavage product with potential biological relevance allow an improved understanding of the mechanisms underlying the range of functionalities of Vg. The findings have many implications for Vgs of other species as well as for members of the LLTP superfamily.

Contributions from the author: Processed and analyzed the cryo-EM data and built the model for full-length Vg and the cleavage product. Performed the analysis of the structure. Wrote the manuscript. Produced all figures except for Figure S1.

Results and Discussion

In this section, a general overview of the results and discussion for this thesis is presented. Since most of the results and discussions are part of the manuscripts, they are not presented here. Reading the manuscripts first is recommended to get a full overview of this section.

On GbpA sample preparation methods

Structural biochemistry studies on protein usually require high amounts of pure protein. For most targets, a non-native expression system is required in order to produce protein recombinantly and in sufficient amounts. The most common and arguably simple expression systems are based on *E. coli*, a widely studied and fast-growing model organism. In **Manuscript I**, we exploited the Vmax™ X2 expression system to produce GbpA. Given its faster growth rate, this expression system produced even better yields of the protein. Furthermore, *Vibrio natriegens* , the bacterial species from which Vmax™ X2 has been developed, belongs to the same Genus as *Vibrio cholerae* and contains the equivalent secretion systems. In our case, the T2SS can recognize an epitope in GbpA and secrete it to the growth media. Thus, recovery and purification of the protein is easier than for the E.coli-based methods, in which the protein was extracted from the periplasm. These results facilitate sample production for GbpA for its use in structural biology analysis and assays, with yields increasing 6-fold in Vmax™ X2 compared to *E. coli*-based expression systems.

Moreover, isotope labeling is required for some structural biology techniques such as NMR spectroscopy and neutron scattering techniques. Perdeuteration, or the almost complete substitution of hydrogen for deuterium in a molecule, is expensive but required for the contrast matching experiments that we want to perform using SANS. Contrast matching allows the separation of the signal obtained from two interacting molecules, allowing for structural studies of a protein — such as GbpA — and a binding partner — such as chitin. In addition, perdeuteration allows the use of smaller crystals for neutron crystallography. Obtaining large enough crystals is usually the bottleneck to using this technique. Moreover, removing all hydrogens from the sample prevents incoherent contributions to scattering from such atoms. **Manuscript II** describes our efforts and success in producing cost-efficient methods for perdeuterating $GbpA_{FL}$ and $GbpA_{LPMO}$, both in-house and scaled up in international facilities. This sets the stage for neutron-scattering experiments of GbpA.

On GbpA binding to chitin

As described in the previous section, the results from **Manuscript II** allowed contrast-matching SANS experiments on GbpA and chitin described in **Manuscript III**. We could see how chitin is completely decorated with GbpA along its longitudinal axis while the biding is not uniform with some protein aggregation occurs between chitin fibers. In addition, we could confirm such findings with visual insight from EM. All these data suggest a mechanism by which GbpA starts the colonization process by completely coating chitin surfaces preparing the ground for other colonization factors, responsible for secondary firm anchorage to the polymer and microcolony formation.

Additional EM studies on chitin-GbpA$_{LPMO}$

In addition to the EM analysis reported in **Manuscript III** for GbpA$_{FL}$ binding to chitin, negative-stain EM studies were also done for GbpA$_{LPMO}$ alone. Importantly, for such a construct, the CBM73 chitin-binding domain of GbpA is missing. These additional results are shown in **Figure 6**. The results already presented in **Manuscript III** for chitin and chitin with GbpA$_{FL}$ are also shown here for comparison.

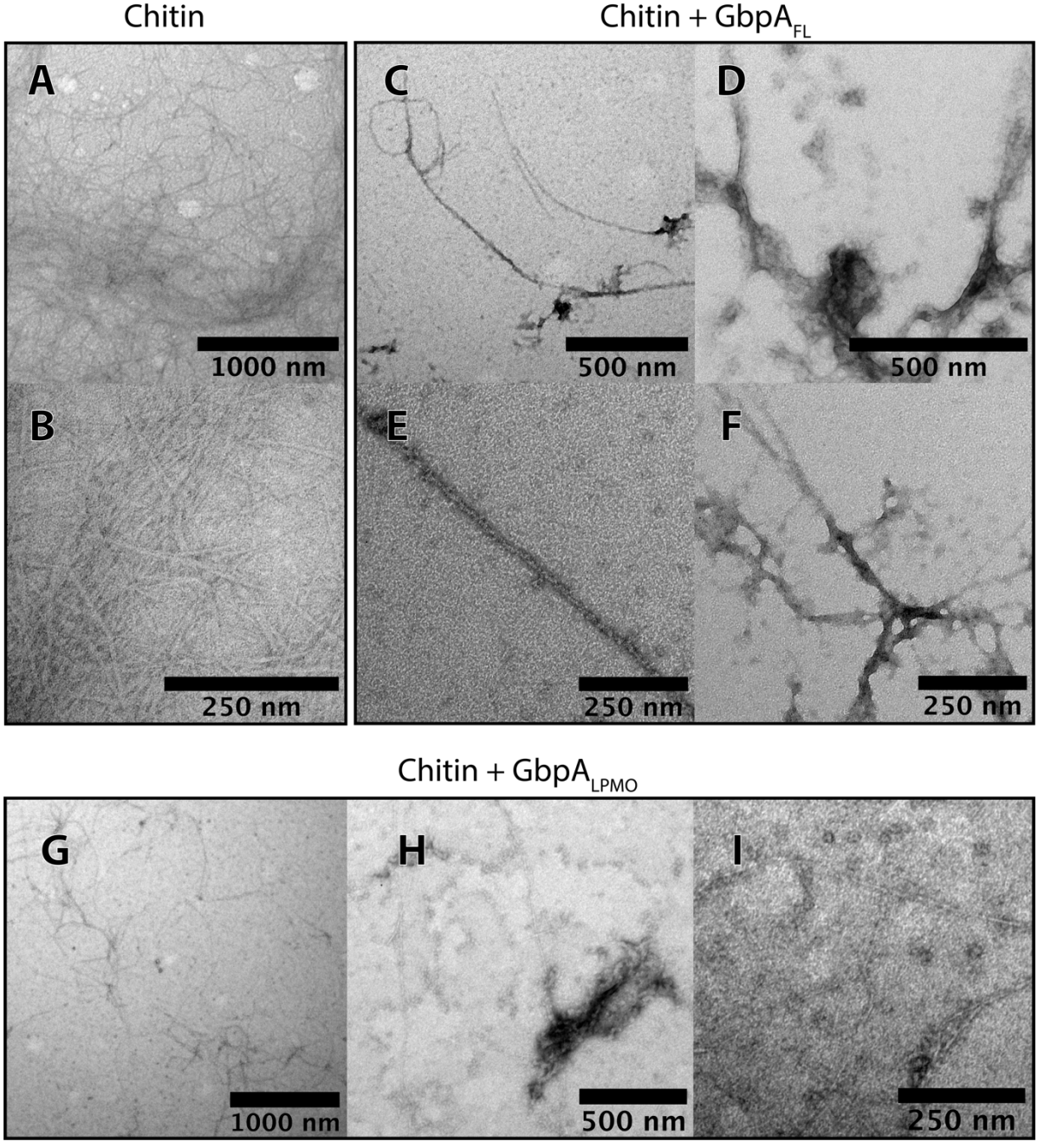


***Figure 6: Negative-stain EM chitin binding studies**. Micrographs of suspended chitin (**A**, **B**) in the presence of GbpA$_{FL}$ (**C**, **D**, **E**, **F**) or GbpA$_{LPMO}$ (**G**, **H**, **I**). Panels A-F are part of Figure 4 of **Manuscript III**.*

Interestingly, the aggregates that $GbpA_{FL}$ seems to produce among chitin fibers are not observed for $GbpA_{LPMO}$ alone. In addition, analysis of the width of fibers shows no significant increase upon addition of $GbpA_{LPMO}$ (**Figure 7**), although, upon visual inspection, the fibers seem to be more electron-dense compared to chitin alone. This, however, could be a result of just a slightly longer incubation with the negative stain. It is therefore not completely clear if $GbpA_{LPMO}$ can bind to the chitin fibers alone. If $GbpA_{LPMO}$ is binding along the chitin fibers forming a single layer, the increase in width in the fibers might be difficult to detect since the LPMO domain measures around 3.5 nm from the carbohydrate-binding surface to the opposite end of the domain. The results are not completely unexpected given that many LPMOs require CBMs in order to increase their catalytic efficiency [34]. This is also the case for GbpA, for which the presence of its small fourth chitin-binding domain (CBM73) might be required for stable binding to chitin.

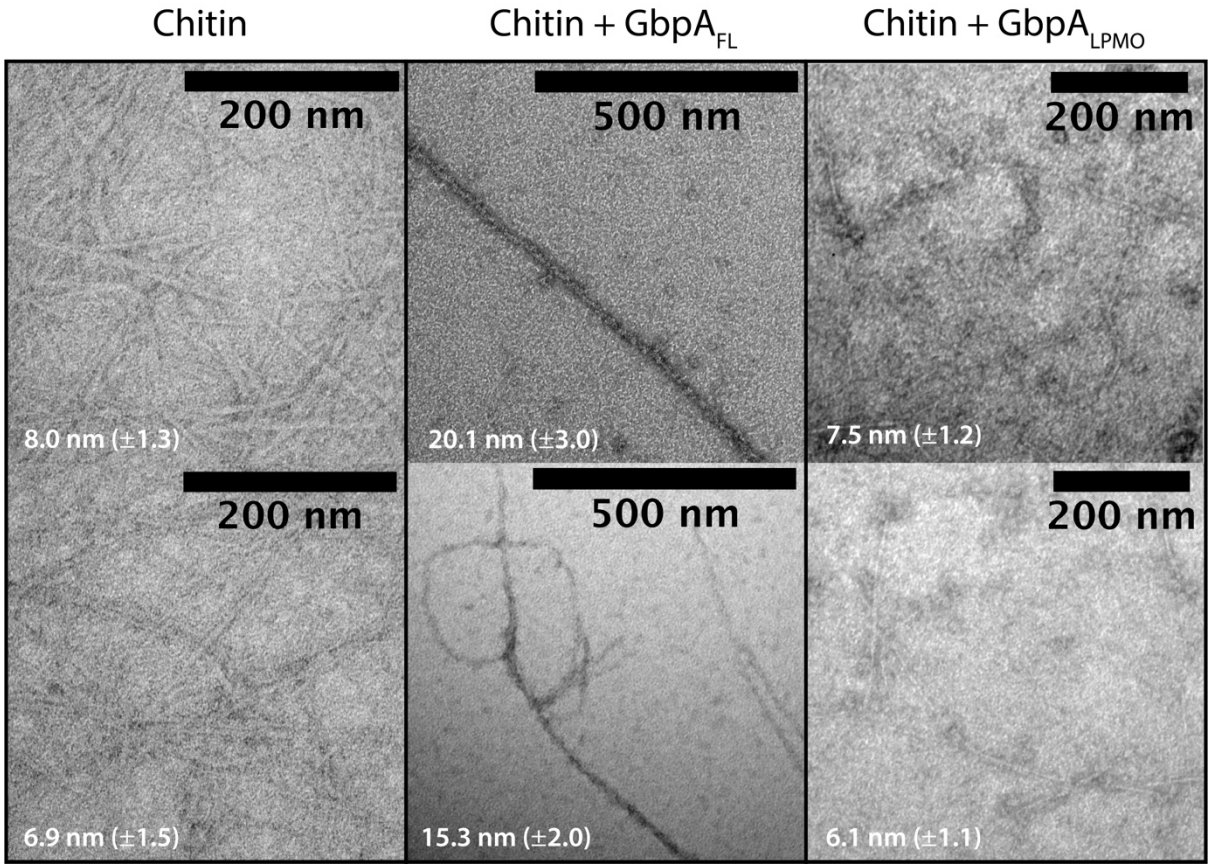


Figure 7: Measurements of chitin fibers thickness*. Fibers were measured from micrographs of chitin alone, chitin in the presence of $GbpA_{FL}$ or $GbpA_{LPMO}$. The average thickness of fibers in each micrograph is shown on the bottom left. Standard deviation values are shown in brackets. Fifteen measurements were done for each micrograph using Image J [105]. The panels corresponding to chitin alone and chitin + $GbpA_{FL}$ are part of Figure 5 of* ***Manuscript III****.*

On the GbpA LPMO enzymatic mechanism

The results from **Manuscript II** set the stage for future neutron crystallography experiments with $GbpA_{LPMO}$. Neutron crystallography allows the study of the protonation state of key residues and the

presence of water molecules and copper ions as well as oxygen species in the active site, allowing a better understanding of the LPMO mechanism. In addition, neutron scattering produces negligible radiation damage, allowing data collection at room temperature and preventing photoreduction of the active site copper(II). First, given that neutrons interact less with matter than X-rays, larger crystals need to be grown in order to be able to collect neutron scattering data from them. Ideally of at least 0.1 mm^3 for a perdeuterated crystal [106]. In addition, to be able to perform meaningful neutron crystallography in which biologically relevant reaction intermediates are trapped in the active site, preliminary X-ray crystallography experiments that allow a better understanding of the studied system are needed. Experiments trying to achieve these two goals are described and discussed in the following two sub-sections.

Optimization of $GbpA_{LPMO}$ crystals for neutron crystallography

$GbpA_{LPMO}$ can crystalize in different crystal forms that yield high-resolution diffraction data. The first crystal form described and characterized in **Manuscript II** was obtained from variations of condition B7 of the Structure I and II screen from Molecular Dimensions (100 mM sodium cacodylate pH 6.5, 200 mM zinc acetate and 18% w/v polyethylene glycol (PEG) 8000). Even though this crystal form corresponds to a high symmetry space group (*P* 2_12_12), which is generally desired for neutron crystallography, the presence of zinc ions close to the active site copper limits the biological significance of any structure obtained. The second crystal form is obtained from variations of condition G2 of the Structure I and II screen from Molecular Dimensions. This condition contains 200 mM 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 6.5, 200 mM $(NH_4)_2SO_4$ and 30 % w/v PEG 5000 monomethyl ether (MME)). This second crystal form has been used in the soaking and high-pressure freezing studies described below. The small volume of the asymmetric unit makes it suitable for neutron crystallography, even with the lack of symmetry in the unit cell for this crystal form (*P* 1) [106].

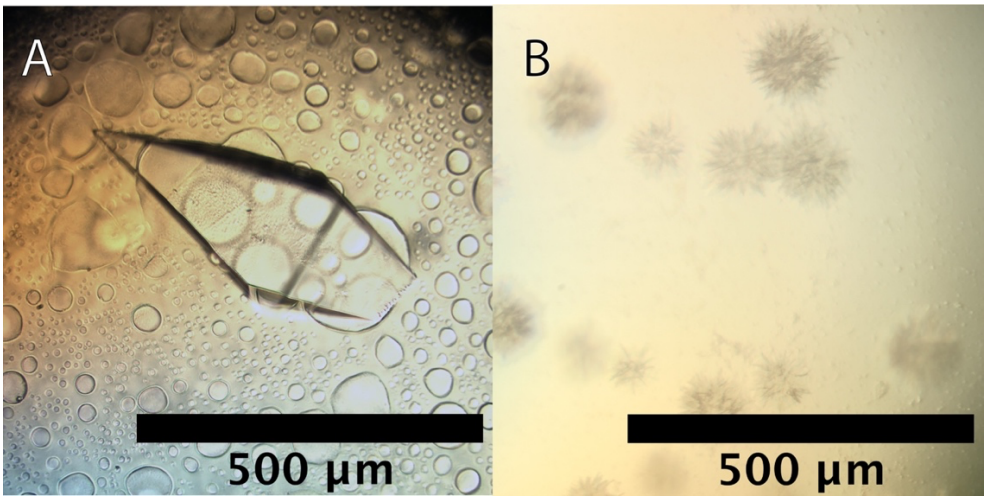


Figure 8: Different morphologies of $GbpA_{LPMO}$ crystals from the crystal form 2. A. *Large crystal of hydrogenated $GbpA_{LPMO}$.* ***B.*** *$GbpA_{LPMO}$ crystals with a "sea urchin" morphology.*

An issue with the second crystal from(200 mM MES pH 6.5, 200 mM $(NH_4)_2SO_4$ and 30 % w/v PEG 5000 MME) is reproducibility. Often, crystals were obtained with a sea urchin-like morphology, not suitable for diffraction (**Figure 8B**). Microseeding with crushed single crystals such as the one shown in **Figure 8A** has been used to try to address this reproducibility problem, without success. The seeds do not trigger any growth faster than spontaneous nucleation. Interestingly, in very few cases, the condition described produces crystals in a third crystal form. These crystals diffract at high resolution and belong to the high-symmetry space group $P\,4_32_12$, desirable for neutron crystallography. However, no microseeds have yet been obtained for these crystals.

The biggest crystals obtained of the second crystal form are about 0.5 mm on their longest edge and with an approximate volume of 0.004 mm^3 such as the one shown in **Figure 8A**. This crystal was obtained from mixing 0.5 μL of protein at 13 mg/mL with 0.5 μL of crystallization solution without any extra feeding steps. The crystal reached its final size in 4 months. Optimization of the individual components of the crystallization conditions is likely to produce better crystals that can grow bigger. GbpA crystallizes in conditions with very different PEGs being used as precipitants. Thus, a two-dimensional screen of PEGs with different molecular weights and varying concentrations could improve the crystals obtained. Another two-dimensional screen for salts could also be helpful. Since acetate binds close to the active site for the crystals grown in zinc acetate, different salts of acetate and formate could be explored. Using cations that can bind GbpA in the metal-binding site described in **Manuscript IV** could also help produce better crystals. Ideally, several feeding steps with highly concentrated protein in a small volume would allow the crystals to grow to the desired size. It could also be beneficial to exchange the protein buffer from tris(hydroxymethyl)aminomethane (Tris) pH 8.0 to MES pH 6.5 from before crystallization.

Given that deuterated protein is generally less soluble, the final crystallization conditions should then be slightly adjusted for growth of crystals with the perdeuterated material described in **Manuscript II**. Direct growth in D_2O based mother liquor at the equivalent pD would probably be less aggressive and require less crystal handling, making them safer options for the growth of larger crystals of better quality.

Capturing snapshots of LPMO reaction mechanism using X-ray crystallography

In an effort to obtain snapshots of the LPMO catalytic mechanism in GbpA, $GbpA_{LPMO}$ crystals were subjected to soaking with hydrogen peroxide (H_2O_2) and azide (N_3^-) as well as high-pressure freezing in oxygen (O_2). Both hydrogen peroxide and oxygen are known to be co-substrates of LPMOs, while azide has been used as an analogue to the hydrogen peroxide anion in spectroscopy studies [107]. Soaking with hydrogen peroxide was expected to not produce any usable data given its sensitivity to radiation damage. This, however, proved to not be a significant problem. **Table 1** contains the relevant data collection and processing statistics for the collected datasets. The refinement is preliminary; more details about it can be found in the Methods section. The results obtained are shown in **Figure 9**.

In all three cases, extra density was observed in the proximity of the active site in a prebinding state similar to the findings by O'Dell *et al*. [47]. The modelled azide, molecular oxygen and hydrogen peroxide are at 3.6, 3.3 and 4.1 Å away from the copper ion respectively. Whereas oxygen originating from atmospheric concentrations was found in the active site in the prebinding state by O'Dell *et al*. [47], for $GbpA_{LPMO}$ this was only observed when the crystal was subjected to oxygen at high pressure (50 bar) and frozen in liquid oxygen. Our results are comparable to most crystallography studies of LPMO

where oxygen species in the active site originating from atmospheric oxygen are only observed upon reduction of the active site copper by soaking with a reductant [46–48].

Table 1. Data collection and refinement parameters for $GbpA_{LPMO}$ soaking experiments

(a) Data collection

	$GbpA_{LPMO}$	$GbpA_{LPMO}$-O_2	$GbpA_{LPMO}$-H_2O_2	$GbpA_{LPMO}$-N_3^-
Beamline	ID23-2 (ESRF)	ID30A-3 (ESRF)	ID23-2 (ESRF)	ID30A-3 (ESRF)
Wavelength (Å)	0.8731	0.9677	0.8731	0.9677
Resolution range	41.6 – 1.6 (1.67 – 1.61)	34.5 – 1.5 (1.53 – 1.48)	50.2 – 1.7 (1.71 – 1.65)	34.3 – 2.1 (2.20 – 2.12)
Space group	*P 1*	*P 1*	*P 1*	*P 1*
Unit cell parameters				
a, b, c (Å)	44.3 46.3 51.4	43.7 45.9 51.1	43.7 45.9 51.2	44.1 46.2 51.3
α, β, γ (°)	82.2 79.6 71.6	81.3 80.1 71.2	81.5 80.1 71.3	81.7 79.9 71.4
R_{merge} (%)	8.8 (35.0)	13.4 (97.8)	12.9 (88.0)	8.8 (24.4)
$CC_{1/2}$	0.97 (0.85)	0.98 (0.38)	0.89 (0.29)	0.93 (0.85)
Mean I/σ	8.9 (3.2)	6.9 (0.9)	4.7 (0.9)	7.3 (3.3)
Completeness (%)	95.9 (95.2)	83.4 (44.5)	58.9 (2.9)	91.2 (93.0)
Multiplicity	2.6 (2.6)	2.6 (2.6)	1.8 (1.2)	2.0 (2.0)
Unique reflections	47206 (4680)	51220 (2745)	32105 (133)	19350 (1994)

(b) Refinement

	$GbpA_{LPMO}$	$GbpA_{LPMO}$-O_2	$GbpA_{LPMO}$-H_2O_2	$GbpA_{LPMO}$-N_3^-
R_{work}/ R_{free}	0.172/0.194	0.206/0.232	0.199/0.231	0.187/0.241
Macromolecules/a.u.	2	2	2	2
Number of non-hydrogen atoms	3187	3324	3001	3215
Protein	2812	3063	2798	3073
Ligands	2	4	4	8
Waters	373	257	199	134
B-factors (Å^2)				
Protein	11.6	14.5	13.7	20.4
Ligands	31.2	29.0	35.3	23.9
Waters	21.8	23.7	18.5	21.7
R.M.S.D. from ideal values				
Bond length (Å)	0.009	0.010	0.010	0.009
Bond angles (deg)	1.57	1.59	1.64	1.58
Ramachandran Plot				
Favored (%)	97.2	96.6	96.6	95.8
Outliers (%)	0.0	0.3	0.0	0.6

Data reported treating Friedel pairs as a single reflection/ Statistics for the highest-resolution shell shown in parentheses

The chemical species observed in our results seem to be stabilized by E67, a residue likely to be key for proton transfer during the LPMO reaction [42], as well as E63 from the other $GbpA_{LPMO}$ molecule present in the asymmetric unit (**Figure 9A**). Although this represents a non-native contact, it could be equivalent to an hydroxyl group from chitin. Schröder *et al*. [48] observed a crystal packing very similar as the one observed here for $GbpA_{LPMO}$, with the carbohydrate-binding surface of the two molecules

in the asymmetric unit facing each other. $GbpA_{LPMO}$ molecules are also facing each other but rotated 180° along the carbohydrate-binding surface in our structures. It was suggested that in the LPMO-binding surface, proline rings emulate pyranose rings in carbohydrates [48]. Therefore, the carbohydrate-binding surfaces of LPMOs could mimic carbohydrates themselves, allowing binding and producing crystal packing as the one observed for GbpA. The similarity could then be extended to the hydroxyl groups available for interaction with the co-substrates that were here observed, assuming that a carbohydrate-like binding surface could establish similar contacts to the ones the polysaccharide chains establish between each other. The similarities with the crystal packing observed by Schröder *et al.* [48] extend to the presence of co-substrates/oxygen species in only one of the molecules present in the asymmetric unit. In our structures, this is the case for oxygen and hydrogen peroxide, with only azide being observed binding the two $GbpA_{LPMO}$ molecules in the asymmetric unit. Schröder *et al*. [48] suggested that only the active site in one of the molecules resembles the native active site in presence of carbohydrate given the position of polysaccharide interacting side chains.

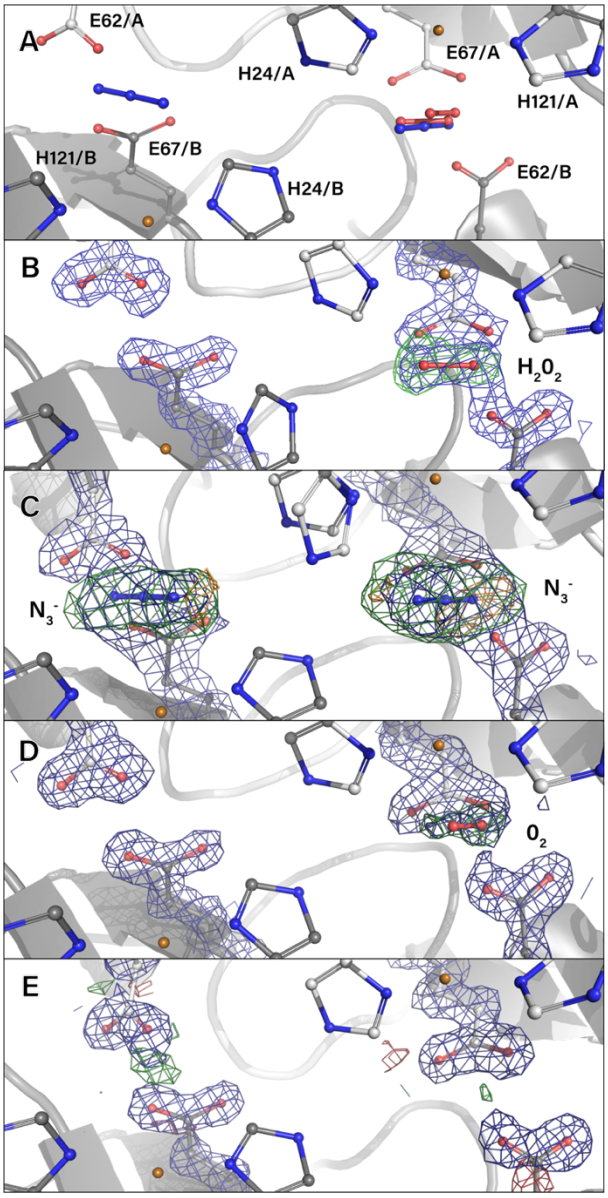


Figure 9: *(Previous page)* ***Soaking and high-pressure freezing experiments on GbpA****. The interface between the two* $GbpA_{LPMO}$ *molecules in the asymmetric unit is shown, with different shades of gray for each molecule and the histidine side chains in the histidine brace – H24 and H121 – as well as D62 and D67 visible. The copper ion in the histidine brace is colored in bronze. Panel* ***A*** *shows the equivalent position of* H_2O_2, O_2 *and* N_3^- *between the D67 and D62 of opposing GbpA molecules. The electron densities corresponding to D67 and D62 as well as the different ligands are shown for* H_2O_2 (***B***), N_3^- (***C***), O_2 (***D***) *and in the absence of ligand (****E****) at 1.5*σ*. The difference map (*$F_o - F_c$*) before the modelling of each ligand is shown in green and red. Orange density corresponds to leftover positive density in the difference map (*$F_o - F_c$*) after the modelling of the ligands, observed only for* N_3^-*. Difference maps are shown at 3*σ*.*

LPMOs are activated by reducing the Cu^{2+} in the active site to Cu^{+} [41], which allows the binding of oxygen species to the metal and its subsequent activation. Reduction of the copper can be achieved by adding a reducing agent. In crystals this can be achieved with a soak in ascorbate. In addition, photoreduction of copper upon exposure to moderate X-ray doses is well characterized [108]. The angles (**Figure 10**) and parameters relevant for assessing copper photoreduction in LPMOs for each of the copper active site in the datasets collected are shown in **Table 2**. Given the lack of symmetry in the unit cell X-ray doses are high. This, together with the lack of coordinating water molecules (or they increased distance from the copper) and the low θ_3 angles suggests that the copper in the active site is reduced. The values for θ_T are puzzling, since they are generally positive for LPMOs and there seems to be low correspondence between the different molecules in the asymmetric unit for a single dataset, suggesting they might not be good indicators of photoreduction for GbpA. More studies on photoreduction specific to $GbpA_{LPMO}$ would be required to better understand the histidine brace of this particular LPMO.

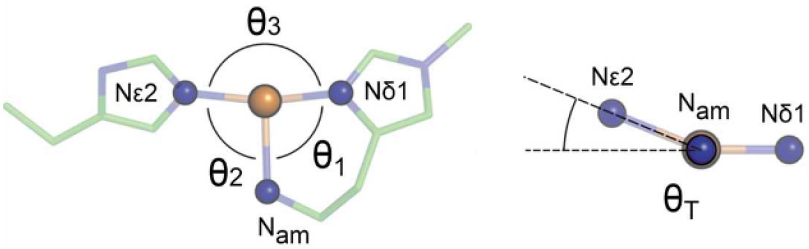


Figure 10: Nomenclature of the angles in the histidine brace relevant for assessing copper photoreduction. *Reproduced from Tandrup et al.* [108] *under creative commons license.*

Even though the copper(II) seems to be reduced by X-rays, the oxygen species are still found in a pre-binding state. This is often reported for LPMO structures[46–48], where the chemical reduction of the active site prior to freezing and exposure to the X-ray beam is required, allowing the cryo-trapping of the oxygen species bound to copper. This represents the obvious next step for the experiments reported here.

Table 2. Parameters relevant to assess X-ray photoreduction in the $GbpA_{LPMO}$ active site.

	X-ray dose (Gy)	a.u. site	H_2O_{eq} (Å)	H_2O_{ax} (Å)	θ_3(°)	θ_T (°)
Empty	$7.18 \cdot 10^7$	1	No	No	163.6	-10.6
		2	No	No	156.7	-12.6
O_2	$5.3 \cdot 10^7$	1	No	No	171.2	4.5
		2	3.3 Å	No	166.7	-6.3
N_3^-	$7.09 \cdot 10^7$	1	No	No	161.4	-6.3
		2	No	No	151.3	-10.3
H_2O_2	$1.35 \cdot 10^8$	1	No	No	159.9	-3.9
		2	No	No	146.3	22.3

Parameters are analyzed according to Tandrup et al. [108]. Parameters are shown for each active site present in the asymmetric unit of the datasets collected. Moderate to high X-ray doses, the absence of water ligands (H_2O_{eq} and H_2O_{ax}) as well as values lower than 170° for θ_3 and higher than 3° for θ_T are strong indicators of a Cu^+ site.

In order to better characterize the formation of oxygen species, it was initially planned to use on-line *in crystallo* Raman spectroscopy to catch the binding of oxygen species to the copper ion during X-ray data collection. This is likely not going to be possible given that oxygen species are found in a prebound state even if photoreduction of copper occurs. However, this could maybe be changed by room temperature data collection that would allow the movement of the oxygen species from the prebound state to the copper upon reduction of the metal. General radiation damage and the increased sensitivity to photoreduction would have to be accounted for in such an experiment. So far, preliminary Raman spectra upon excitation with a 785 nm laser were collected of $GbpA_{LPMO}$ crystals soaked in H_2O_2 and compared to a crystal that had not been soaked (**Figure 11**). By comparing the two spectra, peaks that could correspond to H_2O_2 were identified. However, in order to unambiguously assign peaks corresponding to H_2O_2, isotopically labelled hydrogen peroxide should be used.

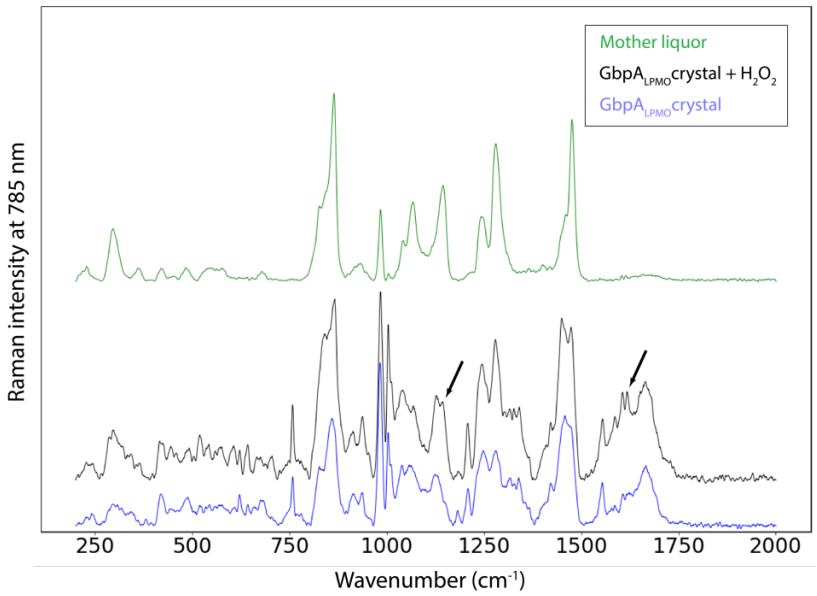


Figure 11:* In crystallo *Raman spectroscopy for $GbpA_{LPMO}$ crystals upon H_2O_2 soaking. *Raman spectra for the mother liquor (green), a $GbpA_{LPMO}$ crystal (blue) and a similar crystal soaked in H_2O_2. The black arrows indicate the additional peaks observed only in the presence of H_2O_2 in the crystal.*

On the effects of salts on GbpA

In **Manuscript IV** we identified a new metal-binding site in $GbpA_{LPMO}$ that binds calcium and potassium. The binding site is located in a loop adjacent to the active site that is part of the chitin binding surface. In addition, *V. cholerae* and GbpA in particular encounter calcium in different concentrations during environmental survival and intestinal colonization. Calcium precipitates are present in large amounts in chitin, whereas calcium ions are also important for the stability and fold of intestinal mucins. Thus, it is likely that the metal site plays a role in regulating GbpA binding and activity towards its substrates.

The hydrogenated and deuterated $GbpA_{LPMO}$ structures reported in **Manuscript II** also featured an ion in the newly identified metal-binding site. However, the description of the metal-binding site in the $GbpA_{LPMO}$ structures reported in **Manuscript II** is not included in **Manuscript IV**. The discussion around the metal-binding site from the structures in **Manuscript II** is less relevant biologically given the close presence of a zinc ion originating from the crystallization solution next to the metal-binding site (**Figure 12**). However, the metal-binding site as it is observed in the $GbpA_{LPMO}$ structures from **Manuscript II** is presented and discussed here.

The metal-binding site of GbpA observed in the structures from Manuscript II

The metal-binding site is shown in **Figure 12** for hydrogenated $GbpA_{LPMO}$. The metal species identified in the structures from **Manuscript II** is likely not calcium given the different geometry of the coordination compared to the calcium-bound structure in **Manuscript IV**. The geometry of the coordination is octahedral as described for potassium in **Manuscript IV**. Here, the side chain of D70 is not only involved in coordinating the metal ion but also a zinc ion located in the proximity. Additionally, in this crystal form the coordination sphere is not completed by a water molecule, but the side chain of E173 from the neighboring asymmetric unit. Interestingly, for the second $GbpA_{LPMO}$ molecule in the asymmetric unit, the site is occupied by a water molecule in close proximity to another zinc ion.

Both sodium and magnesium at full occupancy and potassium at lower occupancy (0.66) can be accurately modelled in the active site. X-ray scattering cannot distinguish between the three two ions. Both sodium and magnesium ions are isoelectronic and produce negligible anomalous scattering at the wavelength at which the datasets were collected (0.8731 and 0.9763 Å). The anomalous scattering of potassium should be slightly higher but still very low. No anomalous scattering signal at all was observed in the metal-binding site. Sodium was present in the protein storage buffer (50 mM), while magnesium and potassium were only present in the bacterial growth medium. In an attempt to disprove the presence of magnesium in the crystals, they were subjected to energy dispersive X-ray spectroscopy (EDS) under excitation from an electron beam in a scanning transmission electron microscope (STEM). However, the results were inconclusive given that the spectra of magnesium overlaps with that of zinc, which is present in large quantities in the crystal. Sodium was modelled in the structures based on its much higher concentration in the mother liquor, although the presence of magnesium or potassium cannot be ruled out.

Even if sodium or magnesium are proven to bind the newly identified metal-binding site, the ability to do so is likely not relevant biologically, since its presence does not produce specific stabilizing effects on GbpA as described in **Manuscript IV**. The binding might then be a result of the restrains imposed by crystallization. However, it remains a possibility that the binding of ions different to calcium in the

metal-binding site is a form of competitive inhibition as is suggested for potassium in the discussion of **Manuscript IV**. In such hypothesis, metal ions more abundant than calcium might bind to the metal site, regulating the binding of calcium by only allowing its binding to the metal site when calcium concentrations reach a certain threshold.

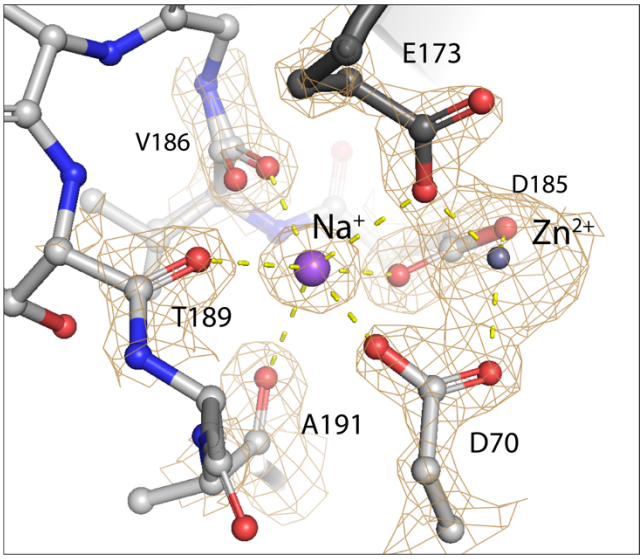


Figure 12: The newly identified metal-binding site in GbpA observed in the structures from Manuscript II, likely coordinating sodium*. The metal-binding site observed in the structure included in* ***Manuscript II*** *for hydrogenated* $GbpA_{LPMO}$ *(PDB ID: 8CC3). Density is shown at* 2σ*. Octahedral coordinating geometry is observed with contacts including E173 from the neighboring asymmetric unit and a zinc ion is located in the proximity. The full coordination sphere is not shown for the zinc ion, with its identity confirmed by anomalous scattering analysis as described in* ***Manuscript II****.*

On Vg structural studies

The structure prediction and, particularly, the cryo-EM structure of full-length Vg from the honey bee described in **Manuscript V** and **Manuscript VI** represent a major step forward in our understanding of the molecular mechanisms behind the multiple functions of Vg.

Interestingly, the results described in **Manuscript VI** are a good example showcasing the strengths and limitations of the protein structure software prediction AlphaFold. The AI-based software did an excellent job at predicting both the domain organization of the full-length Vg as well as the specific position of the backbone chain in problem areas such as loops that have no real homology to any previously experimentally solved structures. The backbone R.M.S.D. value between prediction and cryo-EM structure is surprisingly low (1.9 Å). The prediction also provided structural information for the CTCK domain whereas no experimental data could be obtained for it given the flexible linker that connects it to the rest of Vg. The prediction was also good enough to allow the identification of the previously uncharacterized domain as a CTCK domain through structural alignment. However, the prediction provides limited information regarding metal-binding sites, showing only which residues could be good candidates, but not allowing the accurate prediction of side chain conformations nor

the presence of ions. In addition, no information regarding PTMs or lipid binding could be obtained, although the lipid-binding cavity structure and shape were predicted accurately.

Importantly, the biological relevance of the cleavage product identified remains an open question, even after experimental structure determination.

Conclusions and Outlook

Methods to produce GbpA samples suitable for structural biology analysis to better understand the functions of the protein have been developed. First, GbpA production yields have been increased and the expression and purification protocols of the protein have been simplified with the use of the Vmax^{TM} X2 expression system (**Manuscript I**).

In addition, protocols for the production of perdeuterated GbpA both in-house have been developed and further scaled-up at international facilities, with the protein activity and structure similar to that of the hydrogenated protein (**Manuscript II**). The development of these methods has set the stage for neutron scattering studies, including SANS and neutron crystallography. Perdeuteration of GbpA in the Vmax^{TM} X2 expression system has already been tried successfully by other members of our group and not reported here.

SANS and EM have been used to study binding of GbpA to chitin together with EM (**Manuscript III, Additional EM studies on chitin-GbpA$_{LPMO}$**). Important insight into how the binding occurs has been gained, with implications for chitin colonization by *V. cholerae*. The negative-stain EM results provide a solid starting point for higher-resolution cryogenic electron tomography (cryo-ET) studies on individual fibers coated with GbpA. The binding of GbpA$_{LPMO}$ alone and the function of the CBM of GbpA can be further investigated assessing the binding to chitin of a GbpA construct containing all domains except for the CBM, which we have available in our group. Both SANS and EM could be used to that end. In addition, to further investigate the biological role of GbpA, the EM setup presented in this thesis could be exploited to understand which bacterial molecules might be the binding partners of GbpA$_{D2}$ and GbpA$_{D3}$.

One of our goals is to perform neutron crystallography experiments on GbpA$_{LPMO}$. This technique is generally limited by crystal size and efforts have been directed towards obtaining larger crystals (**Optimization of GbpA$_{LPMO}$ crystals for neutron crystallography**). In addition, in order to obtain biologically relevant snapshots of the LPMO reaction mechanism from neutron crystallography, know-how into the trapping of oxygen species in the active site is required. Initial insight into pre-binding states has been obtained (**Capturing snapshots of LPMO reaction mechanism using X-ray crystallography**), and the way is paved for further studies in which the in active site copper is chemically reduced and can bind to oxygen species. In addition, it might be feasible to track the binding of oxygen species to copper using Raman spectroscopy and X-ray photoreduction while the crystal is exposed to the X-ray beam by collecting data at room temperature.

Finally, using X-ray crystallography a new metal-binding site in the proximity of the GbpA LPMO active site and carbohydrate-binding surface has been identified (**Manuscript IV, The metal-binding site of GbpA observed in the structures from Manuscript II**). Thermal stability data and structural alignment provide clues into the relevance of the active site for the role of GbpA in environmental survival and pathogenesis of *V. cholerae*. In the near future, we plan to obtain the binding constants for the different salts to include in the manuscript. Further studies into protein activity in the presence of different salts as well as a complete phylogenetic study of the active site are required to better understand the role of the newly discovered metal-binding site. In order to unambiguously identify the metal binding in the hydrogenated and deuterated GbpA$_{LPMO}$ structures from **Manuscript II**, either

atomic absorption (AA) spectroscopy or inductively coupled plasma MS (ICP-MS) of dissolved crystals would be required.

For the honey bee Vg, the aim was solving its structure in order to tackle many of the open questions regarding its structure to function relationship. To this end, a full-length structure prediction constrained by low-resolution experimental data was reported (**Manuscript V**) and, subsequently, the high-resolution cryo-EM structure (3.2 Å) was solved (**Manuscript VI**). The model has been the first experimental structure of a full-length Vg, with the first experimental model of a vWFD domain from an LLTP. The Vg structure has provided insight into domain organization, flexible regions, post-translational modifications and zinc binding sites. The accuracy of the prediction by AlphaFold has showcased how the software represents a powerful tool to study Vg from different species and other Vg-like LLTPs. Such studies would provide valuable insight into neo-functionalization in different taxa. Further single-particle cryo-EM studies on Vg can be designed in order to identify how it interacts with PAMPs and better understand Vg immune fuctions. Furthermore, the identity of its cleavage products could be easily studied by subjecting fresh samples of hemolymph and fat-body purified Vg to cryo-EM structure determination. X-ray crystallography and simple biophysical analysis could provide valuable insights into the newly identified CTCK domain and its functions. The domain is small and compact, making a good target for not only crystallography, but also NMR spectroscopy.

Overall, significant progress has been made in understanding the structure-function relationship of both GbpA and Vg, uncovering new and exciting details of their involvement in different fronts of molecular warfare.

Materials and Methods

Protein parameters

GbpA

```
        10         20         30         40         50         60
MKKQPKMTAI ALILSGISGL AYGHGYVSAV ENGVAEGRVT LCKFAANGTG EKNTHCGAIQ
        70         80         90        100        110        120
YEPQSVEGPD GFPVTGPRDG KIASAESALA AALDEQTADR WVKRPIQAGP QTFEWTFTAN
       130        140        150        160        170        180
HVTKDWKYYI TKPNWNPNQP LSRDAFDLNP FCVVEGNMVQ PPKRVSHECI VPEREGYQVI
       190        200        210        220        230        240
LAVWDVGDTA ASFYNVIDVK FDGNGPVLPD WNPAGQIIPS MDLSIGDTVY TRVFDNDGEN
       250        260        270        280        290        300
PAYRTELKID SETLTKANQW SYALATKINQ TQKQQRAGQL NGDQFVPVYG TNPIYLKEGS
       310        320        330        340        350        360
GLKSVEIGYQ IEAPQPEYSL TVSGLAKEYE IGEQPIQLDL TLEAQGEMSA ELTVYNHHQK
       370        380        390        400        410        420
PLASWSQAMT DGELKSITLE LSEAKAGHHM LVSRIKDRDG NLQDQQTLDF MLVEPQTPPT
       430        440        450        460        470        480
PGDYDFVFPN GLKEYVAGTK VLASDGAIYQ CKPWPYSGYC QQWTSNATQY QPGTGSHWEM

  AWDKR
```

Domain coloring: signal peptide, LPMO, D2, D3, D4/CBM73, flexible linkers

Uniprot ID: Q9KLD5

	$GbpA_{FL}$	$GbpA_{LPMO}$
Molecular mass (Da):	31,254.08	19,774.01
Theoretical pI:	4.81	5.05
Extinction coefficient (L mol^{-1} cm^{-1}):	97,665	36,690

Parameters computed for the protein without the signal peptide and extinction coefficients for all cysteines forming cystines. Computed using the Expasy ProtParam tool.

Vg

```
        10         20         30         40         50         60
MLLLLTLLLF AGTVAADFQH NWQVGNEYTY LVRSRTLTSL GDLSDVHTGI LIKALLTVQA
        70         80         90        100        110        120
KDSNVLAAKV WNGQYARVQQ SMPDGWETEI SDQMLELRDL PISGKPFQIR MKHGLIRDLI
       130        140        150        160        170        180
VDRDVPTWEV NILKSIVGQL QVDTQGENAV KVNSVQVPTD DEPYASFKAM EDSVGGKCEV
       190        200        210        220        230        240
LYDIAPLSDF VIHRSPELVP MPTLKGDGRH MEVIKIKNFD NCDQRINYHF GMTDNSRLEP
       250        260        270        280        290        300
GTNKNGKFFS RSSTSRIVIS ESLKHFTIQS SVTTSKMMVS PRLYDRQNGL VLSRMNLTLA
       310        320        330        340        350        360
KMEKTSKPLP MVDNPESTGN LVYIYNNPFS DVEERRVSKT AMNSNQIVSD NSLSSSEEKL
       370        380        390        400        410        420
KQDILNLRTD ISSSSSSISS SEENDFWQPK PTLEDAPQNS LLPNFVGYKG KHIGKSGKVD
       430        440        450        460        470        480
VINAAKELIF QIANELEDAS NIPVHATLEK FMILCNLMRT MNRKQISELE SNMQISPNEL
       490        500        510        520        530        540
KPNDKSQVIK QNTWTVFRDA ITQTGTGPAF LTIKEWIERG TTKSMEAANI MSKLPKTVRT
       550        560        570        580        590        600
PTDSYIRSFF ELLQNPKVSN EQFLNTAATL SFCEMIHNAQ VNKRSIHNNY PVHTFGRLTS
       610        620        630        640        650        660
KHDNSLYDEY IPFLERELRK AHQEKDSPRI QTYIMALGMI GEPKILSVFE PYLEGKQQMT
       670        680        690        700        710        720
VFQRTLMVGS LGKLTETNPK LARSVLYKIY LNTMESHEVR CTAVFLLMKT NPPLSMLQRM
       730        740        750        760        770        780
AEFTKLDTNR QVNSAVKSTI QSLMKLKSPE WKDLAKKARS VNHLLTHHEY DYELSRGYID
       790        800        810        820        830        840
EKILENQNII THMILNYVGS EDSVIPRILY LTWYSSNGDI KVPSTKVLAM ISSVKSFMEL
       850        860        870        880        890        900
SLRSVKDRET IISAAEKIAE ELKIVPEELV PLEGNLMINN KYALKFFPFD KHILDKLPTL
       910        920        930        940        950        960
ISNYIEAVKE GKFMNVNMLD TYESVHSFPT ETGLPFVYTF NVIKLTKTSG TVQAQINPDF
       970        980        990       1000       1010       1020
AFIVNSNLRL TFSKNVQGRV GFVTPFEHRH FISGIDSNLH VYAPLKISLD VNTPKGNMQW
      1030       1040       1050       1060       1070       1080
KIWPMKGEEK SRLFHYSVVP FVSNHDILNL RPLSMEKGTR PMIPDDNTSL ALPKNEGPFR
      1090       1100       1110       1120       1130       1140
LNVETAKTNE EMWELIDTEK LTDRLPYPWT MDNERYVKVD MYMNLEGEQK DPVIFSTSFD
      1150       1160       1170       1180       1190       1200
SKVMTRPDTD SENWTPKMMA VEPTDKQANS KTRRQEMMRE AGRGIESAKS YVVDVRVHVP
      1210       1220       1230       1240       1250       1260
GESESETVLT LAWSESNVES KGRLLGFWRV EMPRSNADYE VCIGSQIMVS PETLLSYDEK
      1270       1280       1290       1300       1310       1320
MDQKPKMDFN VDIRYGKNCG KGERIDMNGK LRQSPRLKEL VGATSIIKDC VEDMKRGNKI
      1330       1340       1350       1360       1370       1380
LRTCQKAVVL SMLLDEVDIS MEVPSDALIA LYSQGLFSLS EIDNLDVSLD VSNPKNAGKK
      1390       1400       1410       1420       1430       1440
KIDVRAKLNE YLDKADVIVN TPIMDAHFKD VKLSDFGFST EDILDTADED LLINNVFYED
      1450       1460       1470       1480       1490       1500
ETSCMLDKTR AQTFDGKDYP LRLGPCWHAV MTTYPRINPD NHNEKLHIPK DKSVSVLSRE
      1510       1520       1530       1540       1550       1560
NEAGQKEVKV LLGSDKIKFV PGTTSQPEVF VNGEKIVVSR NKAYQKVEEN EIIFEIYKMG
      1570       1580       1590       1600       1610       1620
DRFIGLTSDK FDVSLALDGE RVMLKASEDY RYSVRGLCGN FDHDSTNDFV GPKNCLFRKP
      1630       1640       1650       1660       1670       1680
EHFVASYALI SNQCEGDSLN VAKSLQDHDC IRQERTQQRN VISDSESGRL DTEMSTWGYH
      1690       1700       1710       1720       1730       1740
HNVNKHCTIH RTQVKETDDK ICFTMRPVVS CASGCTAVET KSKPYKFHCM EKNEAAMKLK
      1750       1760       1770
KRIEKGANPD LSQKPVSTTE ELTVPFVCKA
```

Region coloring: signal peptide, N-sheet, PolySerine, α-Helical, C-sheet, A-sheet, vWFD, CTCK, flexible linkers

Uniprot ID: Q868N5

Molecular mass: 199,405.58 Da

Theoretical pI: 6.25

Extinction coefficient: 173,260 L mol^{-1} cm^{-1}

Parameters computed for the protein without the signal peptide and extinction coefficients for all cysteines forming cystines. Computed using the Expasy ProtParam tool, therefore excluding PTMs.

Media recipes

LB media (for 100 mL):

- 0.5 g yeast extract (VWR-J850)
- 1 g tryptone (VWR-84610)
- 1 g NaCl (VWR-27788.460)

LB-v2 media (for 100 mL)

- 0.5 g yeast extract (VWR-J850)
- 1 g tryptone (VWR-84610)
- 2.2 g NaCl (VWR-27788.460)
- 0.031 g KCl (Sigma/Merck-60132)
- 0.220 g $MgCl_2$ – weight for the anhydrous salt (Sigma/Merck-63064)

Trace element solution 1000x (for 100mL)

- 0.6 g $FeSO_4 \cdot 7H_2O$ (Sigma/Merck-215422)
- 0.6 g $CaCl_2 \cdot 2H_2O$ (Sigma/Merck-21097)
- 0.12 g $MnCl_2 \cdot 4H_2O$ (Sigma/Merck-63535)
- 0.08 g $CoCl_2 \cdot 6H_2O$ (Sigma/Merck-255599)
- 0.07 g $ZnSO_4 \cdot 7H_2O$ (Sigma/Merck-Z4750)
- 0.03 g $CuCl_2 \cdot 2H_2O$ (Sigma/Merck-459097)
- 0.02 g H_3BO_4 (Sigma/Merck-B9645)
- 0.025g $(NH4)_6Mo_7O_{24} \cdot 4H_2O$ (Sigma/Merck-O9878)
- 17 mM EDTA (Sigma/Merck-E5134)

$M9_{glyc+}$ (for 1L)

- 19 g K_2HPO_4 (VWR-26932.290)
- 5 g KH_2PO_4 (VWR-26925.295)
- 9 g Na_2HPO_4 (VWR-28026.292)
- 2.4 g K_2SO_4 (VWR-P9458)
- 5.0 g NH_4Cl (Sigma/Merck-0621)
- 18 g glycerol – autoclaved separately (VWR-24388.295)
- 10 mM $MgCl_2$ – sterile filtered separately and added from stock solution (Sigma/Merck-63064)
- 1x MEM vitamins (Sigma/Merck-M6895)
- 1x Trace element solution

$M9_{max}$ (for 1L)

- 21.3 g Na_2HPO_4 (VWR-28026.292)
- 10.2 g KH_2PO_4 (VWR-26925.295)
- 0.58 g NaCl (VWR-27810.364)
- 1.5 g NH_4Cl (Sigma/Merck-0621)
- 18 g glycerol – autoclaved separately (VWR-24388.295)
- 10 mM $CaCl_2$ – sterile filtered separately and added from stock solution (Sigma/Merck-21097)
- 5 mM $MgSO_4$ – sterile filtered separately and added from stock solution (Sigma/Merck-25164.265)
- 1x MEM vitamins (Sigma/Merck-M6895)
- 1x Trace element solution

For the production of deuterated minimal media deuterium oxide was used as solvent (Euroisotop:D214L) together with anhydrous $MgCl_2$ (Sigma/Merch-M8266).

Sample production

GbpA constructs

$GbpA_{FL}$ and $GbpA_{LPMO}$ were cloned in either pET-22b(+) or pET-26b(+) vectors, both producing the same final protein products. The pET-22b(+) vectors were provided by the Vaaje-Kolstad laboratory (Norwegian University of Life Sciences) and are described by Wong *et al.* [10]. Briefly, the sequences were cloned from the genomic DNA of *V. cholerae* from the N1RB3 strain between the NdeI and XhoI restriction sites. The construct corresponding for $GbpA_{LPMO}$ was obtained by site-directed mutagenesis with the addition of two stop codons at position 203. The natural secretion tag of GbpA was thus kept and responsible for periplasmic export when expressed in *E. coli* expression systems. The pET-26(+) expression constructs were produced by Genscript. Here, the *E. coli* codon optimized sequences coding for the residues 24-485 for $GbpA_{FL}$ and 24-202 for $GbpA_{LPMO}$ (UniProt ID: Q9KLD5) were cloned between the restriction sites NcoI and XhoI. This introduced the PelB leader sequence present from the vector immediately upstream of the N-terminus of the protein, signaling for periplasmatic export. The C-terminal His-tag present in the vector was omitted by the inclusion of a stop codon at the end of the insert.

GbpA variants

$GbpA_{FL}$ variants were produced using the NEB Q5 Site-Directed Mutagenesis Kit for the construct in the pET-26b(+) vector. The primers were designed using the NEBaseChanger online tool.

$GbpA_{FL}$D70A:	Forward	5´– GAAGGTCCGGcTGGTTTCCCG –3´
	Reverse	5´– AACGCTTTGCGGTTCGTAC –3´
$GbpA_{FL}$D70K:	Forward	5´– TGAAGGTCCGaagGGTTTCCCGG –3´
	Reverse	5´– ACGCTTTGCGGTTCGTAC –3´

The standard protocol supplied by NEB was used (E0554). During amplification cycles, annealing was performed at 67 °C for 20 s while elongation lasted 90 s. Successful mutagenesis was assessed by sequencing.

GbpA production in *E. coli*

pET-22b(+) constructs were transformed into BL21-star (DE3) cells and selection was achieved with 37 μg/mL sodium ampicillin (Applichem-A0839), while the pET-26b(+) constructs were transformed in BL21 (DE3) cells and selection was achieved with 50 μg/mL kanamycin sulphate (Applichem-A1493). Both bacterial strains were grown following the same protocol, both for $GbpA_{FL}$ and $GbpA_{LPMO}$. Interestingly, GbpA production yields were found to be higher in minimal media compared to rich media [109].

The cells were scratched off a glycerol stock or an individual plated colony in a 2.5 mL LB medium pre-culture for 6 h (220 rpm, 37 °C). Then, 200μL of the pre-culture were used to inoculate a growth culture of 25 mL $M9_{glyc+}$ medium, which was incubated overnight for 16 h (120 rpm, 37 °C). Next, 225 mL of fresh $M9_{glyc+}$ medium were added to the culture, and incubation (120 rpm, 37 °C) continued until an OD_{600} between 2 and 3 was reached. The culture was induced with 1 mM IPTG (VWR-43714-4N) and incubated for 20 h to allow protein expression (120 rpm, 25 °C). Scaling up was performed by parallel growth of several 225 mL cultures. Normal yields of protein are around 30 mg/L for $GbpA_{FL}$ and 7 mg/L for $GbpA_{LPMO}$.

GbpA production in *V. natriegens*

Only the GbpA constructs in the pET-26b(+) vectors were transformed into *V. natriegens* strain Vmax™ X2 following the manufacturer guidelines. Selection was achieved using 200 μg/mL kanamycin sulphate (Applichem-A1493). Interestingly, when produced in Vmax™ X2, GbpA was secreted in a process analogous to that of *V. cholerae* [110], most likely through the T2SC. This facilitates the harvesting process and increases purity by avoiding periplasmatic extraction. However, it complicates the scaling-up of the protocol since the purification process starts with larger volumes of sample compared to periplasmic extraction when GbpA is produced in *E. coli*.

The cells were scratched off a glycerol stock or an individual plated colony in a 2.5 mL LB-v2 medium pre-culture for 4 h (220 rpm, 30 °C). Then, 200μL of the pre-culture were used to inoculate a growth culture of 10 mL $M9_{max}$ medium, which was incubated overnight for 16 h (120 rpm, 30 °C). Next, 90 mL of fresh $M9_{max}$ medium was added to the culture and incubation (120 rpm, 30 °C) continued for 3 h. The culture was induced with 1 mM IPTG (VWR-43714-4N) and incubated for 20 h to allow protein expression (120 rpm, 30 °C). Scaling up was performed by parallel growth of several 100 mL cultures. Normal yields of protein are around 50 mg/L for $GbpA_{FL}$ and 12 mg/L for $GbpA_{LPMO}$.

Cells were harvested at 18,500 x g for 30 min at 4 °C for $GbpA_{LPMO}$ and 8,500 x g for 30 min at 4 °C for $GbpA_{FL}$. Both $GbpA_{FL}$ and $GbpA_{LPMO}$ are secreted. After harvesting, the GbpA containing supernatants were diluted 7x in buffer without salt (20mM Tris-HCl pH 8.0 (Sigma/Merck-93352) to decrease the ionic strength.

Deuterated $GbpA_{LPMO}$ production in *E. coli*

The pET-26b(+) construct containing the coding sequence of $GbpA_{LPMO}$ was transformed in BL21 (DE3) cells. For expression, cells were scratched off a glycerol stock or an individual plated colony and used to inoculate a 2.5 mL pre-culture of LB medium for 3 h (220 rpm, 37 °C). Then, 200µL of the pre-culture were used to inoculate a deuterated pre-culture of 2.5 mL LB deuterated medium, which was incubated for 6 h (220 rpm, 37 °C). Next, all the deuterated pre-culture was added to 25 mL of deuterated $M9_{glyc+}$ medium, which was incubated overnight for 16 h (120 rpm, 37 °C). In the morning, 225 mL of fresh deuterated $M9_{glyc+}$ medium were added to the culture and incubation (120 rpm, 37 °C) continued until an OD_{600} between 2 and 3 was reached 25 h later. The culture was then induced with 1 mM isopropyl-β-D-thiogalactoside (IPTG) and incubated for 20 h to allow expression (120 rpm, 25 °C). Scaling up was performed by parallel growth of several 225 mL cultures. Normal yields of protein are around 3 mg per liter of culture. Since the yields obtained were not high enough for neutron crystallography, scaling up of the protocol was performed by international facilities as described in **Manuscript II**.

GbpA periplasmatic extraction

$GbpA_{FL}$ and $GbpA_{LPMO}$ were harvested from the *E. coli* periplasm using an osmotic shock. First, the bacterial cells were harvested by centrifugation (30 min, 10,000 x g, 4 °C). The pellet was resuspended in an hypertonic solution (4-5 mL per gram of pelleted cells) containing 25% w/v sucrose (VWR-27483.363), 20 mM Tris-HCl pH 8.0 (Sigma/Merck-93352) and 5 mM ethylenediaminetetraacetic acid (EDTA; Sigma/Merck-E5134), and incubated for 30 min at 4 °C. Thereafter, the cells were once again pelleted (30 min, 10,000 x g, 4 °C) and resuspended an hypotonic solution (4-5 mL per gram of pelleted cells) containing 5 mM MgCl2 (Sigma/Merck-63064), 1 mM phenylmethylsulfonyl fluoride (PMSF; Sigma/Merck-78830) and 0.1 mg/mL lysozyme (Sigma/Merck-L4919). The suspension was incubated for 30 min on at 4 °C and centrifuged again (30 min, 10,000 x g, 4 °C). The supernatants resulting from the two incubations contained GbpA, and were subjected to further purification.

GbpA purification

Generally, $GbpA_{FL}$ and $GbpA_{LPMO}$ can be stored at -20 °C without the need for flash freezing and are stable at room temperature for days. Proteins were purified by anion-exchange chromatography (AEX) and size-exclusion chromatography (SEC). AEX was performed in a HiTrap Q HP column (Cytiva) equilibrated in 20 mM Tris-HCl pH 8.0 (Sigma/Merck-93352) and 50 mM NaCl (VWR-27810.364). Elution was achieved with a salt gradient up to 400 mM NaCl. SEC was performed on Superdex 75 Increase 10/300 or Superdex 200 Increase 10/300 columns equilibrated in 20 mM Tris-HCl pH 8.0 (Sigma/Merck-93352) and 50 mM NaCl (VWR-27810.364).

GbpA copper saturation

Copper saturation of $GbpA_{LPMO}$ was achieved by incubation of GbpA for 20 min at 4 °C with a 3x molar excess of $CuCl_2$. Further desalting was achieved by using a HiTrap desalting column (Cytiva). The protein was copper saturated prior to crystallization and activity assays (not described in this thesis, but part of **Manuscript II**).

Vg

Vg was purified from the hemolymph of worker bees by our collaborators at the Amdam lab (Norwegian University of Life Sciences) as described in **Manuscript V** and **Manuscript VI**.

Methods for the biophysical characterization of Vg using BN-PAGE and SEC are described in **Manuscript V**.

Protein crystallization

Crystallization of $GbpA_{LPMO}$

$GbpA_{LPMO}$ crystals were produced in conditions corresponding to the condition B7 from the Structure I and II screen (Molecular Dimensions). More details can be found in the Methods section of **Manuscript I**.

$GbpA_{LPMO}$ crystals for high-pressure freezing and soaking studies were produced by sitting drop in MRC MAXI 48-well Plates (SwissCl) in conditions corresponding to the condition G2 from the Structure I and II screen (Molecular Dimensions). $GbpA_{LPMO}$ was concentrated to 15 mg/mL and 1 µL of the concentrated protein was mixed with 1 µL of crystallization solution containing 200 mM MES pH 6.5 (Sigma/Merck-M825), 200 mM $(NH_4)_2SO_4$ (Sigma/Merck-A4915), 30 % w/v PEG 5000 MME (Sigma/Merck-81323). Crystals take time to appear but grow in just a couple days. They reach their final size after 10-12 days at 20 °C.

High-pressure freezing and soaking studies of $GbpA_{LPMO}$ crystals

For soaking, crystals were placed for a few seconds in solutions containing crystallization solutions supplemented with glycerol 10 % w/v for cryoprotection. In addition, the solutions contained 100 mM H_2O_2 (Sigma/Merck-51681-3) or 100 mM sodium azide (Sigma/Merck-71289). The crystals were flash-cooled in liquid N_2 immediately after cryoprotection.

High-pressure freezing in O_2 was performed at the High-Pressure Freezing Laboratory (HPMX) at European Synchrotron Radiation Facility (ESRF, Grenoble) with the help of Philippe Carpentier. The system used allowed for the exposure of $GbpA_{LPMO}$ crystals to pressurized O_2 (50 bar) and fast cryocooling in liquid O_2, allowing the freezing of crystals without cryoprotectant.

Optimization of $GbpA_{LPMO}$ crystals for neutron crystallography

Described in the Results section.

In crystallo Raman spectroscopy

Raman spectroscopy on $GbpA_{LPMO}$ crystals was performed at the *in crystallo* Optical Spectroscopy (icOS lab) at the European Synchrotron Radiation Facility (ESRF, Grenoble) with the help of Philippe Carpentier and Gabriele Cordara. An excitation source at 785 nm was used for the system described in Carpentier et al. [111]. Hans-Petter Hersleth helped design the experiments.

X-ray crystallography

GbpA crystal characterization

Details on data collection and refinement for $GbpA_{LPMO}$ are described in the Methods section of **Manuscript I**. In addition, the characterization of anomalous scatterers (zinc and copper) in the crystal using datasets collected at the K absorption edges of these metals are also included in **Manuscript I**.

$GbpA_{LPMO}$ soaking and high-pressure freezing experiments

Details on data collection and refinement for the $GbpA_{LPMO}$ soaking and high-pressure freezing experiments are summed up in **Table 1** in the Results section. Structures were solved by molecular replacement using and PHASER [112] from the CCP4 program suite [113]. The LPMO domain of GbpA from the structure solved by Wong *et al.* [10] (PDB ID:2XWX) was used as a search model. Refinement is still preliminary and at different stages for different datasets, with only the protein backbone and copper ions built in accurately in all datasets. REFMAC5 [114] was used iteratively together with manual real-space refinement. Water molecules were included later with Coot [115] using the "Find waters" tool with default parameters and not manually inspected for all the datasets. Occupancies have not yet been refined. Angles and distances were measured in Pymol (Schrödinger). Radiation doses were calculated with RADDOSE [116].

SAXS

Details on the SAXS data collection and analysis can be found in the Methods section of **Manuscript IV**.

Negative-Stain EM of GbpA-Chitin

β-chitin fibers were kindly provided the Vaaje-Kolstad laboratory (Norwegian University of Life Sciences) and originally bought from France Chitine. 10 mL of 3 mg/mL chitin were prepared by sonicating it in 20 mM acetic acid pH 3.2 (Sigma/Merck-1.00063). Sonication was performed in a Q500 Sonicator (Qsonica) using the thinnest tip at 30% intensity in 3/3 sec on/off pulses until the sample was completely suspended (around 30 min). Chitin was dialyzed overnight against 20 mM acetic acid pH 5.0 (Sigma/Merck-1.00063) and sonicated again to ensure it remained suspended. A few microliters of concentrated $GbpA_{FL}$ or $GbpA_{LPMO}$ in 20 mM Tris-HCl pH 8.0 (Sigma/Merck-93352) and 50 mM NaCl (VWR-27810.364) were added to 200 μL of the chitin suspension to a final concentration of 2 mg/mL of the protein and mixed thoroughly with the use of a pipette. Staining and imaging was performed immediately thereafter.

Samples were applied to carbon-coated TEM grids and blotted. Immediately, they were washed twice in water and stained with uranyl acetate 1 % for a few seconds. Imaging was done in a JEM-1400Plus microscope operating at 120 kV. Images were processed using ImageJ [105].

Cryo-EM of Vg

Grid freezing, data collection, data processing and model building for Vg are described in detail in the Methods section of **Manuscript VI**.

The figures in this thesis have been produced using Adobe Illustrator, pyMOL (Schrödinger), UCSF Chimera X [117], ImageJ [105] and Matplotlib [118] for python.

Abbreviations

AA	atomic absorption spectroscopy
a.u.	asymmetric unit
AEX	anion-exchange chromatography
apoB	apolipoprotein B
BN-PAGE	blue native polyacrylamide gel electrophoresis
CBM	carbohydrate-binding module
CCD	colony collapse disorder
ChiRP	chitin regulated pili
Cryo-EM	cryogenic electron microscopy
Cryo-ET	cryogenic electron tomography
CTCK	C-terminal cystine knot domain
DAMP	danger associated molecular patterns
DFT	density functional theory
EDS	energy dispersive X-ray spectroscopy
EM	electron microscopy
fbVg	fat-body vitellogenin
HAA	hydrogen atom abstraction
GbpA	GlcNAc binding protein A
$GbpA_{FL}$	full-length GbpA
$GbpA_{LPMO}$	LPMO domain of GbpA (equivalent to domain 1)
$GbpA_{D1}$	domain 1 of GbpA (equivalent to the LPMO)
$GbpA_{D2}$	domain 2 of GbpA
$GbpA_{D3}$	domain 3 of GbpA
$GbpA_{D4}$	domain 4 of GbpA
GH	glycosyl hydrolase
GlcNAc	*N*-acetylglucosamine
ICP-MS	inductively coupled plasma mass spectrometry
IPTG	isopropyl-β-D-thiogalactoside
JH	juvenile hormone
LLTP	large lipid transfer protein superfamily
LPMO	lytic polysaccharide monooxygenase
Mam7	outer membrane adhesion factor multivalent molecule 7
MES	2-(*N*-morpholino)ethanesulfonic acid
MME	monomethyl ether
MS	mass spectrometry
MSHA	mannose sensitive hemagglutinin
MTP	triglyceride transfer protein
NMR	nuclear magnetic resonance
PAGE	polyacrylamide gel electrophoresis
PAMP	pathogen-associated molecular pattern
PEG	polyethylene glycol
PTM	post-translational modification
R.M.S.D.	root mean square deviation/difference

SANS	small-angle neutron scattering
SAS	small-angle scattering
SAXS	small-angle X-ray scattering
SEC	size-exclusion chromatography
STEM	scanning transmission electron microscope
PMSF	phenylmethylsulfonyl fluoride
TCP	toxin-coregulated pilus
T2SC	type II secretion system
Vg	vitellogenin
vWFD	von Willebrand factor type D domain

Bibliography

(1) WHO. (2020) The top 10 causes of death.

(2) Kanungo, S., Azman, A. S., Ramamurthy, T., Deen, J., and Dutta, S. (2022) Cholera. *Lancet 399*, 1429–1440.

(3) WHO. (2017) Cholera vaccines: WHO position paper – August 2017. *Relev. Epidemiol. Hebd.*

(4) Grant, T. A., Balasubramanian, D., and Almagro-Moreno, S. (2021) JMM Profile: *Vibrio cholerae*: an opportunist of human crises. *J. Med. Microbiol. 70*, 1–3.

(5) Holmner, Å., Mackenzie, A., and Krengel, U. (2010) Molecular basis of cholera blood-group dependence and implications for a world characterized by climate change. *FEBS Lett. 584*, 2548–2555.

(6) Haley, B. J., Chen, A., Grim, C. J., Clark, P., Diaz, C. M., Taviani, E., Hasan, N. A., Sancomb, E., Elnemr, W. M., Islam, M. A., Huq, A., Colwell, R. R., and Benediktsdóttir, E. (2012) *Vibrio cholerae* in a historically cholera-free country. *Environ. Microbiol. Rep. 4*, 381–389.

(7) Julie E. Heggelund, Bjørnestad, V. A., and Krengel, U. (2016) *Vibrio cholerae* and *Escherichia coli* heat-labile enterotoxins and beyond, in *The comprehensive sourcebook of bacterial protein toxins*, pp 195–229.

(8) Almagro-Moreno, S., Pruss, K., and Taylor, R. K. (2015) Intestinal colonization dynamics of *Vibrio cholerae*. *PLoS Pathog. 11*, 1–11.

(9) Kirn, T. J., Jude, B. A., and Taylor, R. K. (2005) A colonization factor links *Vibrio cholerae* environmental survival and human infection. *Nature 438*, 863–866.

(10) Wong, E., Vaaje-Kolstad, G., Ghosh, A., Hurtado-Guerrero, R., Konarev, P. V., Ibrahim, A. F. M., Svergun, D. I., Eijsink, V. G. H., Chatterjee, N. S., and van Aalten, D. M. F. (2012) The *Vibrio cholerae* colonization factor GbpA possesses a modular structure that governs binding to different host surfaces. *PLoS Pathog. 8*, 1–12.

(11) Bhowmick, R., Ghosal, A., Das, B., Koley, H., Saha, D. R., Ganguly, S., Nandy, R. K., Bhadra, R. K., and Chatterjee, N. S. (2008) Intestinal adherence of *Vibrio cholerae* involves a coordinated interaction between colonization factor GbpA and mucin. *Infect. Immun. 76*, 4968–4977.

(12) Banse, A. V., VanBeuge, S., Smith, T. J., Logan, S. L., and Guillemin, K. (2023) Secreted *Aeromonas* GlcNAc binding protein GbpA stimulates epithelial cell proliferation in the zebrafish intestine. *Gut Microbes 15*, 1–15.

(13) Vezzulli, L., Guzmán, C. A., Colwell, R. R., and Pruzzo, C. (2008) Dual role colonization factors connecting *Vibrio cholerae´*s lifestyles in human and aquatic environments open new perspectives for combating infectious diseases. *Curr. Opin. Biotechnol. 19*, 254–259.

(14) Sakib, S. N., Reddi, G., and Almagro-Moreno, S. (2018) Environmental role of pathogenic traits in *Vibrio cholerae*. *J. Bacteriol. 200*.

(15) Lutz, C., Erken, M., Noorian, P., Sun, S., and McDougald, D. (2013) Environmental reservoirs and mechanisms of persistence of *Vibrio cholerae*. *Front. Microbiol. 4*, 1–15.

(16) Almagro-Moreno, S., and Taylor, R. K. (2013) Cholera: Environmental reservoirs and impact on disease transmission. *Microbiol. Spectr. 1*.

(17) Hood, M. A., and Winter, P. A. (1997) Attachment of *Vibrio cholerae* under various environmental conditions and to selected substrates. *FEMS Microbiol. Ecol. 22*, 215–223.

(18) Nahar, S., Sultana, M., Naser, M. N., Nair, G. B., Watanabe, H., Ohnishi, M., Yamamoto, S.,

Endtz, H., Cravioto, A., Sack, R. B., Hasan, N. A., Sadique, A., Huq, A., Colwell, R. R., and Alam, M. (2012) Role of shrimp chitin in the ecology of toxigenic *Vibrio cholerae* and cholera transmission. *Front. Microbiol. 2*, 1–8.

(19) Meiborn, K. L., Li, X. B., Nielsen, A. T., Wu, C. Y., Roseman, S., and Schoolnik, G. K. (2004) The *Vibrio cholerae* chitin utilization program. *Proc. Natl. Acad. Sci. U. S. A. 101*, 2524–2529.

(20) Pruzzo, C., Vezzulli, L., and Colwell, R. R. (2008) Global impact of *Vibrio cholerae* interactions with chitin. *Environ. Microbiol. 10*, 1400–1410.

(21) Waldor, M. K., and Mekalanos, J. J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science. 272*, 1910–1913.

(22) Meibom, K. L., Blokesch, M., Dolganov, N. A., Wu, C. Y., and Schoolnik, G. K. (2005) Chitin induces natural competence in *Vibrio cholerae*. *Science. 310*, 1824–1827.

(23) Tamayo, R., Patimalla, B., and Camilli, A. (2010) Growth in a biofilm induces a hyperinfectious phenotype in *Vibrio cholerae*. *Infect. Immun. 78*, 3560–3569.

(24) Colwell, R. R., Huq, A., Islam, M. S., Aziz, K. M. A., Yunus, M., Huda Khan, N., Mahmud, A., Bradley Sack, R., Nair, G. B., Chakraborty, J., Sack, D. A., and Russek-Cohen, E. (2003) Reduction of cholera in Bangladeshi villages by simple filtration. *Proc. Natl. Acad. Sci. U. S. A. 100*, 1051–1055.

(25) Landman, N. H., and Harries, P. J. (2020) Chitin; Formation and diagenesis.

(26) Stauder, M., Vezzulli, L., Pezzati, E., Repetto, B., and Pruzzo, C. (2010) Temperature affects *Vibrio cholerae* O1 El Tor persistence in the aquatic environment via an enhanced expression of GbpA and MSHA adhesins. *Environ. Microbiol. Rep. 2*, 140–144.

(27) Fan, Y., Saito, T., and Isogai, A. (2008) Preparation of chitin nanofibers from squid pen ß-chitin by simple mechanical treatment under acid conditions 1919–1923.

(28) Chiavelli, D. A., Marsh, J. W., and Taylor, R. K. (2001) The mannose-sensitive hemagglutinin of *Vibrio cholerae* promotes adherence to zooplankton. *Appl. Environ. Microbiol. 67*, 3220–3225.

(29) Stauder, M., Huq, A., Pezzati, E., Grim, C. J., Ramoino, P., Pane, L., Colwell, R. R., Pruzzo, C., and Vezzulli, L. (2012) Role of GbpA protein, an important virulence-related colonization factor, for *Vibrio cholerae*'s survival in the aquatic environment. *Environ. Microbiol. Rep. 4*, 439–445.

(30) Loose, J. S. M., Forsberg, Z., Fraaije, M. W., Eijsink, V. G. H., and Vaaje-Kolstad, G. (2014) A rapid quantitative activity assay shows that the *Vibrio cholerae* colonization factor GbpA is an active lytic polysaccharide monooxygenase. *FEBS Lett. 588*, 3435–3440.

(31) Vaaje-Kolstad, G., Westereng, B., Horn, S. J., Liu, Z., Zhai, H., Sørlie, M., and Eijsink, V. G. H. (2010) An oxidative enzyme boosting the enzyatic conversion of recalcitrant polysaccharides. *Science. 219*, 219–223.

(32) Drula, E., Garron, M., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022) The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res. 50*, 571–577.

(33) Vaaje-Kolstad, G., Forsberg, Z., Loose, J. S., Bissaro, B., and Eijsink, V. G. (2017) Structural diversity of lytic polysaccharide monooxygenases. *Curr. Opin. Struct. Biol. 44*, 67–76.

(34) Forsberg, Z., and Courtade, G. (2022) On the impact of carbohydrate-binding modules (CBMs) in lytic polysaccharide monooxygenases (LPMOs). *Essays Biochem. 0*, 561–574.

(35) Tamburrini, K. C., Terrapon, N., Lombard, V., Bissaro, B., Longhi, S., and Berrin, J. G. (2021) Bioinformatic analysis of lytic polysaccharide monooxygenases reveal the pan-families occurrence of intrinsically disordered C-terminal extensions. *Biomolecules 11*.

(36) Chylenski, P., Bissaro, B., Sørlie, M., Røhr, Å. K., Várnai, A., Horn, S. J., and Eijsink, V. G. H. (2019) Lytic polysaccharide monooxygenases in enzymatic processing of lignocellulosic biomass. *ACS*

Catal. 9, 4970–4991.

(37) Shah, F., Nicolás, C., Bentzer, J., Ellström, M., Smits, M., Rineau, F., Canbäck, B., Floudas, D., Carleer, R., Lackner, G., Braesel, J., Hoffmeister, D., Henrissat, B., Ahrén, D., Johansson, T., Hibbett, D. S., Martin, F., Persson, P., and Tunlid, A. (2016) Ectomycorrhizal fungi decompose soil organic matter using oxidative mechanisms adapted from saprotrophic ancestors. *New Phytol. 209*, 1705–1719.

(38) Blanco-Ulate, B., Morales-Cruz, A., Amrine, K. C. H., Labavitch, J. M., Powell, A. L. T., and Cantu, D. (2014) Genome-wide transcriptional profiling of *Botrytis cinerea* genes targeting plant cell walls during infections of different hosts. *Front. Plant Sci. 5*, 1–16.

(39) Polonio, Á., Fernández-Ortuño, D., de Vicente, A., and Pérez-García, A. (2021) A haustorial-expressed lytic polysaccharide monooxygenase from the cucurbit powdery mildew pathogen *Podosphaera xanthii* contributes to the suppression of chitin-triggered immunity. *Mol. Plant Pathol. 22*, 580–601.

(40) Askarian, F., Uchiyama, S., Masson, H., Sørensen, H. V., Golten, O., Bunæs, A. C., Mekasha, S., Røhr, Å. K., Kommedal, E., Ludviksen, J. A., Arntzen, M., Schmidt, B., Zurich, R. H., van Sorge, N. M., Eijsink, V. G. H., Krengel, U., Mollnes, T. E., Lewis, N. E., Nizet, V., and Vaaje-Kolstad, G. (2021) The lytic polysaccharide monooxygenase CbpD promotes *Pseudomonas aeruginosa* virulence in systemic infection. *Nat. Commun. 12*, 1–19.

(41) Kjaergaard, C. H., Qayyum, M. F., Wong, S. D., Xu, F., Hemsworth, G. R., Walton, D. J., Young, N. A., Davies, G. J., Walton, P. H., Johansen, K. S., Hodgson, K. O., Hedman, B., and Solomon, E. I. (2014) Spectroscopic and computational insight into the activation of O_2 by the mononuclear Cu center in polysaccharide monooxygenases. *Proc. Natl. Acad. Sci. U. S. A. 111*, 8797–8802.

(42) Caldararu, O., Oksanen, E., Ryde, U., and Hedegård, E. D. (2019) Mechanism of hydrogen peroxide formation by lytic polysaccharide monooxygenase. *Chem. Sci. 10*, 576–586.

(43) Bissaro, B., Røhr, Å. K., Müller, G., Chylenski, P., Skaugen, M., Forsberg, Z., Horn, S. J., Vaaje-Kolstad, G., and Eijsink, V. G. H. (2017) Oxidative cleavage of polysaccharides by monocopper enzymes depends on H_2O_2. *Nat. Chem. Biol. 13*, 1123–1128.

(44) Hedegård, E. D., and Ryde, U. (2018) Molecular mechanism of lytic polysaccharide monooxygenases. *Chem. Sci. 9*, 3866–3880.

(45) Frommhagen, M., Koetsier, M. J., Westphal, A. H., Visser, J., Hinz, S. W. A., Vincken, J. P., Van Berkel, W. J. H., Kabel, M. A., and Gruppen, H. (2016) Lytic polysaccharide monooxygenases from *Myceliophthora thermophila* C1 differ in substrate preference and reducing agent specificity. *Biotechnol. Biofuels 9*, 1–17.

(46) Bacik, J. P., Mekasha, S., Forsberg, Z., Kovalevsky, A. Y., Vaaje-Kolstad, G., Eijsink, V. G. H., Nix, J. C., Coates, L., Cuneo, M. J., Unkefer, C. J., and Chen, J. C. H. (2017) Neutron and atomic resolution X-ray structures of a lytic polysaccharide monooxygenase reveal copper-mediated dioxygen binding and evidence for N-terminal deprotonation. *Biochemistry 56*, 2529–2532.

(47) O'Dell, W. B., Agarwal, P. K., and Meilleur, F. (2017) Oxygen activation at the active site of a fungal lytic polysaccharide monooxygenase. *Angew. Chemie - Int. Ed. 56*, 767–770.

(48) Schröder, G. C., O'Dell, W. B., Webb, S. P., Agarwal, P. K., and Meilleur, F. (2022) Capture of activated dioxygen intermediates at the copper-active site of a lytic polysaccharide monooxygenase. *Chem. Sci. 118*.

(49) Tandrup, T., Lo Leggio, L., and Meilleur, F. (2023) Joint X-ray/neutron structure of *Lentinus similis* AA9_A at room temperature. *Acta Crystallogr. Sect. F Struct. Biol. Commun. 79*, 1–7.

(50) Banerjee, S., Muderspach, S. J., Tandrup, T., Frandsen, K. E. H., Singh, R. K., Ipsen, J. Ø., Hernández-Rollán, C., Nørholm, M. H. H., Bjerrum, M. J., Johansen, K. S., and Lo Leggio, L. (2022) Protonation state of an important histidine from high resolution structures of lytic polysaccharide monooxygenases. *Biomolecules 12*.

(51) Frandsen, K. E. H., Simmons, T. J., Dupree, P., Poulsen, J. C. N., Hemsworth, G. R., Ciano, L., Johnston, E. M., Tovborg, M., Johansen, K. S., Von Freiesleben, P., Marmuse, L., Fort, S., Cottaz, S., Driguez, H., Henrissat, B., Lenfant, N., Tuna, F., Baldansuren, A., Davies, G. J., Lo Leggio, L., and Walton, P. H. (2016) The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases. *Nat. Chem. Biol. 12*, 298–303.

(52) Simmons, T. J., Frandsen, K. E. H., Ciano, L., Tryfona, T., Lenfant, N., Poulsen, J. C., Wilson, L. F. L., Tandrup, T., Tovborg, M., Schnorr, K., Johansen, K. S., Henrissat, B., Walton, P. H., Lo Leggio, L., and Dupree, P. (2017) Structural and electronic determinants of lytic polysaccharide monooxygenase reactivity on polysaccharide substrates. *Nat. Commun. 8*.

(53) Brander, S., Tokin, R., Ipsen, J., Jensen, P. E., Hernández-Rollán, C., Nørholm, M. H. H., Lo Leggio, L., Dupree, P., and Johansen, K. S. (2021) Scission of glucosidic bonds by a *Lentinus similis* lytic polysaccharide monooxygenases is strictly dependent on H2O2 while the oxidation of saccharide products depends on O2. *ACS Catal. 11*, 13848–13859.

(54) Vaaje-Kolstad, G., Horn, S. J., Van Aalten, D. M. F., Synstad, B., and Eijsink, V. G. H. (2005) The non-catalytic chitin-binding protein CBP21 from *Serratia marcescens* is essential for chitin degradation. *J. Biol. Chem. 280*, 28492–28497.

(55) Vaaje-Kolstad, G., Houston, D. R., Riemen, A. H. K., Eijsink, V. G. H., and van Aalten, D. M. F. (2005) Crystal structure and binding properties of the *Serratia marcescens* chitin-binding protein CBP21. *J. Biol. Chem. 280*, 11313–11319.

(56) Zhou, Y., Wannapaiboon, S., Prongjit, M., and Pornsuwan, S. (2023) Structural and binding studies of a new chitin-active AA10 lytic polysaccharide monooxygenase from the marine bacterium *Vibrio campbellii*. *Acta Crystallogr. Sect. D Biol. Crystallogr. D79*, 479–497.

(57) Madland, E., Forsberg, Z., Wang, Y., Lindorff-Larsen, K., Niebisch, A., Modregger, J., Eijsink, V. G. H., Aachmann, F. L., and Courtade, G. (2021) Structural and functional variation of chitin-binding domains of a lytic polysaccharide monooxygenase from *Cellvibrio japonicus*. *J. Biol. Chem. 297*, 101084.

(58) 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., Nikolov, 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., and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature 596*, 583–589.

(59) Smolenaars, M. M. W., Madsen, O., Rodenburg, K. W., and Van Der Horst, D. J. (2007) Molecular diversity and evolution of the large lipid transfer protein superfamily. *J. Lipid Res. 48*, 489–502.

(60) Byrne, B. M., Gruber, M., and Ab, G. (1989) The evolution of egg yolk proteins. *Prog. Biophys. Mol. Biol. 53*, 33–69.

(61) Trenczek, T., and Engels, W. (1986) Occurrence of vitellogenin in drone honeybees (*Apis mellifica*). *Int. J. Invertebr. Reprod. Dev. 10*, 307–311.

(62) Piulachs, M. D., Guidugli, K. R., Barchuk, A. R., Cruz, J., Simões, Z. L. P., and Bellés, X. (2003) The vitellogenin of the honey bee, Apis mellifera: Structural analysis of the cDNA and expression studies. *Insect Biochem. Mol. Biol. 33*, 459–465.

(63) Rakhshandehroo, M., Gijzel, S. M. W., Siersbæk, R., Broekema, M. F., De Haar, C., Schipper, H. S., Boes, M., Mandrup, S., and Kalkhoven, E. (2014) CD1d-mediated presentation of endogenous lipid antigens by adipocytes requires microsomal triglyceride transfer protein. *J. Biol. Chem. 289*, 22128–22139.

(64) Hoeger, U., and Harris, J. R. (2020) Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins.

(65) Zhang, S., Wang, S., Li, H., and Li, L. (2011) Vitellogenin, a multivalent sensor and an

antimicrobial effector. *Int. J. Biochem. Cell Biol. 43*, 303–305.

(66) Li, Z., Zhang, S., and Liu, Q. (2008) Vitellogenin functions as a multivalent pattern recognition receptor with an opsonic activity. *PLoS One 3*, 4–10.

(67) Liu, Q. H., Zhang, S. C., Li, Z. J., and Gao, C. R. (2009) Characterization of a pattern recognition molecule vitellogenin from carp (*Cyprinus carpio*). *Immunobiology 214*, 257–267.

(68) Li, Z., Zhang, S., Zhang, J., Liu, M., and Liu, Z. (2009) Vitellogenin is a cidal factor capable of killing bacteria via interaction with lipopolysaccharide and lipoteichoic acid. *Mol. Immunol. 46*, 3232–3239.

(69) Liu, M., Pan, J., Ji, H., Zhao, B., and Zhang, S. (2011) Vitellogenin mediates phagocytosis through interaction with FcγR. *Mol. Immunol. 49*, 211–218.

(70) Sun, C., Hu, L., Liu, S., Gao, Z., and Zhang, S. (2013) Functional analysis of domain of unknown function (DUF) 1943, DUF1944 and von Willebrand factor type D domain (VWD) in vitellogenin2 in zebrafish. *Dev. Comp. Immunol. 41*, 469–476.

(71) Garcia, J., Munro, E. S., Monte, M. M., Fourrier, M. C. S., Whitelaw, J., Smail, D. A., and Ellis, A. E. (2010) Atlantic salmon (*Salmo salar L.*) serum vitellogenin neutralises infectivity of infectious pancreatic necrosis virus (IPNV). *Fish Shellfish Immunol. 29*, 293–297.

(72) Du, X., Wang, X., Wang, S., Zhou, Y., Zhang, Y., and Zhang, S. (2017) Functional characterization of Vitellogenin_N domain, domain of unknown function 1943, and von Willebrand factor type D domain in vitellogenin of the non-bilaterian coral *Euphyllia ancora*. *Dev. Comp. Immunol. 67*, 485–494.

(73) Havukainen, H., Münch, D., Baumann, A., Zhong, S., Halskau, Ø., Krogsgaard, M., and Amdam, G. V. (2013) Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. *J. Biol. Chem. 288*, 28369–28381.

(74) Salmela, H., Amdam, G. V., and Freitak, D. (2015) Transfer of immunity from mother to offspring is mediated via egg-yolk protein vitellogenin. *PLoS Pathog. 11*, 1–12.

(75) Amdam, G. V., Norberg, K., Hagen, A., and Omholt, S. W. (2003) Social exploitation of vitellogenin. *Proc. Natl. Acad. Sci. U. S. A. 100*, 1799–1802.

(76) Harwood, G., Amdam, G., and Freitak, D. (2019) The role of Vitellogenin in the transfer of immune elicitors from gut to hypopharyngeal glands in honey bees (*Apis mellifera*). *J. Insect Physiol. 112*, 90–100.

(77) Marco Antonio, D. S., Guidugli-Lazzarini, K. R., Do Nascimento, A. M., Simões, Z. L. P., and Hartfelder, K. (2008) RNAi-mediated silencing of vitellogenin gene function turns honeybee (*Apis mellifera*) workers into extremely precocious foragers. *Naturwissenschaften 95*, 953–961.

(78) Amdam, G. V., and Omholt, S. W. (2003) The hive bee to forager transition in honeybee colonies: The double repressor hypothesis. *J. Theor. Biol. 223*, 451–464.

(79) Guidugli, K. R., Nascimento, A. M., Amdam, G. V., Barchuk, A. R., Omholt, S., Simões, Z. L. P., and Hartfelder, K. (2005) Vitellogenin regulates hormonal dynamics in the worker caste of a eusocial insect. *FEBS Lett. 579*, 4961–4965.

(80) Nelson, C. M., Ihle, K. E., Fondrk, M. K., Page, R. E., and Amdam, G. V. (2007) The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol. 5*, 0673–0677.

(81) Seehuus, S. C., Norberg, K., Gimsa, U., Krekling, T., and Amdam, G. V. (2006) Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc. Natl. Acad. Sci. U. S. A. 103*, 962–967.

(82) Corona, M., Velarde, R. a, Remolina, S., Moran-lauter, A., Wang, Y., Hughes, K. a, and Robinson, G. E. (2007) Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc.*

Natl. Acad. Sci. U. S. A. 104, 7128–7133.

(83) Süren-Castillo, S., Abrisqueta, M., and Maestro, J. L. (2012) FoxO inhibits juvenile hormone biosynthesis and vitellogenin production in the german cockroach. *Insect Biochem. Mol. Biol. 42*, 491–498.

(84) Philip Kohlmeier, Barbara Feldmeyer, S. F. (2017) *Vitellogenin*6-dependent shifts in social cue responsiveness regulate behavioral task specialization in an ant *4*, 4–7.

(85) Nakamura, A., Yasuda, K., Adachi, H., Sakurai, Y., Ishii, N., and Goto, S. (1999) Vitellogenin-6 is a major carbonylated protein in aged nematode, *Caenorhabditis elegans*. *Biochem. Biophys. Res. Commun. 264*, 580–583.

(86) Ando, S., and Yanagida, K. (1999) Susceptibility to oxidation of copper-induced plasma lipoproteins from japanese eel: Protective effect of vitellogenin on the oxidation of very low density lipoprotein. *Comp. Biochem. Physiol. - C Pharmacol. Toxicol. Endocrinol. 123*, 1–7.

(87) Tufail, M. (2005) Biosynthesis and processing of insect vitellogenins. *Reprod. Biol. Invertebr. Vol. 12, Part B* 17–48.

(88) Dainat, B., Evans, J. D., Chen, Y. P., Gauthier, L., and Neumann, P. (2012) Predictive markers of honey bee colony collapse. *PLoS One 7*.

(89) Tufail, M., and Takeda, M. (2008) Molecular characteristics of insect vitellogenins. *J. Insect Physiol. 54*, 1447–1458.

(90) Sullivan, C. V., and Yilmaz, O. (2018) Vitellogenesis and yolk proteins, fish. *Encycl. Reprod.* Second Edi. Elsevier.

(91) Thompson, J. R., and Banaszak, L. J. (2002) Lipid-protein interactions in lipovitellin. *Biochemistry 41*, 9398–9409.

(92) David S. Auld, Kenneth H. Falchuk, Ke Zhang, M. M. and B. L. V. (1996) X-ray absorption fine structure as a monitor of zinc coordination sites during oogenesis of *Xenopus laevis*. *Proc. Natl. Acad. Sci. U. S. A. 93*, 3227–3231.

(93) Martin, D. J., and Rainbow, P. S. (1998) The kinetics of zinc and cadmium in the haemolymph of the shore crab *Carcinus maenas (L.)*. *Aquat. Toxicol. 40*, 203–231.

(94) Anderson, T. A., Levitt, D. G., and Banaszak, L. J. (1998) The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. *Structure 6*, 895–909.

(95) Raag, R., Appelt, K., Xuong, N. H., and Banaszak, L. (1988) Structure of the lamprey yolk lipid-protein complex lipovitellin-phosvitin at 2.8 Å resolution. *J. Mol. Biol. 200*, 553–569.

(96) Li, A., Sadasivam, M., and Ding, J. L. (2003) Receptor-ligand interaction between vitellogenin receptor (VtgR) and vitellogenin (Vtg), implications on low density lipoprotein receptor and apolipoprotein B/E. The first three ligand-binding repeats of VtgR interact with the amino-terminal region of Vtg. *J. Biol. Chem. 278*, 2799–2806.

(97) Roth, Z., Weil, S., Aflalo, E. D., Manor, R., Sagi, A., and Khalaila, I. (2013) Identification of receptor-interacting regions of vitellogenin within evolutionarily conserved ß-sheet structures by using a peptide array. *ChemBioChem 14*, 1116–1122.

(98) Biterova, E. I., Isupov, M. N., Keegan, R. M., Lebedev, A. A., Sohail, A. A., Liaqat, I., Alanen, H. I., and Ruddock, L. W. (2019) The crystal structure of human microsomal triglyceride transfer protein. *Proc. Natl. Acad. Sci. U. S. A. 116*, 17251–17260.

(99) Javitt, G., Khmelnitsky, L., Albert, L., Bigman, L. S., Elad, N., Morgenstern, D., Ilani, T., Levy, Y., Diskin, R., and Fass, D. (2020) Assembly mechanism of mucin and von Willebrand factor polymers. *Cell 183*, 717-729.e16.

(100) Springer, T. A. (2014) Von Willebrand factor, Jedi knight of the bloodstream. *Blood*.

(101) Havukainen, H., Underhaug, J., Wolschin, F., Amdam, G., and Halskau, Ø. (2012) A vitellogenin polyserine cleavage site: Highly disordered conformation protected from proteolysis by phosphorylation. *J. Exp. Biol. 215*, 1837–1846.

(102) Havukainen, H., Halskau, Ø., Skjaerven, L., Smedal, B., and Amdam, G. V. (2011) Deconstructing honeybee vitellogenin: Novel 40 kDa fragment assigned to its N terminus. *J. Exp. Biol. 214*, 582–592.

(103) Salmela, H., Harwood, G. P., Münch, D., Elsik, C. G., Herrero-Galán, E., Vartiainen, M. K., and Amdam, G. V. (2022) Nuclear translocation of vitellogenin in the honey bee (*Apis mellifera*). *Apidologie 53*, 1–17.

(104) Callaway E. (2022) The Entire protein universe: AI predicts shape of nearly every protein. *Nature 608*, 15–16.

(105) Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods 9*, 671–675.

(106) Blakeley, M. P., Hasnain, S. S., and Antonyuk, S. V. (2015) Sub-atomic resolution X-ray crystallography and neutron crystallography: Promise, challenges and potential. *IUCrJ*.

(107) Chen, P., Bell, J., Eipper, B. A., and Solomon, E. I. (2004) Oxygen activation by the noncoupled binuclear copper site in peptidylglycine α-hydroxylating monooxygenase. Spectroscopic definition of the resting sites and the putative Cu^{II}_{M}-OOH intermediate. *Biochemistry 43*, 5735–5747.

(108) Tandrup, T., Muderspach, S. J., Banerjee, S., Santoni, G., Ipsen, J., Hernández-Rollán, C., Nørholm, M. H. H., Johansen, K. S., Meilleur, F., Lo Leggio, L., and Moffat, K. (2022) Changes in active-site geometry on X-ray photoreduction of a lytic polysaccharide monooxygenase active-site copper and saccharide binding. *IUCrJ 9*, 666–681.

(109) Sørensen, H. V. (2021) Of shellfish and men - Applying X-ray and neutron techniques to surface-active bacterial colonization factors.

(110) Tislevoll, A. (2021) Isotope labelling and interaction studies of the *Vibrio cholerae* colonization factor GbpA.

(111) Carpentier, P., Royant, A., Ohana, J., and Bourgeois, D. (2007) Advances in spectroscopic methods for biological crystals. 2. Raman spectroscopy. *J. Appl. Crystallogr. 40*, 1113–1122.

(112) McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr. 40*, 658–674.

(113) Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr. 67*, 235–242.

(114) Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr. 67*, 355–367.

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

(116) Bury, C. S., Brooks-Bartlett, J. C., Walsh, S. P., and Garman, E. F. (2018) Estimate your dose: RADDOSE-3D. *Protein Sci. 27*, 217–228.

(117) Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., and Ferrin, T. E. (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. 30*, 70–82.

(118) Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng. 9*, 90–95.

Manuscripts

II

ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Perdeuterated GbpA Enables Neutron Scattering Experiments of a Lytic Polysaccharide Monooxygenase

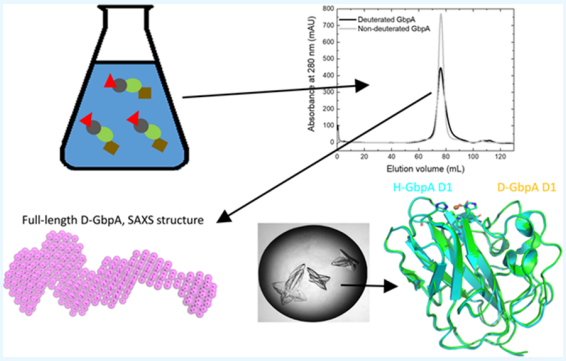
H. V. Sørensen, Mateu Montserrat-Canals, Jennifer S. M. Loose, S. Zoë Fisher, Martine Moulin, Matthew P. Blakeley, Gabriele Cordara, Kaare Bjerregaard-Andersen, and Ute Krengel*

Cite This: *ACS Omega* 2023, 8, 29101–29112

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Lytic polysaccharide monooxygenases (LPMOs) are surface-active redox enzymes that catalyze the degradation of recalcitrant polysaccharides, making them important tools for energy production from renewable sources. In addition, LPMOs are important virulence factors for fungi, bacteria, and viruses. However, many knowledge gaps still exist regarding their catalytic mechanism and interaction with their insoluble, crystalline substrates. Moreover, conventional structural biology techniques, such as X-ray crystallography, usually do not reveal the protonation state of catalytically important residues. In contrast, neutron crystallography is highly suited to obtain this information, albeit with significant sample volume requirements and challenges associated with hydrogen's large incoherent scattering signal. We set out to demonstrate the feasibility of neutron-based techniques for LPMOs using *N*-acetylglucosamine-binding protein A (GbpA) from *Vibrio cholerae* as a target. GbpA is a multifunctional protein that is secreted by the bacteria to colonize and degrade chitin. We developed an efficient deuteration protocol, which yields >10 mg of pure 97% deuterated protein per liter expression media, which was scaled up further at international facilities. The deuterated protein retains its catalytic activity and structure, as demonstrated by small-angle X-ray and neutron scattering studies of full-length GbpA and X-ray crystal structures of its LPMO domain (to 1.1 Å resolution), setting the stage for neutron scattering experiments with its substrate chitin.



INTRODUCTION

Lytic polysaccharide monooxygenases (LPMOs) are surface-active enzymes that introduce breaks in chitin, cellulose, or other polysaccharide layers through a copper-dependent redox reaction. Conserved among all active LPMOs is the histidine-brace motif, where a copper-ion is coordinated by the amino group and a side-chain nitrogen of the N-terminal histidine together with a side-chain nitrogen of a second histidine residue.[1−3] The enzymatic reaction is initiated by the reduction of Cu(II) to Cu(I) by an electron donor and subsequent activation of an oxygen co-substrate. It is unclear if the oxygen co-substrate is O_2 or H_2O_2, with the growing consensus being that H_2O_2 is the more likely co-substrate.[4−6] The natural electron donor likely varies significantly between organisms.[7]

Interestingly, the function of LPMOs often goes beyond polysaccharide conversion. Several LPMOs are part of multi-domain proteins, often with an additional carbohydrate-binding module. Some LPMOs have shown to be key virulence factors for pathogenic bacteria, with importance for colonization. One example is the AA10 class LPMO, *N*-acetylglucosamine-binding protein A (GbpA) of *Vibrio cholerae*. GbpA is a four-domain protein with chitin affinity of the first and fourth domains, and LPMO activity in the first domain.[8,9] This domain in addition exhibits affinity for human intestinal mucin,[8,10] a feature likely important for the colonization of the small intestine by *V. cholerae*.

While the structures of several LPMOs have been determined to atomic resolution by X-ray crystallography, including the first three domains of GbpA,[8] and all four domains of the close homologue *Vh*LPMO10A from *Vibrio campbellii*,[11] the functional information has been limited, in part due to the inability to map hydrogen atoms in the structures. For redox-active enzymes like LPMOs, neutron macromolecular crystallography (NMX) is a strong complementary method to X-ray crystallography that can reveal additional information.[12] With NMX, functionally relevant hydrogens can be modeled even at relatively low (~2.5 Å) resolution.[13] Furthermore, neutrons are non-ionizing,[13] leaving the active site of LPMOs intact throughout the diffraction experiment. Unfortunately, neutron sources have a

Received: March 31, 2023
Accepted: July 14, 2023
Published: July 31, 2023

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

very low flux, and hydrogen atoms have a high incoherent scattering; this diminishes the signal-to-noise ratio for the crystallographic data. Longer data collection times (e.g., multiple days) and large crystal volumes (0.1−1.0 mm^3) can therefore be required to determine a neutron protein crystal structure. Also, hydrogen has a negative coherent scattering length compared to the positive scattering lengths of carbon, nitrogen, sulfur, and oxygen, which may cause density cancellation effects in neutron maps at intermediate resolution.[13] These issues can be resolved with protein deuteration. Deuterium has strong positive coherent scattering and incoherent scattering that is essentially negligible compared to that of hydrogen.[13] This alleviates the sample volume requirement by as much as a factor of ten. A few neutron crystal structures of LPMOs have already been determined;[14−17] however, none of them is perdeuterated, enabling mapping of only select hydrogen atoms.

Other neutron techniques that can benefit from protein deuteration are small-angle neutron scattering (SANS) and neutron reflectometry (NR), where contrast-matching can help visualize individual components in a complex, provided that the components have sufficiently different scattering-length densities (SLD). Deuterated biomolecules have very different SLDs from non-deuterated biomolecules. Using SANS or NR on deuterated LPMOs can therefore yield information on the interactions with other biomolecules, most obviously the polysaccharide substrates. An extensive review of deuteration methodologies/protocols/applications for neutron scattering has been given by Haertlein et al.[18]

Perdeuteration of proteins by expression in deuterated bacterial cultures can be a challenging and time-consuming process. Protocols usually require adaptation of the cells to increasing amounts of D_2O; however, Cai et al.[19] recently developed a faster and simpler protocol, yielding high amounts of perdeuterated proteins. We adapted this protocol for both full-length (FL) GbpA (GbpA-FL) and for its first domain (GbpA-D1), scaled up production at international deuteration facilities, and characterized perdeuterated GbpA-FL (D-GbpA-FL) by small-angle X-ray scattering (SAXS) and SANS. The protein retained its catalytic activity. We also crystallized and determined the structure of deuterated GbpA-D1 (D-GbpA-D1) using X-ray crystallography and compared it with the X-ray structure of the hydrogenated protein (H-GbpA-D1).

These results demonstrate the feasibility of perdeuterating LPMOs for neutron-based structural biology studies, with the promise of increased knowledge for the functional mechanisms of this important class of enzymes.

MATERIALS AND METHODS

Materials. Glycerol-d_8 (99% D) and deuterium oxide (99.9% D) were bought from ChemSupport AS (Hommelvik, Norway). For the experiments at D-lab and DEMAX, these chemicals were purchased from Eurisotop and Sigma-Aldrich, respectively. K_2HPO_4, KH_2PO_4, Na_2HPO_4, NH_4Cl, and glycerol were from VWR (Oslo, Norway). All other chemicals and chemical competent cells were purchased from Sigma-Aldrich (Merck Life Science AS, Oslo, Norway).

Production of Hydrogenated GbpA (H-GbpA). Unless specified otherwise, all the work described in this study was carried out with expression constructs provided by the Vaaje-Kolstad laboratory (Norwegian University of Life Sciences). They are described in detail by Wong et al.[8] Briefly, the GbpA-FL sequence was cloned from the genomic DNA of *V. cholerae* strain N1RB3 into a pET-22b vector between the *NdeI* and *XhoI* sites. The construct for expressing GbpA-D1 was obtained by the addition of two codon stops at position 203. The natural tag for protein secretion (amino acids 1−23) is cleaved off by the *E. coli* expression system. The pET-22b vector contains an ampicillin resistance gene, exploited for selection in growth and expression phases.

H-GbpA-FL was expressed in BL21(DE3) STAR cells transfected with the GbpA-encoding plasmid using Terrific Broth (TB), Luria Bertani (LB), or non-deuterated M9glyc+ media (Table 1, but with non-deuterated glycerol). For expression in TB or LB media, cells were first grown in 50 mL pre-cultures on INFORS culture shakers at 37 °C and 130 rpm to an optical density at 600 nm (OD_{600}) of 6−8 absorption units (AU), then transferred to 1 L cultures, and grown at 37 °C to an OD_{600} of 0.8 before induction with isopropyl β-D-1-thiogalactopyranoside (IPTG) at a final concentration of 1 mM. The temperature was lowered to 20 °C after induction, and H-GbpA-FL was expressed for 18 h.

For expression of hydrogenated protein (H-GbpA-FL and H-GbpA-D1) in non-deuterated M9glyc+ medium, the cells were grown for 8 h at 37 °C (INFORS shaker, 130 rpm) in 2.5 mL LB/H_2O medium in 15 mL tubes. This culture was transferred to 25 mL non-deuterated M9glyc+ medium in a 250 mL baffled

Table 1. Composition of Deuterated Media (M9+ Compared to M9glyc+ and ModC1; Ingredients for 1 L Media)

M9+ medium[19]		M9glyc+ medium (this work)	
K_2HPO_4	19.0 g	K_2HPO_4	19.0 g
KH_2PO_4	5.0 g	KH_2PO_4	5.0 g
Na_2HPO_4	9.0 g	Na_2HPO_4	9.0 g
K_2SO_4	2.4 g	K_2SO_4	2.4 g
D-Glucose-d_7	18.0 g	Glycerol-d_8	18.0 g
NH_4Cl	5.0 g	NH_4Cl	5.0 g
Trace element solution[a]	1.0 mL	Trace element solution[a]	1.0 mL
MEM	10.0 mL	MEM[b]	10.0 mL
$MgCl_2$	0.95 g	$MgCl_2$	0.95 g[c]

ModC1 medium[20,21]	
NH_4Cl	2.58 g
KH_2PO_4	2.54 g
Na_2HPO_4	4.16 g
K_2SO_4	1.94 g
Glycerol-d_8	2 g
d-algal extract	10 mL
Trace element solution[d]	1.0 mL
Vitamin mix[e]	1.0 mL
$MgSO_4(7H_2O)$	0.67 g
$FeSO_4(7H_2O)$	20 mg
Trisodium citrate	88 mg

[a]The trace element solution was made by dissolving the following ingredients in 100 mL H_2O: 0.6 g $FeSO_4$ ($7H_2O$), 0.6 g $CaCl_2$ ($2H_2O$), 0.12 g $MnCl_2$ ($4H_2O$), 0.08 g $CoCl_2$ ($6H_2O$), 0.07 g $ZnSO_4$ ($7H_2O$), 0.03 g $CuCl_2$ ($2H_2O$), 2 mg H_3BO_4, 0.025 g $(NH_4)_6Mo_7O_{24}$ ($4H_2O$), 0.5 g EDTA. [b]MEM vitamin solution from Sigma-Aldrich. [c]MgCl was dissolved in 10 mL D_2O prior to addition, which prevented precipitation. [d]The trace element solution was made by dissolving the following ingredients in 1 L of D_2O to prepare a 1000× stock solution: 5.1 g $MnSO_4$ (H_2O), 8.6 g $ZnSO_4$ ($7H_2O$), 0.75 g $CuSO_4$ ($5H_2O$). [e]The vitamin mix solution was made by dissolving the following ingredients in 1 L of D_2O to prepare a 1000× stock solution: 25 mg biotin, 135 mg vitamin B12, 335 mg thiamine.

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

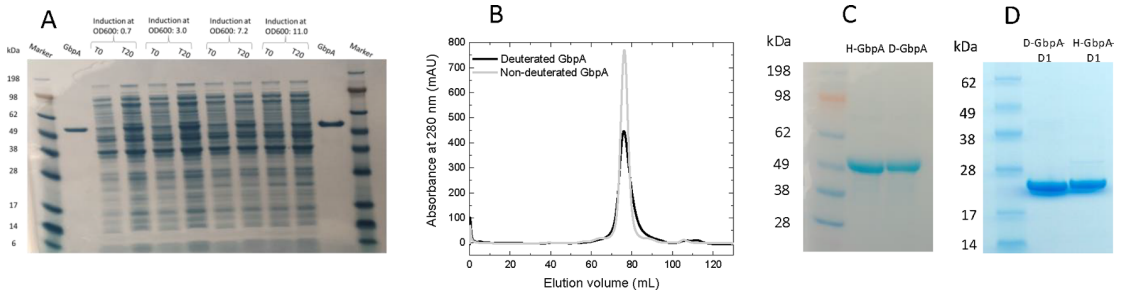


Figure 1. Expression of GbpA in deuterated M9glyc+ medium. (A) SDS-PAGE analysis of full cell proteome, pre-induction (T0), and post-induction (T20), compared for the four different induction points. Seeblue Plus2 marker and purified full-length GbpA (FL) are included for comparison. (B) SEC elution profiles for H-GbpA-FL and D-GbpA-FL (both produced at UiO). Similar elution peaks at the same retention volumes indicate highly comparable hydrodynamic radii. (C, D) SDS-PAGE of H-GbpA-FL and D-GbpA-FL (C) as well as LMPO domain GbpA-D1 (produced at UiO) (D) after SEC.

flask and incubated for another 15 h. Subsequently, we added the culture to 225 mL M9glyc+ media in a 2 L baffled flask. We tested induction with 1 mM IPTG at four different OD_{600} levels (0.7, 3.0, 7.2 and 11.0; Figure S1B) and obtained the best yield at OD_{600} = 3.0 (Figure 1A). After addition of IPTG, the temperature was lowered to 25 °C for 20 h of expression. For a scheme of the protocol, see Figure 2. 100 μM sodium ampicillin was used for selection.

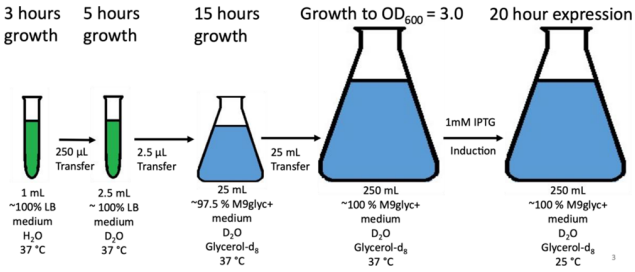


Figure 2. Optimized GbpA deuteration protocol. *E. coli* BL21 Star cells are initially grown in LB medium for 3 h and then transferred to LB medium containing D_2O, in which the cells are grown for 5 h. Cells are then added to a small M9glyc+ D_2O pre-culture, in which they are allowed to grow for 15 h until they are transferred to a larger culture. When OD_{600} reaches 3.0, expression is induced by addition of IPTG, and the temperature is lowered from 37 to 25 °C.

Production of Deuterated Protein (D-GbpA). When producing deuterated protein (D-GbpA-FL and D-GbpA-D1), we used the same conditions as for H-GbpA production in minimal media, except for one additional step: After growing BL21(DE3) Star cells containing GbpA-encoding plasmid for 3 h in 1 mL LB/H_2O medium in 15 mL tubes, this culture was diluted ten times into 2.5 mL LB/D_2O medium in 15 mL tubes, before incubation for another 5 h. Subsequently, 2.5 mL of this culture were transferred to 25 mL deuterated M9glyc+ medium (Table 1) in a 250 mL baffled flask. This culture was incubated for 15 h and subsequently added to 225 mL deuterated M9glyc+ medium in a 2 L baffled flask, where it remained until OD_{600} reached 3.0 at 37 °C. The production of GbpA was induced by addition of IPTG to a concentration of 1 mM. The culture was left for expression for 20 h at 25 °C. A final concentration of 100 μM sodium ampicillin was used for selection. The cultures remained on culture shakers (INFORS Multitron Standard) at 130 rpm throughout the growth and expression phases. Since we only obtained insufficient yields of D-GbpA-D1 for NMX (~3 mg/L of media), we contacted international facilities to help us scale up production.

For the production of D-GbpA-D1 at D-lab (ILL) and DEMAX (ESS), a different construct was used (but coding for the same amino acid sequence; UniProt ID: Q9KLD5, residues 24−485). The gbpa gene was codon-optimized and cloned into pET vector pET-26b by GenScript (Leiden, The Netherlands) using the restriction sites *Nco*I and *Xho*I. The pelB leader sequence is cleaved off during post-translational translocation to the periplasmic space to instate the catalytically important His24 as N-terminal amino acid. The C-terminal His-tag was omitted by inclusion of a stop codon at the end of the insert. The pET-26b vector contains a kanamycin resistance gene, which we exploited for selection in growth and expression phases.

For D-GbpA-D1 production at the D-lab deuteration facility of Institut Laue-Langevin (ILL, Grenoble, France; proposal number DL-03-223), the protein was over-produced in *E. coli* strain BL21 (DE3) adapted to growth in deuterated minimal medium,[18] using the expression construct cloned in the pET-26b plasmid. A 1.9 L (final volume) deuterated high-cell-density fed-batch fermenter culture was grown at 30 °C. Feeding with glycerol-d_8 was started at an OD_{600} value of about 5. Expression of D-GbpA-D1 was induced at an OD_{600} of about 13 by addition of 1 mM IPTG (final concentration). Cells were harvested at OD_{600} = 15.8, yielding 40 g perdeuterated cell paste (wet weight), thus approximately 20 g/L media. The cell paste was flash frozen and stored at −80 °C to prevent proteolysis before transport to Oslo for further processing.

D-GbpA D1 production in the DEMAX biodeuteration labs of the European Spallation Source (ESS, Lund, Sweden; proposal number 890320) followed a different protocol, without adaptation, using an approach described in Koruza et al.[21] D-GbpA D1 was expressed in BL21 (DE3) Tuner cells after transformation with GbpA-D1-encoding plasmid using selective growth conditions (i.e., in the presence of kanamycin) on LB agar plates. A 5 mL LB starter culture was grown with kanamycin (50 μg/mL final concentration) after inoculation from a single colony, and was left to grow while shaking at 180 rpm at 37 °C overnight. From this overnight culture, small-scale expression tests were performed in 50 mL LB cultures and a glycerol stock was prepared. For scaling up the yield, the cells were pre-grown in LB medium, a 100 mL overnight culture was started from the glycerol stock, and kanamycin was added (50 μg/mL final concentration). The next morning, 6 × 1 L cultures were inoculated with 10 mL of the overnight culture and fresh antibiotic was added. The cultures were grown in Tunair baffled

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

Table 2. SAXS and SANS

(a) Sample details	
Organism	*Vibrio cholerae*
Expression host	*E. coli* BL21(DE3) Star
Uniprot sequence ID (residues in construct)	Q9KLD5 (24−485)
Extinction coefficient [A_{280}, 0.1% (=1 g/L)]	1.906
$\overline{\nu}$ from chemical composition ($cm^3\ g^{-1}$)	0.73
Particle contrast, $\Delta\overline{\rho}$ ($10^{10}\ cm^{-2}$)	2.99
MM from chemical constituents (kDa)	51.3
Protein concentration (mg/mL)	46.4 μM for H-GbpA, 36.0 for μM D-GbpA
Solvent	100 mM NaCl, 20 mM Tris−HCl, pH 8.0
(b) SAXS data-collection parameters	
Instrument	Bruker Nanostar with InCoatec Cu microsource and Våntec-2000 detector (RECX, University of Oslo, Norway)
Wavelength (Å)	1.54
Beam size (μm)	750 × 750
Sample to detector distance (cm)	109
q measurement range ($Å^{-1}$)	0.00925−0.29866
Absolute scaling method	Milli-Q water standard measurement
Normalization	Transmitted intensities through semi-transparent beam-stop
Exposure time (h)	1
Capillary size (mm)	1.5
Sample temperature	24 °C
(c) Software employed for SAXS data processing, analysis and interpretation	
SAXS data processing	*SUPERSAXS* (CLP Oliveira and JS Pedersen, unpublished)
Extinction coefficient estimate	ProtParam[32]
Calculation of contrast and specific volume	*MULCh1.1*[33]
Basic analysis	*PRIMUS* (*ATSAS*)[27,28]
Shape reconstruction	*DAMMIF*[29]/*DAMAVER*[30]/*DAMMIN*[31]
Representation	PyMOL

(d) Structural parameters

	H-GbpA	D-GbpA
Guinier analysis		
$I(0)$ (cm^{-1})	0.0406 ± 0.0008	0.03776 ± 0.0008
R_g (Å)	37.35 ± 1.02	36.78 ± 1.25
q_{min} ($Å^{-1}$)	0.0114	0.0107
qR_g max	1.26	1.24
R^2	0.92	0.88
MM from $I(0)$ (kDa) (ratio to predicted)	52.0 (1.01)	48.2 (0.94)
$P(r)$ analysis		
$I(0)$ (cm^{-1})	0.0416 ± 0.0001	0.03944 ± 0.0007
R_g (Å)	40.8 ± 1.0	40.6 ± 0.9
d_{max} (Å)	142.5	142.5
q range ($Å^{-1}$)	0.0114−0.2109	0.0107−0.2165
χ^2	0.99	0.94
Total quality estimate from *PRIMUS*	0.65	0.70
MM from $I(0)$ (kDa) (ratio to predicted)	53.2 (1.04)	50.5 (0.98)

(e) Shape model-fitting results

	H-GbpA	D-GbpA
DAMMIF		
q range ($Å^{-1}$)	0.0114−0.2109	0.0107−0.21086
Symmetry, anisotropy assumptions	*P*1, None	*P*1, None
NSD (Standard deviation)	0.695 (0.043)	0.811 (0.054)
Resolution (Å)	31 ± 3	32 ± 3
MM from *DAMMIF* (kDa) (ratio to predicted)	38.2 (0.74)	40.4 (0.79)
χ^2	0.984−1.003	0.797−0.804
DAMAVER/DAMMIN		
q range	0.0114−0.2109	0.0107−0.21086
Symmetry, anisotropy assumptions	*P*1, none	*P*1, none

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

Table 2. continued

χ^2	0.993		0.796	
Constant adjustment	Skipped		Skipped	
(f) SANS data analysis				
	H-GbpA 100% D_2O	D-GbpA 0% D_2O	D-GbpA 45% D_2O	D-GbpA 100% D_2O
Guinier analysis				
$I(0)$ (cm^{-1})	0.075 ± 0.001	0.302 ± 0.004	0.118 ± 0.002	0.019 ± 0.001
R_g (Å)	36.2 ± 0.5	36.1 ± 0.7	34.7 ± 0.8	37.2 ± 2.5
q_{min} ($Å^{-1}$)	0.0133	0.0159	0.0149	0.0181
qR_g max	1.26	1.22	1.30	1.18
R^2	97.5	95.6	91.9	76.6

flasks in a 37 °C incubator while shaking at 200 rpm. When OD_{600} reached 2, the cells were harvested by centrifugation at 5000 × *g* for in a JLA8.100 rotor for 10 min. The cells were gently resuspended in deuterated ModC1 (Table 1) medium and transferred to 6 × 1 L deuterated ModC1 medium in fresh flasks. These media were prepared according to the protocol reported by Koruza et al.[21] with the addition of 10 mL of *Botryococcus braunii* deuterated algal extract and a reduction in the amount of glycerol-d_8 to 2 g/L of media. At this point, fresh antibiotic was added, and the cells were allowed to recover for 1 h at 25 °C while shaking at 100 rpm. Thereafter, the shaking was increased to 200 rpm, 1 mM IPTG (final concentration) was added, and expression was left to continue overnight for up to 20 h. Finally, the cells were harvested by centrifugation at 10,000 × *g* for 15 min in a JLA8.1000 rotor and immediately further processed for periplasmic protein extraction as described below. Using this approach, it was possible to obtain 60 g of wet cell paste (i.e., 10 g/L of media) for further processing. The Certificate of Analysis of the material provided is deposited with DOI 10.5281/zenodo.6631673. The periplasmic fraction was frozen at −20 °C and shipped on ice packs for further purification.

Autolysis Procedure for Deuterated Algal Extract. The procedure for algal autolysis was adjusted and modified from literature descriptions.[22,23] Microalgae *B. braunii* (UTEX Showa strain, Culture Collection of Algae at the University of Texas at Austin) were continuously cultivated in perdeuterated modified Bolds 3 N medium.[24] Cells were grown in a 12 h:12 h light−dark cycle and illuminated with 240 μmol photons/m^2s LED lights in a Multitron Pro incubator (INFORS HT). Cultures were agitated by shaking at 60 rpm, and the atmosphere was kept at 2% CO_2. The cells were periodically harvested, typically every 14−21 days, by centrifugation in a JLA8.100 rotor (Beckman) at 5000 × *g* for 15 min. The pelleted cells were frozen at −80 °C until further processing. To prepare the autolysate from frozen cells, 400 mL of 99.9% D_2O was added to 100 g of frozen wet microalgae cells. The cells were thawed and resuspended to a uniform suspension and then incubated at 50 °C in a water bath for 24 h. The digested cell product was centrifuged in a JA-21 rotor (Beckman) at 10,000 × *g* for 20 min. The supernatant was aliquoted in 10 mL tubes and frozen at −20 °C until needed. For modified ModC1 medium preparation, 10 mL of d-algal extract was used per liter of culture media.

Periplasmic Lysis and Protein Purification. GbpA (both FL and D1, deuterated or hydrogenated) was harvested from the *E. coli* periplasm using the following periplasmic lysis protocol: First, the cell culture was pelleted by centrifugation (10,000 × *g*). The pellet was resuspended in a solution containing 25% sucrose, 20 mM tris(hydroxymethyl)aminomethane (Tris)-HCl pH 8.0 and 5 mM ethylenediaminetetraacetic acid (EDTA) (4−5 mL/g cells), and incubated for 30 min. Thereafter, the cells were once again pelleted and resuspended in ∼50 mL (4−5 mL/g cells) solution of 5 mM $MgCl_2$, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 0.25 mg/mL lysozyme. The suspension was incubated for 30 min on ice, pelleted, and the GbpA-containing supernatant was subjected to purification.

GbpA (both FL and D1, deuterated or hydrogenated) was purified by anion-exchange and size-exclusion chromatography. Anion-exchange chromatography (AEX) was performed using a HiTrap Q HP 5 mL column (Cytiva) connected to an ÄKTA Start protein purification system (GE Healthcare). After loading, the protein was eluted by a salt gradient from 100 to 400 mM NaCl buffered with 20 mM Tris−HCl pH 8.0. Size-exclusion chromatography (SEC) was performed using a HiLoad Superdex 200 prep grade column on an ÄKTA Purifier system (GE Healthcare) in a running buffer containing 20 mM Tris−HCl pH 8.0 and 100 mM NaCl. For GbpA-D1, we used a salt gradient for elution during AEX from 50 to 400 mM and performed SEC using a Superdex 75 Increase 10/300 column (Cytiva) on an ÄKTA pure system (GE Healthcare).

Determination of Deuterium Content. To determine the deuteration level, deuterated and non-deuterated GbpA-FL were dialyzed into MilliQ-purified H_2O and measured by MALDI-TOF MS at the proteomics core facilities at UiO (Thiede lab). An ULTRAFLEX II (Bruker Daltonics, Bremen, Germany) MALDI-TOF/TOF mass spectrometer was used after external calibration. The samples were mixed with matrix (20 mg/mL α-cyano-4-hydroxycinnamic acid in 0.3% aqueous trifluoroacetic acid/acetonitrile (1:1)) and applied to a stainless-steel sample holder. Basic settings of the MALDI-TOF/TOF instrument were as follows: Ion source 1, 25 kV; ion source 2: 21.85 kV; lens: 9.60 kV; reflector: 26.3 kV; reflector 2, 13.85 kV; deflector mode, polarity positive. MS spectra were transformed into peak lists by using the software FlexAnalysis version 2.4 (Bruker Daltonics, Bremen, Germany).

Small-Angle X-ray Scattering. SAXS data were acquired on a Bruker NanoStar instrument using 40.0 μM H-GbpA-FL or 32.0 μM D-GbpA-FL in 100 mM NaCl, 20 mM Tris−HCl pH 8.0, with data acquisition times of 1 h per data set. Scattering intensities were recorded as a function of the scattering vector $q = (4\pi/\lambda)\sin\theta$, where 2θ is the scattering angle and λ is the wavelength (λ = 1.54 Å). Data were collected in the q-range: 0.009 to 0.3 $Å^{-1}$.

The scattering intensities were corrected for electronic noise, empty cell scattering, and detector sensitivity. The scattering contribution from the buffer was subtracted, and intensities were calibrated to absolute units with H_2O scattering as standard,

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

using the *SUPERSAXS* program package (CLP Oliveira and JS Pedersen, unpublished; implementation explained in ref 25).

Radii of gyration and pair-distance distribution functions (from inverse Fourier transform[26]) were calculated with *PRIMUS*[27] from the *ATSAS*[28] package. For both H-GbpA-FL and D-GbpA-FL, 20 low-resolution models were calculated by *ab initio* shape determination using the *DAMMIF*[29] software. We built average models with *DAMAVER*[30] and refined them with *DAMMIN*.[31] All three programs are from the *ATSAS*[28] package. SAXS data are summarized in Table 2.

Small-Angle Neutron Scattering. H-GbpA was dialyzed into a buffer containing 100 mM NaCl and 20 mM Tris−HCl pH 8.0 in 100% D_2O. D-GbpA was dialyzed into the same buffer, but using different D_2O/H_2O ratios, i.e., at 100% H_2O, 45% D_2O, or 100% D_2O. SANS data for the four samples were acquired at ILL beamline D11 at λ = 5.6 Å for 2 h, using a protein concentration of 39.2 μM. Subsequently, the data were processed with beamline software (Table 2).

X-ray Crystallography. Hydrogenated and deuterated GbpA-D1 were crystallized using the same protocol and conditions. First, the proteins were saturated with Cu^{2+} by addition of $CuCl_2$ in a molar ratio of 3:1 ($CuCl_2$ to GbpA) and subsequently desalted to a buffer containing 100 mM NaCl, 20 mM Tris−HCl pH 8.0 using a 5 mL HiTrap desalting column. No crystals were obtained when the copper-binding step was omitted. Initial screening yielded crystals under many conditions, but most were either very small, irreproducible, or exclusive to either H-GbpA or D-GbpA. The following paragraph describes the procedure and conditions that yielded reproducible crystals in space group $P2_12_12$ for both proteins, which also diffracted to high resolution.

GbpA-D1 crystals grew from a solution containing the purified protein at 6−10 mg/mL. Sitting-drop vapor diffusion experiments were set up in 96-well 3-drop PS plates (SwissCI). 0.5 μL protein were added to an equal volume of crystallization solution containing 100 mM sodium cacodylate pH 6.5, 200 mM zinc acetate and 18% w/v PEG 8000 (VWR). Crystals grew over the course of two weeks at 20 °C. Crystals were subsequently cryoprotected in mother liquor supplemented with 15% glycerol and flash-cooled in liquid nitrogen before data collection.

Diffraction data were collected at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) at the beamlines ID23−1 (Pilatus 6 M Dectris detector) for D-GbpA (diffraction to 1.1 Å) and ID23-2 (Pilatus3 X 2 M detector) for H-GbpA (diffraction to 1.6 Å resolution). The DOI for the data collection of H-GbpA-D1 is 10.15151/ESRF-ES-541149090 (no DOI was generated for the data collection of D-GbpA-D1).

X-ray data were processed automatically by the EDNA processing pipeline[34] for H-GbpA-D1 and XIA2_DIALS[35] for D-GbpA-D1. In all subsequent steps, the *CCP4* software suite[36] was used. Structures were solved by molecular replacement (with *Phaser*[37]) using domain D1 of the published GbpA structure (PDB ID: 2XWX)[8] as a model. Real-space refinement and model building were performed with *Coot*,[38] and subsequent refinement cycles using *REFMAC5*.[39] Ions and water molecules were added only after the protein chain had been modeled. Finally, occupancies were refined for protein atoms and anomalous scattering ions. Zinc and copper ions were identified with confidence based on data collected at their absorption edges for anomalous scattering (Tables S1 and S2; Figure S3). Full anisotropic refinement was carried out for the 1.1 Å D-GbpA-D1 model, whereas this was not warranted for the lower-resolution H-GbpA-D1 structure (1.6 Å). The refined structures were deposited in the Protein Data Bank (PDB, www.rcsb.org)[40] with PDB IDs: 8CC3 and 8CC5. Data collection and refinement statistics are reported in Table 3.

Table 3. Data Collection and Refinement Statistics[a]

	H-GbpA D1	D-GbpA D1
(a) Data collection		
Beamline	ID23−2 (ESRF)	ID23−1 (ESRF)
Wavelength (Å)	0.8731	0.9763
Resolution range	44.7−1.6 (1.68−1.62)	47.1−1.1 (1.17−1.13)
Space group	$P2_12_12$	$P2_12_12$
Unit cell axes: a, b, c (Å)	75.3 89.4 47.5	74.9 89.1 47.1
R_{merge} (%)	11.4 (>100)	7.9 (>100)
$CC_{1/2}$	1.00 (0.37)	1.00 (0.22)
Mean I/σ	10.8 (0.9)	6.1 (1.2)
Completeness (%)	99.9 (100.0)	98.7 (95.9)
Multiplicity	6.8 (6.7)	2.2 (2.0)
Unique reflections[b]	78,744 (4085)	221,328 (11391)
(b) Refinement		
Resolution range (Å)	44.7−1.6	47.1−1.1
R_{work}/R_{free}[c]	0.182/0.217	0.184/0.201
Macromolecules/a.u.	2	2
Number of non-hydrogen atoms	3240	3526
Protein	3034	3256
Ligands	19	18
Waters	187	252
B-factors (Å^2)		
Protein	27.0	12.6
Ligands	36.0	14.6
Waters	32.1	22.0
R.m.s.d. from ideal values		
Bond length (Å)	0.008	0.009
Bond angles (°)	1.44	1.58
Ramachandran Plot		
Favored (%)	95.5	96.7
Outliers (%)	0.3	0.0
PDB ID	8CC3	8CC5

[a]Statistics for the highest resolution shell shown in parentheses. [b]Data reported treating Bijvoet pairs as separate reflections. [c]R_{free} was calculated from 5% of randomly selected reflections for each dataset.

To unambiguously identify the metal species observed in GbpA-D1, diffraction data were collected at the K absorption edge of zinc (around 9660 eV) and copper (around 8980 eV). Three data sets were collected at the BioMAX beamline of MAX IV (Lund, Sweden; Table S1) using isomorphous crystals grown under the same conditions as the H-GbpA-D1 and D-GbpA-D1 structures described above. The data were integrated and scaled by the *autoPROC* automatic processing pipeline at MAX IV,[41] and subsequently scaled and truncated to 2.5 Å with *XSCALE*, a component of the *XDS* software package.[42] The phases for the highest-resolution dataset (9320 eV) were determined by molecular replacement as described above, and the structure thereafter refined in iterative cycles of maximum-likelihood refinement using *REFMAC5*[39] and manual real-space refinement in *Coot*.[38] Data collection and refinement statistics for this dataset are given in Table S2.

Phase information from the 9320 eV refined structure was used to generate anomalous difference maps for each of the datasets ($D_{ano}^{10.0k}$, $D_{ano}^{9.3k}$, $D_{ano}^{8.5k}$) using the *FFT* tool from the

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

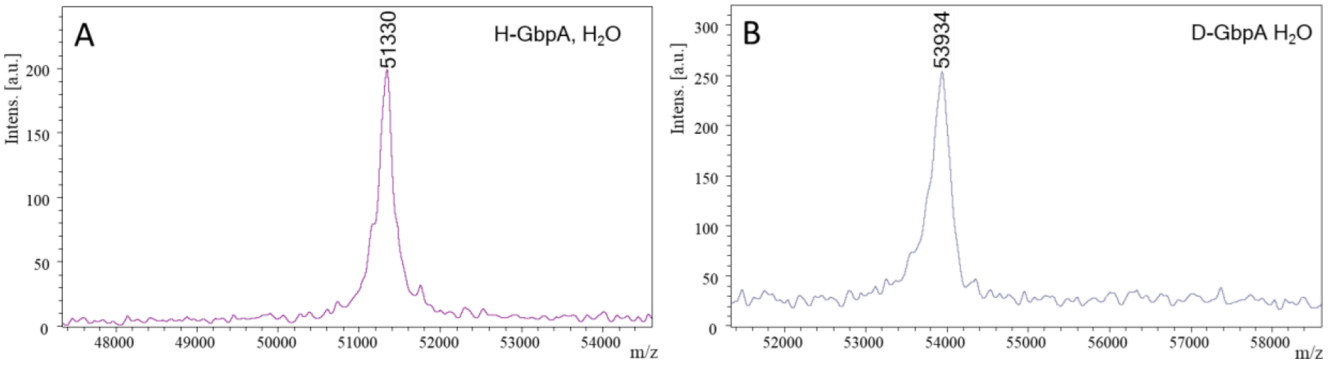


Figure 3. Deuteration. The deuteration levels were quantified with MALDI-TOF MS for full-length H-GbpA (A) and D-GbpA (B), in H_2O. With a theoretical hydrogenated mass of 51,248 Da and a theoretical per-deuterated mass of 53,930 Da, the deuteration level of D-GbpA was determined to be approximately 97%. The slightly higher experimental values compared to theoretical values are likely due to instrument calibration for different mass ranges.

CCP4 program suite.[36] Completeness of the datasets at a resolution higher than 3 Å was limited due to the presence of water ice (see Table S1); however, the anomalous signal at lower resolution was sufficient to allow identification of the positions and identities of the metal ions. Difference density maps of anomalous difference maps were generated using the *CAD* and *FFT* tools to find peaks corresponding to copper atoms ($D_{ano}^{9.3k} - D_{ano}^{8.5k}$) and zinc atoms ($D_{ano}^{10.0k} - D_{ano}^{9.3k}$). For each metal-binding site detected by any of the anomalous difference maps ($D_{ano}^{10.0k}$, $D_{ano}^{9.3k}$, $D_{ano}^{8.5k}$), the combined presence or absence of a peak in the $D_{ano}^{9.3k} - D_{ano}^{8.5k}$ and the $D_{ano}^{10.0k} - D_{ano}^{9.3k}$ maps revealed whether the position was occupied by copper, zinc, or a combination of both. The histidine brace motif only showed a peak in the $D_{ano}^{9.3k}$ and the $D_{ano}^{9.3k} - D_{ano}^{8.5k}$ maps (copper absorption edge) but not in the $D_{ano}^{10.0k} - D_{ano}^{9.3k}$ map calculated around the zinc absorption edge. This confirmed the exclusive presence of copper in its catalytic position.

Activity Studies. GbpA was saturated with Cu^{2+} as described in the previous section, and β-chitin nanofibers from France Chitine (Orange, France) were prepared according to a protocol developed by Loose et al.[9] 1 μM GbpA-FL was mixed with 5 mg/mL β-chitin nanofibers in 20 mM Tris−HCl pH 8.0, and the reaction initiated by addition of 1 mM sodium ascorbate. After incubation on an INFORS shaker for 1.5 h at 37 °C, the reaction was stopped by boiling and subsequent filtering through a 0.2 μm cellulose filter. The products were measured with MALDI-TOF MS. The experiment was performed in triplicates for both H-GbpA and D-GbpA.

RESULTS AND DISCUSSION

High-Yield Production of Deuterated GbpA. In order to optimize expression conditions for GbpA, we first carried out experiments using non-deuterated modified M9 medium, which we refer to as M9glyc+ (Table 1; recipe modified from refs 19,43). A plateau in optical density at 600 nm (OD_{600}) was reached at 12 AU after 12 h without induction (Figure S1A). For comparison, expression in LB medium reached a plateau much earlier (after 5 h) but only at 4 AU (Figure S1A). Expression of GbpA-FL in M9glyc+ was induced by adding 1 mM IPTG at four different cell densities, OD_{600} = 0.7, 3.0, 7.2, and 11.0 (Figure S1B). The optimal induction point was identified to be at OD_{600} = 3.0 based on band intensities over background on SDS-PAGE (Figure 1A). The protein was subsequently purified using a periplasmic isolation protocol, followed by AEX and SEC. The protein eluted at the same retention volume as GbpA-FL expressed in TB (Figure S1C) and exhibited equivalent purity (Figure S1D).

We then grew cell cultures containing GbpA-FL-encoding plasmid in deuterated M9glyc+ medium (Table 1), following a stepwise adaptation protocol, from LB/H_2O to LB/D_2O to D_2O and glycerol-d_8-containing minimal medium (Figure 2). In the final step, cells were grown to OD_{600} = 3.0, and expression was induced. Both H-GbpA-FL and D-GbpA-FL eluted as a single peak from SEC and were shown to be highly pure as evaluated by SDS-PAGE (Figure 1B−D).

We found that using deuterated glycerol instead of the more expensive deuterated glucose used by Cai et al.[19] worked well. A protein yield of 12 mg purified protein from 1 L of expression media was calculated from absorbance at 280 nm. Surprisingly, the yield was approximately two-fold higher than from optimized expression in non-deuterated TB medium, as suggested by final protein contents of the purified samples as compared to the total volume of the expression cultures.

Subsequently, we applied the same protocol to GbpA-D1, in preparation of NMX experiments. A protein yield of 3 mg pure deuterated protein from 1 L of expression media was obtained, as calculated from absorbance at 280 nm (thus 4-fold less than for D-GbpA-FL). Purity was assessed with SDS-PAGE (Figure 1D). The protocol, with small variations, was subsequently scaled up with improved yields by D-lab at ILL and DEMAX at ESS. The procedure for producing deuterated proteins used at DEMAX (ESS) is done in traditional shaker flasks and includes supplementing the deuterated glycerol in the media with deuterated algal extracts. The method described in this work serves as a proof-of-concept, and the optimized formulation and procedure is the subject of a separate publication that is in preparation. The procedures at D-Lab (ILL) use fermenters and as such require far less D_2O (almost 3-fold less) than the shaker method, but the batch-fed approach uses more deuterated glycerol. Despite the fundamental differences in how the deuterium labeling is achieved, both D-Lab and DEMAX produced around 30 mg of D-GbpA-D1 each, indicating that the methods are equivalent for this protein.

An additional difference that could potentially affect the protein quality and yield was the handling after the protein production stage. Usually, the cells were immediately further processed for periplasmic protein extraction, but in one case, the cell paste was frozen and shipped prior to further extraction. This step may have reduced the final yield, but the purified

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

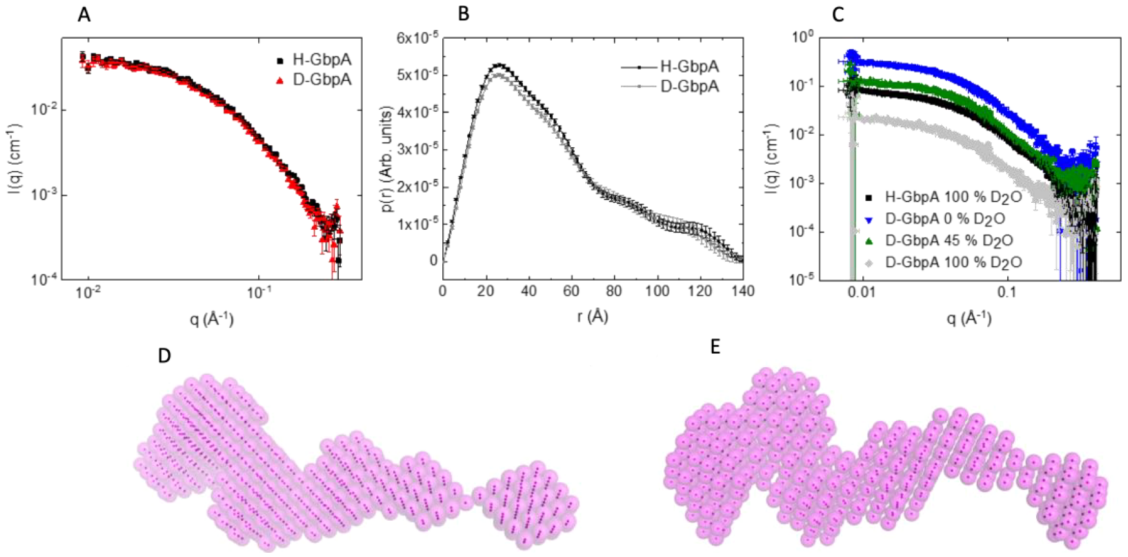


Figure 4. Small-angle scattering experiments on H-GbpA and D-GbpA. (A) SAXS curves of full-length H-GbpA and D-GbpA (intensities normalized to a 1 mg/mL concentration. (B) Pair-distance distribution functions, revealing elongated proteins with maximal dimension of 140 Å. (C) SANS data for H-GbpA and D-GbpA, in buffers containing different levels of D_2O. (D, E) SAXS *ab initio* models averaged from 20 structural models, shown for H-GbpA (D) and D-GbpA (E), respectively.

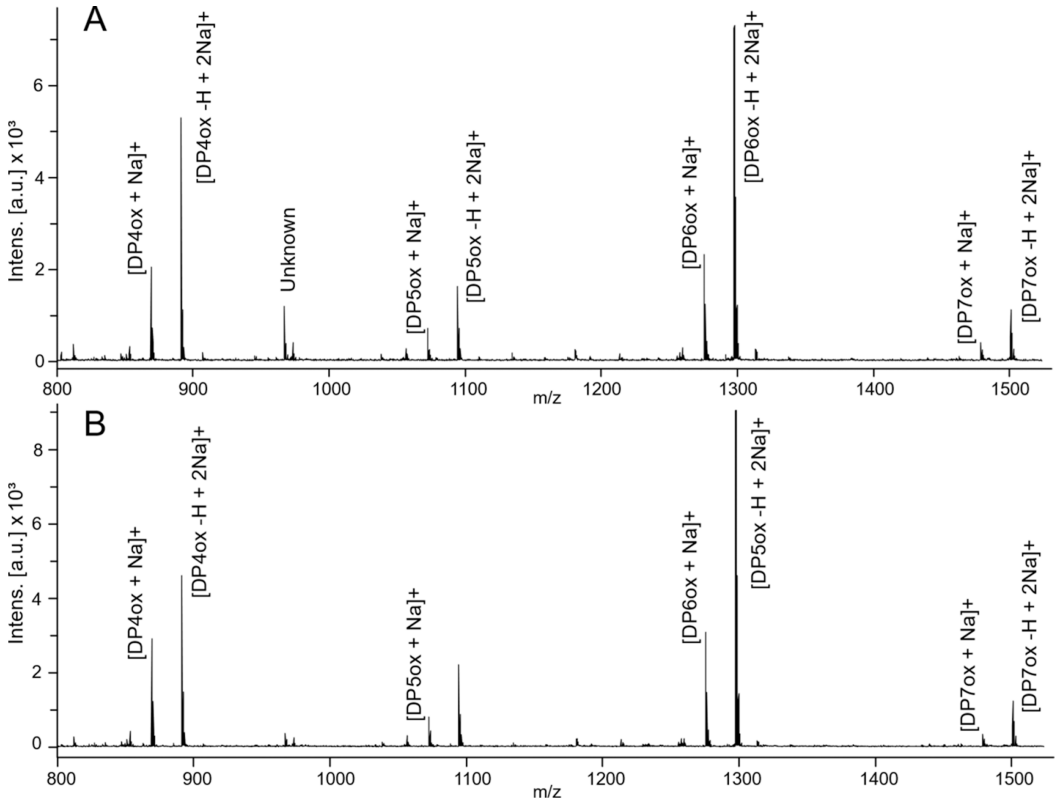


Figure 5. LPMO activity. MALDI-TOF MS spectra of LPMO reaction products after cleavage of β-chitin by H-GbpA (A) and D-GbpA (B). Labeled peaks correspond to chitooligosaccharides of four (869−891), five (1073−1095), six (1276−1298), and seven (1501) units. All labeled peaks correspond to masses of oxidized chitooligosaccharides. Some peaks correspond to saccharides bound to one sodium ion (1276, 1073, 869) and others to two (671, 891, 1095, 1298, 1501). One significant peak is an unknown species (967).

protein could be crystallized, as reported below, suggesting that the sample quality did not suffer significantly.

D-GbpA Is Fully Deuterated, Catalytically Active and Retains Its Fold Compared to H-GbpA. To assess the level of deuteration, intact mass of the full-length protein was quantified by MALDI-TOF Mass Spectrometry (MS) (Figure 3). Masses of non-deuterated (H-GbpA) and deuterated GbpA (D-GbpA) in H_2O were measured to be 51,330 and 53,934 Da, respectively. The theoretical molecular weight of a deuterated protein (MW_{dT}) in H_2O is given by

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

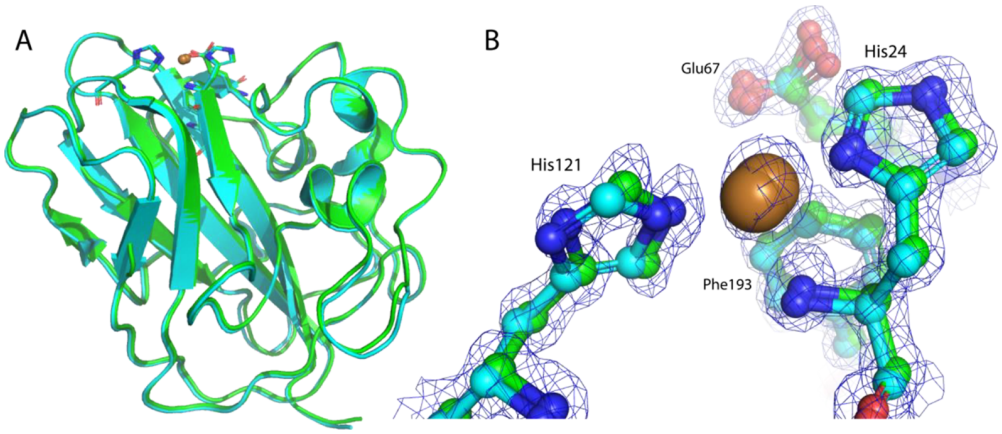


Figure 6. H-GbpA-D1 and D-GbpA-D1 structures. (A) Superimposition of X-ray crystal structures of D-GbpA-D1 (green; PDB ID: 8CC5) and H-GbpA-D1 (cyan, PDB ID: 8CC3) (r.m.s.d. between C_α atoms = 0.2 Å, as calculated with PyMOL, Schrödinger LLC). Both structures are the results from this work. (B) Close-up view of active sites shown in A, including histidine brace and copper ions (bronze spheres), with sigma-A-weighted $2mF_o - DF_c$ electron density map shown for D-GbpA-D1 at 1.5 σ. Figures were prepared with PyMOL (Schrödinger LLC).

$$MW_{dT} = MW_{hT} + NXH \cdot 1.006$$

where MW_{hT} is the theoretical molecular weight of the non-deuterated protein, NXH the number of non-exchangeable hydrogens, and 1.006 is the relative mass difference between hydrogen and deuterium. Using the method of Meilleur et al.[44] for calculating the number of non-exchangeable hydrogens and estimating deuteration levels gives relative molecular weight values of MW_{hT} = 51,250 and MW_{dT} = 53,930 for GbpA. The deuteration level can then be calculated with

$$\text{deuteration level} = \frac{MW_{dE} - MW_{hE}}{MW_{dT} - MW_{hT}}$$

where MW_{dE} and MW_{hE} correspond to the experimentally determined masses of the deuterated and non-deuterated GbpA, respectively. This gives a deuteration level of 97%.

The remaining 3% undesired hydrogen may result from vapor exchange with the air or contamination from the various chemicals used (Table 1). However, 97% deuteration is more than sufficient for neutron-based experiments, and the neutron SLD of 6.12×10^{-6} Å^{-2} (in H_2O) for D-GbpA (compared to 1.97×10^{-6} Å^{-2} for H-GbpA) is expected to be significantly distinguishable from all non-deuterated biomolecules. Interaction studies exploiting contrast matching should therefore be possible with any interaction partner with a defined SLD, most obviously chitin or mucin, but also potential bacterial cell-surface interaction partners or enzyme complex partners yet to be identified.

To confirm that the overall structure of GbpA-FL was unaffected by deuteration, we performed SAXS experiments for both D-GbpA and H-GbpA (Figure 4A). Indeed, the SAXS profiles of both proteins were highly similar, with a radius of gyration (R_g) of 37.4 Å for H-GbpA and 36.8 Å for D-GbpA. Both proteins are monomers in solution, with a maximal diameter of approximately 140 Å (Figure 4B). Structural parameters are summarized in Table 2. *Ab initio* modeling suggests similar, elongated structures of both hydrogenated and deuterated GbpA (Figure 4D,E). In order to test the neutron scattering of D-GbpA at different concentrations of D_2O, we carried out SANS experiments on D-GbpA under three different conditions (Figure 4C), comparing the scattering in buffer at 100% H_2O with 45% D_2O (around the match point of non-deuterated proteins and carbohydrates) and 100% D_2O (where perdeuterated proteins should have the lowest scattering, since the match point is above 100%). We saw that the form factor of D-GbpA was fairly consistent over the three different D_2O concentrations and similar to that of H-GbpA, with comparable R_g for all conditions. We also confirmed that D-GbpA scattered well at 45% D_2O and that the D-GbpA match point was above 100% D_2O, i.e., that D-GbpA could not be matched out even at 100% D_2O due to its high deuteration level. Parameters from SANS are summarized in Table 2.

In order to verify that the catalytic activity was unaffected by deuteration, LPMO activity towards β-chitin was assessed by measuring the masses of chitooligosaccharide products by MS (Figure 5). H-GbpA and D-GbpA yielded the same products, thus activity was maintained after deuteration. We refer to Loose et al.[9] for a more in-depth study of the activity of GbpA and the components involved in catalysis.

Crystal Structures of GbpA-D1 Show no Differences upon Deuteration. The LPMO domains (D1) of both deuterated and non-deuterated GbpA were crystallized (Figure S2), and X-ray diffraction data were collected. Both proteins crystallized in space group $P2_12_12$ with very similar cell parameters. $P2_12_12$ is a relatively high-symmetry space group, which is preferred for NMX, as less angular data are needed to obtain complete data sets. X-ray crystal structures were determined to 1.6 Å resolution and 1.1 Å resolution for H-GbpA-D1 and D-GbpA-D1 (sample produced at ILL), respectively. The high resolution for D-GbpA is especially promising, as crystal packing and diffraction quality are important for NMX.

Both proteins were crystallized in the catalytically active, copper-bound states (Figure 6B), differing from the crystal structure of GbpA (domains 1−3; PDB ID: 2XWX) from Wong et al.,[8] which lacked the copper ion in the active site. Analysis of anomalous scattering from data sets collected at the absorption edges of copper and zinc confirmed the active site to be exclusively occupied by copper (Figure S3 and Tables S1 and S2). The conformation of the histidine brace, which is very similar among copper-free and copper-bound LPMO structures in the Protein Data Bank (see, e.g., PDB IDs: 6IF7,[45] 6RW7,[46] and 5FTZ[47]), is conserved in our structures. Table 3 summarizes the data collection and refinement statistics. Superimposition of the deuterated and hydrogenated GbpA-D1 structures yielded

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

r.m.s.d. values of appr. 0.2 Å for C_α atoms (Figure 6), showing that the fold is not affected by deuteration. In addition, when comparing our X-ray structures with the one from Wong et al.[8] for domains 1−3 of GbpA, the r.m.s.d. values are below 0.3 Å. The crystallization condition thus serves as a good starting point for further optimization for NMX.

CONCLUSIONS

Neutron scattering techniques have great potential to offer deep insights into the molecular structure and function of proteins in complex environments, providing an important complement to other structural biology techniques. Here, we have demonstrated the feasibility of perdeuteration for GbpA, a bacterial colonization factor with LPMO activity, resulting in yields and deuteration levels highly compatible with NMX and SANS without the need for specialized fermentation equipment. We report a new deuteration protocol based on algal extracts and showed that the protein investigated remains structurally uncompromised and active. In addition to first SANS studies of full-length GbpA (GbpA-FL), we succeeded in obtaining well-diffracting crystals of the GbpA LPMO domain (GbpA-D1) in both the deuterated and non-deuterated states. Whereas production of 70% deuterated LPMOs has been achieved before and used for SANS interaction studies,[48] this is to our knowledge the first time that perdeuteration of an LPMO has been achieved, and a more in-depth study of the effect of deuteration of LPMOs has been performed.

This work, alongside numerous other neutron crystallographic studies, further demonstrates the feasibility of NMX and the important role that perdeuteration plays in optimizing the quality of results. Although large crystals are needed for NMX, and optimization of crystal growth is still required, the possibility of using deuterated protein can reduce the volume requirements by up to a factor of 10. With better NMX structures of LPMOs, it is anticipated that the protonation states of key amino acids in and around the active sites will be revealed and that the importance of the water network within the LPMOs can be understood. Given the recently solved crystal structure of full-length VhLPMO10A from *V. campbellii*, a close homologue of GbpA, NMX studies of FL-GbpA and homologues may also become a possibility.[11] Neutron studies with SANS or NR can reveal how LPMOs interact with the carbohydrate substrates and how the LPMO structure adapts to carbohydrate surfaces or fibers.

Our results will thus further enable neutron-based studies of perdeuterated GbpA and LPMOs in general. In addition, other techniques that benefit from isotope labeling, such as NMR spectroscopy, may also benefit from the labeling procedure presented here.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c02168.

> GbpA expression in non-deuterated M9glyc+ medium; LPMO crystals; metal ion identification by anomalous diffraction analysis; anomalous data collection parameters for the characterization of anomalous scatterers in H-GbpA-D1 crystals; and X-ray data collection and refinement statistics for H-GbpA-D1 at 9,320 eV (PDF)

Accession Codes

UniProt ID: Q9KLD5 (Gene gbpA). PDB IDs: 8CC3 and 8CC5.

AUTHOR INFORMATION

Corresponding Author

Ute Krengel − *Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway;* orcid.org/0000-0001-6688-8151; Phone: +47-22855461; Email: ute.krengel@kjemi.uio.no

Authors

H. V. Sørensen − *Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway;* Present Address: Division of Computational Chemistry, Lund University, SE-223 62 Lund, Sweden (H.V.S.)

Mateu Montserrat-Canals − *Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway; Centre for Molecular Medicine Norway, University of Oslo, NO-0318 Oslo, Norway*

Jennifer S. M. Loose − *Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), NO-1340 Ås, Norway*

S. Zoë Fisher − *Science Directorate, European Spallation Source ERIC, SE-221 00 Lund, Sweden; Department of Biology, Lund University, SE-223 62 Lund, Sweden*

Martine Moulin − *Life Sciences Group, Institut Laue-Langevin, 38042 Cedex 9 Grenoble, France*

Matthew P. Blakeley − *Large-Scale Structures Group, Institut Laue-Langevin, 38042 Grenoble, France;* orcid.org/0000-0002-6412-4358

Gabriele Cordara − *Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway;* orcid.org/0000-0001-8029-8043

Kaare Bjerregaard-Andersen − *Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway;* Present Address: Ottilia vej 9, H. Lundbeck A/S, DK-2500 Valby, Denmark (K.B.-A.); orcid.org/0000-0003-3609-3408

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c02168

Author Contributions

K.B.-A. and U.K. conceived the study. H.V.S. developed the protocols at UiO and performed most of the experimental work (including expression, deuteration, SAXS and SANS) with the assistance of M.M.-C., supervised by K.B.-A. and U.K. Deuteration of GbpA-D1 was scaled up at D-Lab (ILL) and DEMAX (ESS) by M.M. and Z.F., respectively. Enzyme activity experiments were carried out by J.S.M.L. at NMBU. H.V.S. performed the SANS experiments. M.M.-C. crystallized hydrogenated and deuterated GbpA-D1, building on previous experiments by K.B.-A. and H.V.S., and solved and refined the X-ray structures, supervised by G.C. and U.K., who also validated the crystal structures. M.M.-C. and G.C. also performed the anomalous diffraction analysis. M.P.B. contributed advice and support in planning the deuteration and neutron scattering experiments. The manuscript was written by H.V.S. and revised by U.K., with additional input from K.B.-A., M.M.-C. and S.Z.F., and was finalized and approved by all authors.

Funding

The project was funded by the Norwegian Research Council (grant no. 272201) and by the University of Oslo (postdoc

position of K.B.-A. and PhD position of M.M.-C.). Most of the work was carried out at the UiO Structural Biology core facilities, which are part of the Norwegian Macromolecular Crystallography Consortium (NORCRYST) and which received funding from the Norwegian INFRASTRUKTUR-program (project no. 245828) as well as from UiO (core facility funds). SAXS experiments were performed at the Norwegian Centre for X-ray Diffraction, Scattering and Imaging (RECX), funded by the Norwegian INFRASTRUKTUR-program (project no. 208896). MS experiments were carried out at the UiO Proteomics Core Facility at the Department of Biosciences, which is a member of the National Network of Advanced Proteomics Infrastructure (NAPI), funded by the Norwegian INFRASTRUKTUR-program (project no. 295910).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Gustav Vaaje-Kolstad (NMBU) for a great collaboration and continued support on the LPMO project and Reidar Lund for supporting the X-ray and neutron scattering aspects of this work. We would further like to acknowledge Michael Haertlein and Trevor Forsyth for advice and comments regarding the experiments at ILL and thank Trevor additionally for comments on the manuscript. X-ray diffraction experiments were performed on beamlines ID23-1 and ID23-2 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France, and on the BioMAX beamline at MAX IV, Lund, Sweden. We are grateful to Local Contacts Alexander Popov and Sylvain Engilberge at the ESRF and Ana Gonzalez at MAX IV for providing assistance in using the beamlines and to Sylvain Prévost for beamline support at Institut Laue-Langevin (ILL). We further thank Tamjidmaa Khatanbaatar for practical help with data collection at MAX IV. We would also like to acknowledge the support of the LU Protein Production Platform for access to labs and equipment. Mass spectrometry-based proteomic analyses were performed by the Proteomics Core Facility, Department of Biosciences, University of Oslo, for which we acknowledge the help of Bernd Thiede. MALDI-TOF MS was performed at NMBU. All other work was performed at the UiO Structural Biology core facilities and RECX.

■ ABBREVIATIONS

AEX, anion-exchange chromatography; EDTA, ethylenediaminetetraacetic acid; ESRF, European Synchrotron Radiation Facility; ESS, European spallation source; GbpA, *N*-acetylglucosamine-binding protein A; D-GbpA, deuterated GbpA; d_{max}, maximum dimension; H-GbpA, hydrogenated (i.e., non-deuterated) GbpA; ILL, Institut Laue-Langevin; IPTG, isopropyl β-D-1-thiogalactopyranoside; LB, Luria Bertani; LPMO, lytic polysaccharide monooxygenase; MALDI-TOF, Matrix-Assisted Laser Desorption/Ionization-Time Of Flight; MM, molecular mass; MS, mass spectrometry; NMR, nuclear magnetic resonance; NMX, neutron macromolecular crystallography; NR, neutron reflectometry; NSD, normalized spatial discrepancy; OD, optical density; PMSF, phenylmethylsulfonyl fluoride; R_g, radius of gyration; SANS, small-angle neutron scattering; SAXS, small-angle X-ray scattering; SDS-PAGE, sodium dodecyl sulphate−polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography; SLD, scattering-length density; TB, terrific broth; Tris, tris(hydroxymethyl)-aminomethane

■ REFERENCES

(1) Vaaje-Kolstad, G.; Houston, D. R.; Riemen, A. H. K.; Eijsink, V. G. H.; van Aalten, D. M. F. Crystal Structure and Binding Properties of the *Serratia marcescens* Chitin-Binding Protein CBP21. *J. Biol. Chem.* **2005**, *280*, 11313−11319.

(2) Vaaje-Kolstad, G.; Westereng, B.; Horn, S. J.; Liu, Z.; Zhai, H.; Sørlie, M.; Eijsink, V. G. H. An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides. *Science* **2010**, *330*, 219−222.

(3) Aachmann, F. L.; Sørlie, M.; Skjåk-Bræk, G.; Eijsink, V. G. H.; Vaaje-Kolstad, G. NMR Structure of a Lytic Polysaccharide Monooxygenase Provides Insight into Copper Binding, Protein Dynamics, and Substrate Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 18779−18784.

(4) Bissaro, B.; Røhr, Å. K.; Müller, G.; Chylenski, P.; Skaugen, M.; Forsberg, Z.; Horn, S. J.; Vaaje-Kolstad, G.; Eijsink, V. G. H. Oxidative Cleavage of Polysaccharides by Monocopper Enzymes Depends on H_2O_2. *Nat. Chem. Biol.* **2017**, *13*, 1123−1128.

(5) Hangasky, J. A.; Iavarone, A. T.; Marletta, M. A. Reactivity of O_2 versus H_2O_2 with Polysaccharide Monooxygenases. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 4915−4920.

(6) Hangasky, J. A.; Marletta, M. A. A Random-Sequential Kinetic Mechanism for Polysaccharide Monooxygenases. *Biochemistry* **2018**, *57*, 3191−3199.

(7) Frommhagen, M.; Koetsier, M. J.; Westphal, A. H.; Visser, J.; Hinz, S. W. A.; Vincken, J.-P.; van Berkel, W. J. H.; Kabel, M. A.; Gruppen, H. Lytic Polysaccharide Monooxygenases from *Myceliophthora thermophila* C1 Differ in Substrate Preference and Reducing Agent Specificity. *Biotechnol. Biofuels* **2016**, *9*, 186.

(8) Wong, E.; Vaaje-Kolstad, G.; Ghosh, A.; Hurtado-Guerrero, R.; Konarev, P. V.; Ibrahim, A. F. M.; Svergun, D. I.; Eijsink, V. G. H.; Chatterjee, N. S.; van Aalten, D. M. F. The *Vibrio cholerae* Colonization Factor GbpA Possesses a Modular Structure that Governs Binding to Different Host Surfaces. *PLoS Pathog.* **2012**, *8*, e1002373.

(9) Loose, J. S. M.; Forsberg, Z.; Fraaije, M. W.; Eijsink, V. G. H.; Vaaje-Kolstad, G. A Rapid Quantitative Activity Assay Shows that the *Vibrio cholerae* Colonization Factor GbpA is an Active Lytic Polysaccharide Monooxygenase. *FEBS Lett.* **2014**, *588*, 3435−3440.

(10) Bhowmick, R.; Ghosal, A.; Das, B.; Koley, H.; Saha, D. R.; Ganguly, S.; Nandy, R. K.; Bhadra, R. K.; Chatterjee, N. S. Intestinal Adherence of *Vibrio cholerae* Involves a Coordinated Interaction between Colonization Factor GbpA and Mucin. *Infect. Immun.* **2008**, *76*, 4968−4977.

(11) Zhou, Y.; Wannapaiboon, S.; Prongjit, M.; Pornsuwan, S.; Sucharitakul, J.; Kamonsutthipaijit, N.; Robinson, R. C.; Suginta, W. Structural and Binding Studies of a New Chitin-Active AA10 Lytic Polysaccharide Monooxygenase from the Marine Bacterium *Vibrio campbellii*. *Acta Crystallogr. D Struct. Biol.* **2023**, *79*, 479−497.

(12) Schröder, G. C.; Meilleur, F. Metalloprotein Catalysis: Structural and Mechanistic Insights into Oxidoreductases from Neutron Protein Crystallography. *Acta Crystallogr. D Struct. Biol.* **2021**, *77*, 1251−1269.

(13) Blakeley, M. P. Neutron Macromolecular Crystallography. *Crystallogr. Rev.* **2009**, *15*, 157−218.

(14) Bacik, J.-P.; Mekasha, S.; Forsberg, Z.; Kovalevsky, A. Y.; Vaaje-Kolstad, G.; Eijsink, V. G. H.; Nix, J. C.; Coates, L.; Cuneo, M. J.; Unkefer, C. J.; Chen, J. C.-H. Neutron and Atomic Resolution X-Ray Structures of a Lytic Polysaccharide Monooxygenase Reveal Copper-Mediated Dioxygen Binding and Evidence for N-Terminal Deprotonation. *Biochemistry* **2017**, *56*, 2529−2532.

(15) O'Dell, W. B.; Agarwal, P. K.; Meilleur, F. Oxygen Activation at the Active Site of a Fungal Lytic Polysaccharide Monooxygenase. *Angew. Chem. Int. Ed.* **2017**, *56*, 767−770.

(16) Schröder, G. C.; O'Dell, W. B.; Webb, S. P.; Agarwal, P. K.; Meilleur, F. Capture of Activated Dioxygen Intermediates at the Copper-Active Site of a Lytic Polysaccharide Monooxygenase. *Chem. Sci.* **2022**, *13*, 13303−13320.

(17) Tandrup, T.; Lo Leggio, L.; Meilleur, F. Joint X-Ray/Neutron Structure of *Lentinus similis* AA9_A at Room Temperature. *Acta Crystallogr. F Struct. Biol. Commun.* **2023**, *79*, 1−7.

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

(18) Haertlein, M.; Moulin, M.; Devos, J. M.; Laux, V.; Dunne, O.; Forsyth, V. T. Biomolecular Deuteration for Neutron Structural Biology and Dynamics. *Methods Enzymol.* **2016**, *566*, 113−157.
(19) Cai, M.; Huang, Y.; Yang, R.; Craigie, R.; Clore, G. M. A Simple and Robust Protocol for High-Yield Expression of Perdeuterated Proteins in *Escherichia coli* Grown in Shaker Flasks. *J. Biomol. NMR* **2016**, *66*, 85−91.
(20) Duff, A. P.; Wilde, K. L.; Rekas, A.; Lake, V.; Holden, P. J. Robust High-Yield Methodologies for ^{2}H and $^{2}H/^{15}N/^{13}C$ Labeling of Proteins for Structural Investigations Using Neutron Scattering and NMR. *Methods Enzymol.* **2015**, *565*, 3−25.
(21) Koruza, K.; Lafumat, B.; Végvári, Á.; Knecht, W.; Fisher, S. Z. Deuteration of Human Carbonic Anhydrase for Neutron Crystallography: Cell Culture Media, Protein Thermostability, and Crystallization Behavior. *Arch. Biochem. Biophys.* **2018**, *645*, 26−33.
(22) Kightlinger, W.; Chen, K.; Pourmir, A.; Crunkleton, D. W.; Price, G. L.; Johannes, T. W. Production and Characterization of Algae Extract from *Chlamydomonas reinhardtii*. *Electron. J. Biotechnol.* **2014**, *17*, 14−18.
(23) Koruza, K. Perdeuteration of Biological Macromolecules: A Case Study of Human Carbonic Anhydrases. Ph.D. Thesis, Lund University, Lund, Sweden, 2019.
(24) Watanabe, M. M. Freshwater Culture Media. In *Algal Culturing Techniques*; Andersen, R. A., Ed.; Elsevier Academic Press, 2005; 13−20.
(25) Pedersen, J. S. A Flux- and Background-Optimized Version of the NanoSTAR Small-Angle X-Ray Scattering Camera for Solution Scattering. *J. Appl. Crystallogr.* **2004**, *37*, 369−380.
(26) Glatter, O. A New Method for the Evaluation of Small-Angle Scattering Data. *J. Appl. Crystallogr.* **1977**, *10*, 415−421.
(27) Konarev, P. V.; Volkov, V. V.; Sokolova, A. V.; Koch, M. H. J.; Svergun, D. I. *PRIMUS*: A Windows PC-Based System for Small-Angle Scattering Data Analysis. *J. Appl. Crystallogr.* **2003**, *36*, 1277−1282.
(28) Franke, D.; Petoukhov, M. V.; Konarev, P. V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H. D. T.; Kikhney, A. G.; Hajizadeh, N. R.; Franklin, J. M.; Jeffries, C. M.; Svergun, D. I. *ATSAS 2.8*: A Comprehensive Data Analysis Suite for Small-Angle Scattering from Macromolecular Solutions. *J. Appl. Crystallogr.* **2017**, *50*, 1212−1225.
(29) Franke, D.; Svergun, D. I. *DAMMIF*, a Program for Rapid *ab-initio* Shape Determination in Small-Angle Scattering. *J. Appl. Crystallogr.* **2009**, *42*, 342−346.
(30) Volkov, V. V.; Svergun, D. I. Uniqueness of *ab initio* Shape Determination in Small-Angle Scattering. *J. Appl. Crystallogr.* **2003**, *36*, 860−864.
(31) Svergun, D. I. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. *Biophys. J.* **1999**, *76*, 2879−2886.
(32) Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.-C.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. Protein Identification and Analysis Tools in the ExPASy Server. In *Methods Mol. Biol.*; 1999; Vol. *112*: 2-D Proteome Analysis Protocols, 531−552.
(33) Whitten, A. E.; Cai, S.; Trewhella, J. *MULCh*: Modules for the Analysis of Small-Angle Neutron Contrast Variation Data from Biomolecular Assemblies. *J. Appl. Crystallogr.* **2008**, *41*, 222−226.
(34) Incardona, M.-F.; Bourenkov, G. P.; Levik, K.; Pieritz, R. A.; Popov, A. N.; Svensson, O. *EDNA*: A Framework for Plugin-Based Applications Applied to X-Ray Experiment Online Data Analysis. *J. Synchrotron Radiat.* **2009**, *16*, 872−879.
(35) Gildea, R. J.; Beilsten-Edmands, J.; Axford, D.; Horrell, S.; Aller, P.; Sandy, J.; Sanchez-Weatherby, J.; Owen, C. D.; Lukacik, P.; Strain-Damerell, C.; Owen, R. L.; Walsh, M. A.; Winter, G. *xia2.multiplex*: A Multi-Crystal Data-Analysis Pipeline. *Acta Crystallogr. D Struct. Biol.* **2022**, *78*, 752−769.
(36) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the *CCP4* Suite and Current Developments. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 235−242.
(37) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *Phaser* Crystallographic Software. *J. Appl. Crystallogr.* **2007**, *40*, 658−674.
(38) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 486−501.
(39) 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. *REFMAC5* for the Refinement of Macromolecular Crystal Structures. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 355−367.
(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(41) Vonrhein, C.; Flensburg, C.; Keller, P.; Sharff, A.; Smart, O.; Paciorek, W.; Womack, T.; Bricogne, G. Data Processing and Analysis with the *autoPROC* Toolbox. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 293−302.
(42) Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125−132.
(43) Anderson, E. H. Growth Requirements of Virus-Resistant Mutants of *Escherichia coli* Strain "B". *Proc. Natl. Acad. Sci. U. S. A.* **1946**, *32*, 120−128.
(44) Meilleur, F.; Weiss, K. L.; Myles, D. A. A. Deuterium Labeling for Neutron Structure-Function-Dynamics Analysis. *Methods Mol. Biol.* **2009**, *544*, 281−292.
(45) Yadav, S. K.; Archana; Singh, R.; Singh, P. K.; Vasudev, P. G. Insecticidal Fern Protein Tma12 Is Possibly a Lytic Polysaccharide Monooxygenase. *Planta* **2019**, *249*, 1987−1996.
(46) Fowler, C. A.; Sabbadin, F.; Ciano, L.; Hemsworth, G. R.; Elias, L.; Bruce, N.; McQueen-Mason, S.; Davies, G. J.; Walton, P. H. Discovery, Activity and Characterisation of an AA10 Lytic Polysaccharide Oxygenase from the Shipworm Symbiont *Teredinibacter turnerae*. *Biotechnol. Biofuels* **2019**, *12*, 232.
(47) Chaplin, A. K.; Wilson, M. T.; Hough, M. A.; Svistunenko, D. A.; Hemsworth, G. R.; Walton, P. H.; Vijgenboom, E.; Worrall, J. A. R. Heterogeneity in the Histidine-Brace Copper Coordination Sphere in Auxiliary Activity Family 10 (AA10) Lytic Polysaccharide Monooxygenases. *J. Biol. Chem.* **2016**, *291*, 12838−12850.
(48) Bodenheimer, A. M.; O'Dell, W. B.; Oliver, R. C.; Qian, S.; Stanley, C. B.; Meilleur, F. Structural Investigation of Cellobiose Dehydrogenase IIA: Insights from Small Angle Scattering into Intra- and Intermolecular Electron Transfer Mechanisms. *Biochim. Biophys. Acta Gen. Subj.* **2018**, *1862*, 1031−1039.

https://doi.org/10.1021/acsomega.3c02168
ACS Omega 2023, 8, 29101−29112

SUPPORTING INFORMATION

Perdeuterated GbpA enables neutron scattering experiments of a lytic polysaccharide monooxygenase

Henrik Vinther Sørensen[1], Mateu Montserrat-Canals[1,2], Jennifer Loose[5], Zoë Fisher[4,5], Martine Moulin[6], Matthew P. Blakeley[7], Gabriele Cordara[1], Kaare Bjerregaard-Andersen[1,§], Ute Krengel[1*]

[1] Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway

[2] Centre for Molecular Medicine Norway, University of Oslo, NO-0318 Oslo, Norway

[3]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), NO-1340 Ås, Norway.

[4] Science Directorate, European Spallation Source ERIC, P.O. Box 176, SE-221 00 Lund, Sweden.

[5] Department of Biology, Lund University, 35 Sölvegatan, SE-223 62 Lund, Sweden

[6] Life Sciences Group, Institut Laue-Langevin, 71 avenue des Martyrs, 38042 Cedex 9, Grenoble, France.

[7] Large-Scale Structures group, Institut Laue-Langevin, 71 avenue des Martyrs, 38042 Grenoble, France.

[§] Present address: Ottilia vej 9, H. Lundbeck A/S, 2500 Valby, Denmark

[*]Correspondence: Ute Krengel (ute.krengel@kjemi.uio.no; +47-22855461)

LIST OF MATERIAL INCLUDED:

Figures S1-S2 (S1, GbpA expression in non-deuterated M9glyc+ medium; S2, Crystals)

SUPPORTING FIGURES

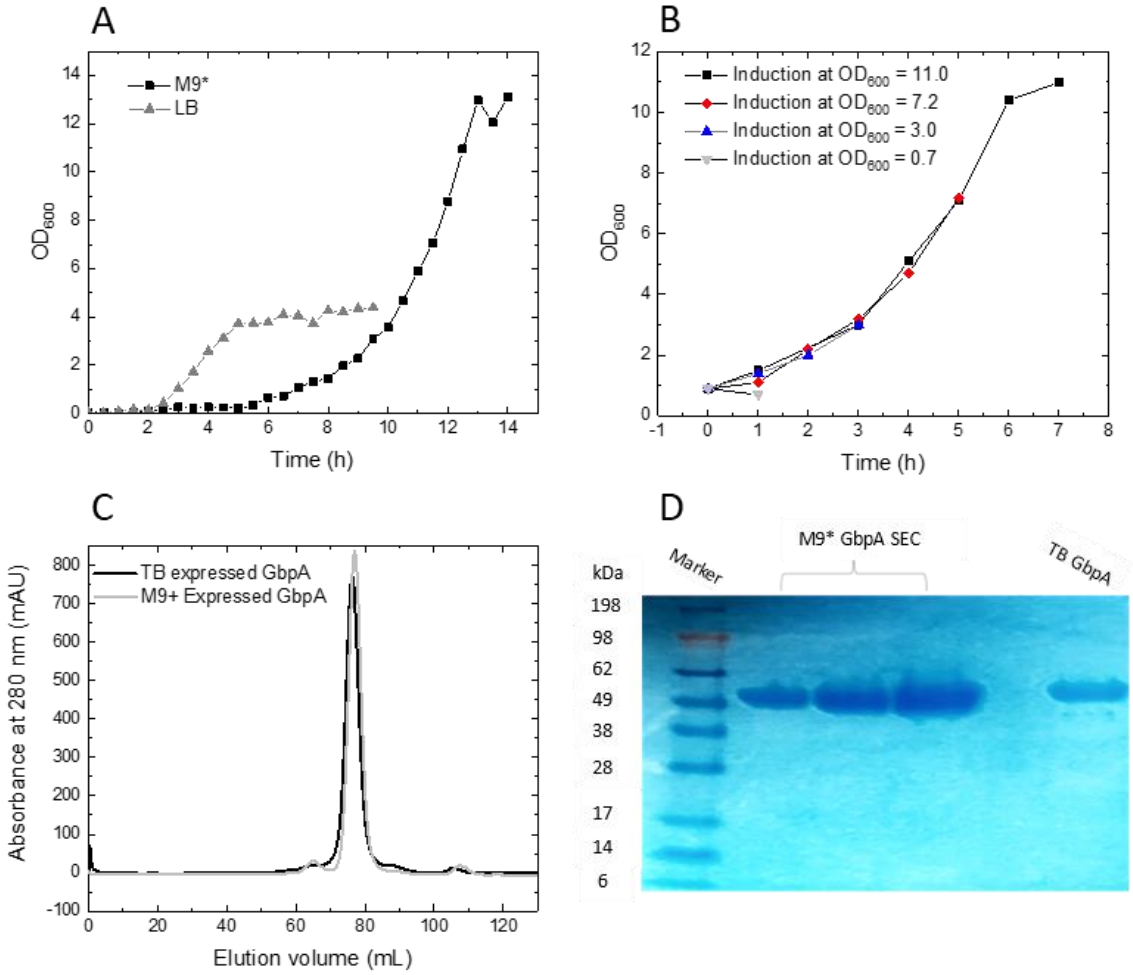


Figure S1. GbpA expression in non-deuterated M9glyc+ medium. A Growth curves for *E. coli* BL21(DE3) cells containing GbpA-FL-encoding plasmid. Luria Bertani (LB) compared to minimal medium (non-deuterated M9glyc+). **B** Growth curves in minimal medium (non-deuterated M9glyc+) up to different induction points. All cultures were "boosted" by LB pre-cultures. Expression was initiated by the addition of IPTG at four different optical densities. **C** SEC elution profile for GbpA expressed in TB or minimal media show comparable elution profiles and retention volumes. **D** SDS-PAGE of GbpA expressed in TB or minimal media, both showing high purity. Marker: SeeBlue plus 2.

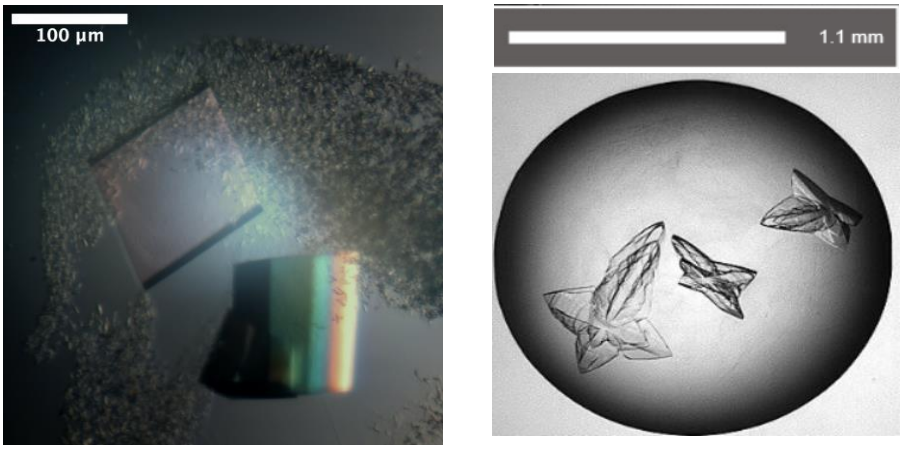


Figure S2: Crystals. Examples of crystals of H-GbpA-D1 (**A**) and D-GbpA-D1 (**B**), corresponding to the same crystallization conditions and space group.

III

IV

V

FEBS openbio

FEBS PRESS science publishing by scientists

Structure prediction of honey bee vitellogenin: a multi-domain protein important for insect immunity

Vilde Leipart[1], Mateu Montserrat-Canals[2], Eva S. Cunha[2], Hartmut Luecke[3], Elías Herrero-Galán[4,*], Øyvind Halskau[5] and Gro V. Amdam[1,6]

1 Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Aas, Norway
2 Norwegian Center for Molecular Medicine, University of Oslo, Norway
3 Department of Physiology and Biophysics, University of California, Irvine, CA, USA
4 Department of Structure of Macromolecules, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain
5 Department of Biological Sciences, University of Bergen, Norway
6 School of Life Sciences, Arizona State University, Tempe, AZ, United States

Keywords
homology modeling; honey bee vitellogenin; rigid-body fitting; von Willebrand factor domain

Correspondence
V. Leipart, Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Høgskoleveien 12, 1430 Ås, Norway
Tel: +47 99444807
E-mail: vilde.leipart@nmbu.no

Present address
E. Herrero-Galan, Molecular Mechanics of the Cardiovascular System Cell and Developmental Biology Area, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Instituto de Salud Carlos III, C/ Melchor Fernández Almagro, Madrid, Spain

(Received 27 May 2021, revised 27 September 2021, accepted 18 October 2021)

doi:10.1002/2211-5463.13316

Edited by Cláudio Soares

Vitellogenin (Vg) has been implicated as a central protein in the immunity of egg-laying animals. Studies on a diverse set of species suggest that Vg supports health and longevity through binding to pathogens. Specific studies of honey bees (*Apis mellifera*) further indicate that the *vitellogenin* (*vg*) gene undergoes selection driven by local pathogen pressures. Determining the complete 3D structure of full-length Vg (flVg) protein will provide insights regarding the structure–function relationships underlying allelic variation. Honey bee Vg has been described in terms of function, and two subdomains have been structurally described, while information about the other domains is lacking. Here, we present a structure prediction, restrained by experimental data, of flVg from honey bees. To achieve this, we performed homology modeling and used AlphaFold before using a negative-stain electron microscopy map to restrict, orient, and validate our 3D model. Our approach identified a highly conserved Ca^{2+}-ion-binding site in a von Willebrand factor domain that might be central to Vg function. Thereafter, we used rigid-body fitting to predict the relative position of high-resolution domains in a flVg model. This mapping represents the first experimentally validated full-length protein model of a Vg protein and is thus relevant for understanding Vg in numerous species. Our results are also specifically relevant to honey bee health, which is a topic of global concern due to rapidly declining pollinator numbers.

Abbreviations
BN-PAGE, blue native polyacrylamide gel electrophoresis; CCS, cross-correlation score; DAMPs, damage-associated molecular patterns; DUF1943/1944, domain of unknown function 1943/1944; EM, electron microscopy; fbVg, fat body Vg; flVg, full-length Vg; LC, lower cavity; MSA, multiple sequence alignment; MTP, microsomal triglyceride transfer protein; ND, N-terminal domain; PAMPs, pathogen-associated molecular patterns; QMEAN, Qualitative Model Energy Analysis; SEC, size exclusion chromatography; SPDBV, Swiss-PdbViewer; UC, upper cavity; VADAR, volume, area, dihedral, angle reporter; Vg, vitellogenin; vWF, von Willebrand factor; Ω, gap region.

Vitellogenin (Vg) belongs to an ancient and phylogenetically broad protein family called large lipid transfer proteins [1]. In most egg-laying animals, Vg contributes to oogenesis by providing lipids. Over the last 20 years, studies of several species have demonstrated additional functions of this superfamily in health and behavior [2]. Many animals with one or more *vg* genes are commercially important, and this has incentivized analyses of reproductive and immune traits in which Vg is likely to play a role. Effects of Vg on host immunity have been studied in animals as diverse as bees and fishes [3,4]. For example, Vg recognizes gram-positive bacteria (i.e., *Staphylococcus aureus*, *Micrococcus luteus,* and *Bacillus subtilis*) and gram-negative bacteria (i.e., *Escherichia coli* and *Vibrio anguillarum*) in nonbilaterian coral (*Euphyllia ancora*) and zebrafish (*Danio rerio*) [5,6]. These studies also show that Vg recognizes general bacterial and fungal pathogen-associated molecular patterns (PAMPs). Antimicrobial activity was not detected in these studies, but the interaction promotes apoptosis. Zhang et al. [4] suggest that Vg in zebrafish functions as an inflammatory acute-phase protein leading to elimination of pathogens. This finding also applies to honey bees (*Apis mellifera*) where Vg appears to have similar immunological binding properties [7]. In addition, the Vg molecule of honey bees recognizes damage-associated molecular patterns (DAMPs) [3] and displays antioxidant activity [8–10].

The honey bee is one of the best studied species in terms of the diverse roles of Vg [8,11,12]. For example, this animal was used to show that via their eggs, females can protect their offspring against diseases using a Vg-mediated transfer mechanism: Fragments of bacterial cell walls (immune elicitors) are recognized by Vg and carried out to the honey bee eggs during oogenesis [7,13]. This phenomenon of trans-generational immune priming without the use of antibody-based (i.e., acquired) immunity was first detected a decade ago [14]. However, the underlying mechanisms were not understood before Vg was proposed as a causal element [7]. The availability of the genomic sequence and some functional genetic technologies in honey bees have also enabled studies of Vg's role in behavior [8,15], and such findings have been extended to ants, cockroaches, and mosquitos [16–18]. Honey bees are globally available due to apiculture and can be obtained in large numbers at low costs. Therefore, honey bees provide a practical and useful model for investigating the structure–function relationship of Vg.

In most egg-laying animals, Vg consists of three conserved domains: The N-terminal domain (ND), a domain of unknown function 1943 (DUF1943) and the von Willebrand factor (vWF) type D domain (Fig. S1). In honey bees, the ND is further subcategorized into two structural subdomains, the β-barrel and the α-helical domains, with a highly disordered polyserine region linking these two domains [19] (Fig. S1A). Circulating Vg in the hemolymph of honey bees has a molecular mass of approximately 180 kDa. Vg is cleaved into a 40 and a 150 kDa fragment in the abdominal fat body tissue, the main site for Vg synthesis and storage, and the polyserine linker has been identified as the cleavage site [19]. During investigation of pathogen recognition of Vg in honey bees, the full-length hemolymph Vg (flVg) and the 150 kDa fat body Vg (fbVg) subunit, together with a recombinant peptide of the α-helical domain, were shown to recognize dead and damaged cells [3]. The authors suggest that the heavily positively charged α-helical domain is the main contributor to pathogen recognition. The same study also includes a recombinant peptide of vWF, but this synthetic domain did not show similar binding activity. Studies in fishes and one coral species confirm that the ND can recognize PAMPs and DAMPs but also show that the DUF1943 and vWF can contribute to pathogen recognition [5,6]. Taken together, these findings indicate that Vg may have multiple pathogen-recognizing domains.

In vertebrates and invertebrates, the three main structural domains of Vg are highly conserved at the structural level [5] despite a low nucleic acid sequence similarity [1]. This conservation indicates that the main features of the Vg amino acid sequence are maintained by natural selection. At the level of nucleic acids, the β-barrel subdomain is the most conserved region of the honey bee *vg* gene, while the presumed lipid-binding region (α-helical domain and DUF1943) undergoes positive selection [20]. In a previous study, five residue positions were identified as candidates of functional polymorphisms (marked in Fig. S1A). Local pathogen pressure can be a significant selective force [21–23], and several studies suggest that Vg structure adapts to more efficiently recognize such local threats [7,12]. This hypothesis relies on structure–function relationships that are not fully understood. In fact, there is no complete and detailed structure of the full-length Vg (flVg) protein in any bee, insect, coral, or modern fish species. The only experimentally solved structure is that of lamprey (*Ichthyomyzon unicuspis*) Vg (PDB ID: 1LSH [24]), which consists only of the lipovitellin light and heavy chain (ca. 76% of the sequence is crystallized; Fig. S1B). Using this information as a resource, the conserved N-terminal subdomains (β-barrel and α-helical) in honey bees were

described using homology modeling [3,25] with lamprey Vg as a template. This approach has not been extended to the less conserved DUF1943 domain that is also present in lamprey. The vWF homologous domain, β-Component, is absent from the lamprey crystallographic structure, which eliminates lamprey as a possible template for homology modeling of the vWF domain in other species like honey bees.

Solving the structure of Vg in more species can increase our understanding of ligand interactions and provide important insights into structure–function relationships. However, even in otherwise well-studied species like honey bees, this centrally important information on the DUF1943 and vWF domain is lacking.

Fortunately, the number of experimentally solved protein structures is growing, and the computational modeling software is becoming more powerful. For example, a crystallographic protein structure of the D'D3 assembly in human vWF protein was resolved in 2019 [26], and the VWD3 domain in this assembly has a pairwise sequence identity slightly above 20% to the honey bee domain, which is sufficient to be used as a template during homology modeling.

In this study, we make progress in describing the structure and interpreting the function of the vWF domain in honey bees. In addition, we compile results from template-based, deep learning modeling methods, and the ground-breaking neural network-based algorithm, AlphaFold [27], to present, for the first time, a full-length model for an invertebrate Vg. We combine this new information with published data to begin to elucidate the domain assembly of flVg. Our findings suggest that vWF contributes to the structural organization and has a previously undescribed and valuable function in the protein. This study contributes to the understanding of a protein that is central to life in many animal species.

Materials and methods

Identification of templates

The full-length honey bee Vg sequence (UniProt ID: Q868N5) was inputted to the HHpred [28] server with default settings, which included 'PDB_mmCIF70_23_Jul' as the target database. HHpred returned 250 hits. Each hit was evaluated based on the sequence identity. For the vWF domain, the structural template was verified by performing a BLAST of honey bee Vg (UniProt ID: Q868N5) against the UniProtKB. The target database was restricted to only include UniProt sequences having a PDB ID. The query was run with default settings (*e*-threshold: 10, matrix: auto, filtering: none, gapped: yes, hits: 1000). This BLAST returned 26 hits, and hits from regions already satisfactorily modeled in earlier work were ignored. The remaining hits included the VWF_HUMAN (UniProt ID: P04275, *e*-value 7.2e-1, and 25.0% sequence identity). Residues 1453–1612 of the vWF domain in Vg were aligned to residues 864–1013 of vWF, *Homo sapiens*. These residues correspond to the WD3 domain in the D'D3 assembly in the human vWF protein.

Structural alignment and homology modeling of the von Willebrand factor domain

Both the target and template sequence are part of two larger assemblies, each comprising 4 and 12 domains, respectively. To identify the correct start and end points of the structural alignments, 16 alignments with different sequence lengths were performed. The highest sequence identity (26.3%) was obtained by aligning residues 1440–1634 (target) with residues 836–1031 (template) using the Emboss Needle pairwise alignment tool [29,30], with default settings (Table S1). To ensure that the functional and important regions were aligned correctly, the pairwise alignment was supplemented with a multiple sequence alignment (MSA). The MSA was executed using BLAST and representative Vg sequences from a wider selection of 16 species [3] (Table S2). To ensure a correct alignment of the full-length vWF *H. sapiens* in the MSA and not cause confusion among the four VWD modules in the protein, we referenced the alignment of the modules in the D assemblies from Dong et al. [26] (Fig. 2). The pairwise alignment was altered so that gaps were in the same positions as in the low-conserved regions of the MSA. The highly conserved residues were correctly aligned and were not altered. To avoid gaps in secondary structures or binding sites, the secondary structure annotations from template 6N29 were added to the alignment.

The homology model was interactively built using Swiss-PdbViewer [31] (SPDBV; v. 4.1.0), a recommended approach when building target models with low sequence identity to the template [32]. To initiate the modeling project, the raw sequence (Q868N5) was fitted onto the 3D coordinates of the template (PDB ID: 6N29). Backbone building was performed automatically after editing the alignment as described above. *Ab initio* loop building was performed to ligate breaks in the backbone caused by gaps in the alignment (insertions/deletions). The loop option with the lowest clash and energy scores was chosen in all cases. In this way, nine loops were inserted into the model (Table S3), leaving three unsolved regions (residues 1494–1504, 1515–1522, and 1537–1541) missing in the model. *Ab initio* and database loop building attempts failed to produce a reasonable output for these three 8–11 residue-long gaps. Side chain conformations of target residues aligned to residues with dissimilar characteristics in the template were

identified by detecting clashes and rearranged into the most optimal rotamer option. Rotamer libraries of the most observed orientations for side chains are included in the program. The entire model was energy minimized through a partial implementation of the GROMOS96 force field [33] integrated in the SPDBV software.

Quality control of the von Willebrand factor homology model

Quality control was performed on the model to determine whether the structural features are consistent with the physiochemical rules. Stereochemical consistency was evaluated residue-by-residue using PROCHECK [34]. Global and local quality estimates were performed using the Qualitative Model Energy Analysis (QMEAN) server [35], powered by SWISS-MODEL. The QMEAN output *Z*-score compares the query to similar values based on X-ray structures. VADAR (v. 1.8) [36] assesses the 3D profile, stereo/packing, accessible surface and residue volume. Based on these quality assessments, manual editing was applied to the residues listed in Table S4. The final model was deposited to ModelArchive and can be accessed at: https://modelarchive.org/doi/10.5452/ma-sfueo (access code: okHs98Pcl2).

The Ca^{2+}-ion was copied from the template to the target model, and the contacts to the binding residues were verified to be reasonable in PYMOL (v. 2.2.2) [37]. All illustrations of the model were made in PYMOL.

Full-length structure prediction of honey bee vitellogenin

The alignments from HHpred with the highest sequence identity were selected and forwarded to the implemented modeling software MODELLER [38]. Models 1–8 were built using the query sequences listed in Table 1. All models were built using default settings. A full-length prediction was also built using the RAPTORX web server [39] with the full-length honey bee Vg sequence (UniProt ID: Q868N5) as input, which generated a structure consisting of six domains, each built using one to five templates or template-free modeling (Table S7 and Fig. S7). The models were visualized with the program PYMOL and aligned, and the final model was assembled and built here.

To run AlphaFold v2.0 ([27], see Jumper et al. (2021) supplementary material for detailed description of the method), a P3.2xlagre instance was provisioned from AWS EC2, using the Deep Learning AMI (Ubuntu 18.04) Version 48.0 and a 300 GB disk. Additionally, a 4TB gp3 EBS volume, with 400 $MB{\cdot}s^{-1}$ of throughput and 3000 IOPS, was provisioned and mounted on the machine. The step-by-step guide (README.md, https://github.com/deepmind/alphafold) was followed for setting up and running AlphaFold using Docker. Dependencies that were not included in the AMI were installed manually using the apt package manager. The input sequence was UniProt ID: Q868N5, and AlphaFold was run with the full_dbs preset. Model parameters, downloaded databases, and the output files were stored on the 4TB EBS volume. The run resulted in five models, ranked by average plDDT (Fig. S8B,C). The PDB-file of the top ranked model is included in Appendix S2.

Rigid-body fitting into the electron microscopy map

The high-resolution full-length model and separate chains, in addition to two previously published homology models [3,25] and lamprey Vg (PDB ID: 1LSH) [24], were fitted into the low-resolution negative-stain electron microscopy (EM) map (Fig. S9, EMDB-22113, deposited) without

Table 1. Structure predictions generated by MODELLER and RAPTORX. The table presents all the models generated using MODELLER and RAPTORX (Figs S6 and S7) and lists the region of the amino acid sequence (aa seq.) that has been modeled and which domain it represents. The template used for the model (protein name, species, and PDB ID) and the sequence identity are listed. For Model 9, several templates have been used to generate the full-length model.

Model	Honey bee Vg aa seq.	Honey bee Vg domain	Template	Seq. iden. (%)
1	21–1059	ND and DUF1943	Lamprey Vg (PDB ID: 1LSH_A)	16
2	1190–1515	Undetermined and partly vWF	Lamprey Vg (PDB ID: 1LSH_B)	15
3	1442–1632	vWF	Human vWF (PDB ID: 6N29)	22
4	21–323	β-barrel	Lamprey Vg (PDB ID: 1LSH_A)	19
5	324–360	Polyserine linker	Honey bee Vg (PDB ID: 2ILC)	97
6	361–756	α-helical	Lamprey Vg (PDB ID: 1LSH_A)	19
7	760–1059	DUF1943	Human MTP (PDB ID: 6I7S)	13
8	760–1059	DUF1943	Lamprey Vg (PDB ID: 1LSH_A)	11
9	1–1770	Full-length Vg	PDB ID: 1LSH_A, 1LSH_B, 6RBF_A, 3WJB_A, 4YU8_A, 4JPH_A, 5BPA, 4NT5_A and 2KD3_A	12, 21, 8, 6, 5, 9, 10, 14 and 7

direct human intervention by using the PowerFit webserver [40,41] and the ADP_EM plugin in CHIMERA [42]. In both methods, the resolution was set to 27 Å based on the Fourier shell correlation curve (Fig. S9C), and for PowerFit, the rotational sampling interval parameter was set to 5.00. The PowerFit algorithm uses the cross-correlation between the EM map and the structure to be fitted to search for optimal fits. Output was provided as the structural model's orientation with a corresponding goodness of fit score. ADP_EM works similarly, but is optimized for low-resolution density maps. The fits were imported to the program UCSF CHIMERA (v. 1.14) [43] to optimize them using the volume data 'Fit-in-map' function. This function calculates a correlation score and an average map value both based on map grid points, but the former calculates overlap, while the latter only focuses on the atoms inside the map. In addition, the number of atoms outside the contour is shown. The setting was left as default, but the resolution of 27 Å was inputted. All resulting scores from both software systems are presented in Tables S5 and S6.

CHIMERA and PYMOL were also used to generate the figures of the fits and apply a hydrophobicity scale [44]. The final assembly was imported to PYMOL, where it was aligned to lamprey Vg (PDB ID: 1LSH). The generate symmetry function in PYMOL was used to produce the dimer formation presented by Anderson et al. [53] of lamprey Vg and aligned the final assembly to this structure to present the dimer of honey bee Vg (Fig. 4E). The conserved residues creating polar contacts in honey bee Vg were identified using the MSA produced by MODELLER (not shown). The distances of polar contacts were measured in PYMOL.

Purification of vitellogenin from honey bees

To obtain purified Vg, we collected 1–10 µL honey bee hemolymph in a 1 : 10 dilution in 0.5 M Tris/HCl pH 7.6, using BD needles (30 G) as described earlier [45]. The dilution was filtered using a 0.2 µm syringe filter. Vg was purified from honey bee hemolymph with ion-exchange chromatography using a HiTrap Q FF 1 mL column 0.5 M Tris/HCl as the sample buffer and 0.5 M Tris/HCl with 0.45 M NaCl as the elution buffer. 400–450 µL diluted hemolymph was manually injected and Vg eluted at a conductivity of 15–22 $mS \cdot cm^{-1}$. All fractions from this peak were collected, pooled and concentrated using an Amicon® Ultracel 100 kDa membrane centrifuge filter (Merck KGaA, Darmstadt, Germany). The fraction purity was verified by running SDS/PAGE, which contained only one band of the correct size (~ 180 kDa). The protein concentration was measured with Qubit.

Native gel and size exclusion chromatography

Blue native polyacrylamide gel electrophoresis (BN-PAGE) was performed at 4 °C in precast 3–12% acrylamide gels (Invitrogen, Waltham, MA, USA) for 2 h at a constant voltage of 150 V. The NativePAGE Novex Bis-Tris Gel System (Life Technologies, Carlsbad, CA, USA) protocol was used both for sample and buffer preparation, and Native-PAGE Running Buffer (1×) and the Dark Blue Cathode Buffer (0.4% Coomassie G-250) were used. Size exclusion chromatography (SEC) was performed of Vg in a Superose 6 Increase 3.2/300 column (GE Healthcare, Chicago, IL, USA) at 4 °C equilibrated with a buffer containing 50 mM Tris pH 7.6 and 225 mM NaCl. The SEC was run on an ÄKTA Pure 25 system (GE Healthcare) in micro configuration that allows the use of very small sample volumes. This modification prevents dilution of the sample by effectively reducing the internal volume since it bypasses the multicolumn valve and the pH flow cell and has a shorter path length between the injection valve and the UV monitor. We injected 50 µL of sample (0.26 $mg \cdot mL^{-1}$) and manually collected fractions directly from the outlet of the UV monitor.

Results

Template search

Increased insight into the tertiary structure of Vg's domains is beneficial to our understanding of how Vg contributes to honey bee immunity. To build a full-length structure prediction of honey bee Vg, we first identified potential templates using HHpred [28] (Fig. 1A) with the complete amino acid sequence as input. HHpred indicated that two templates are available for building the ND and DUF1943 domain, one for an undetermined region (residue 1190–1442), and three for the vWF domain. Except for Template 1 (PDB ID: 6N29_A), the sequence identities fall below 20%. By dividing the query sequence into known subdomains and domain boundaries and repeating the search, we generated more specific alignments. The top two ND subdomain templates increased their sequence identities to 19%. In contrast, the DUF1943 was demonstrated to be more distinct compared to human microsomal triglyceride transfer protein (MTP) and lamprey Vg, having sequence identities of only 13% and 11%, respectively.

Homology modeling of the von Willebrand factor domain

Among the three highly conserved domains, the vWF is a major unknown piece in the structural puzzle of Vg. Our initial search discovered a recently published and promising template for this domain, which we confirmed using BLAST [46]. The WD3 domain in the D'D3 assembly of the vWF protein of *H. sapiens* has

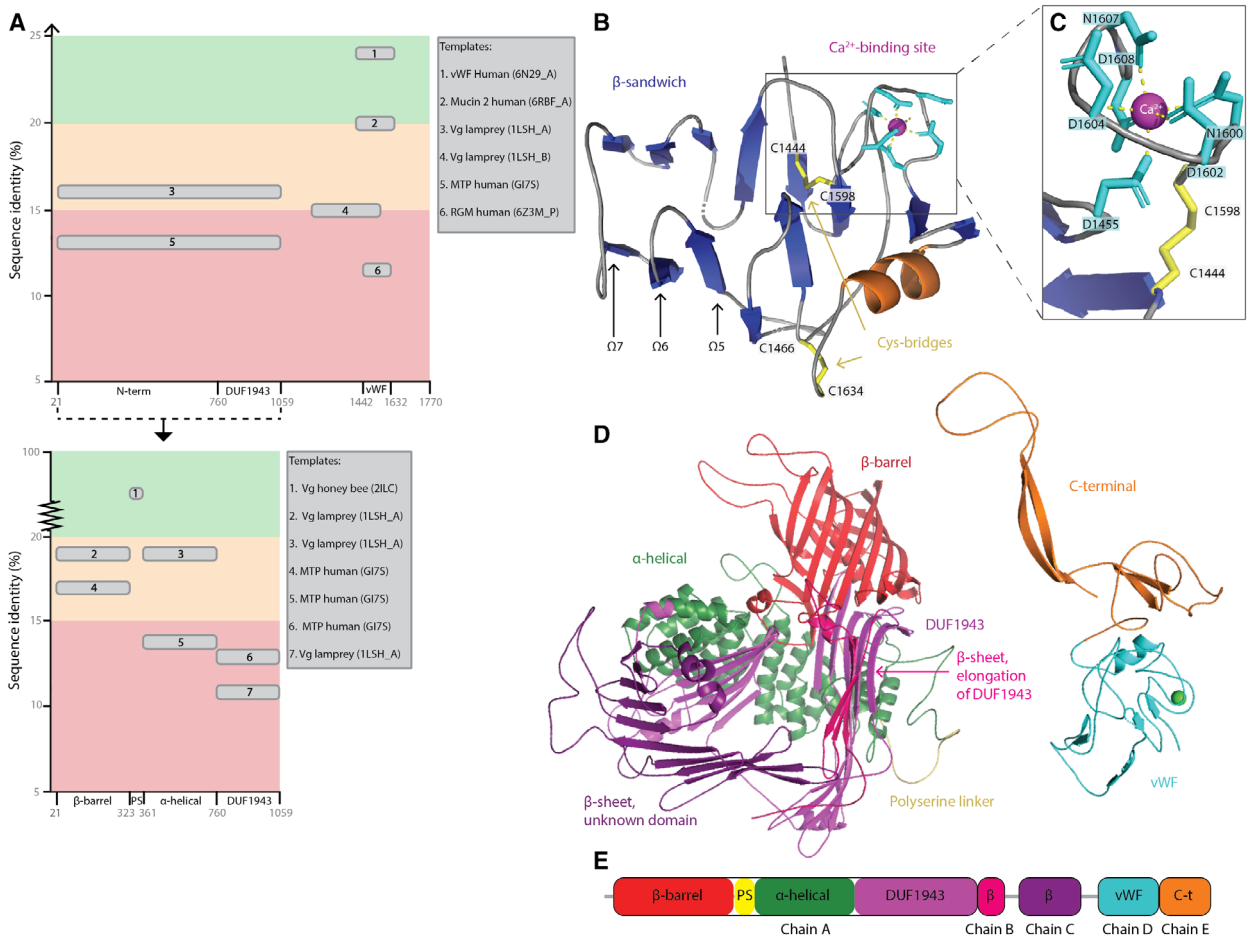


Fig. 1. Structure prediction of honey bee vitellogenin. (A) A graphical illustration of the identified templates using HHpred. On both graphs, the amino acid sequence of honey bee Vg is on the *x*-axis (with the subdomains and domains labeled), and the percentage of sequence identity to the templates is on the *y*-axis. The first graph displays all the templates (gray rounded edge boxes) identified when inputting the full-length sequence of honey bee Vg, while the second shows the templates identified when inputting only the sequence of the separate subdomains. The background colors on both graphs illustrate whether the sequence identity is below 15% (red), between 15% and 20% (orange) or above 20% (green). The templates are numbered according to the sequence identity (highest to lowest), and the protein name, species, and PDB ID are noted in the two large gray boxes. (B) Homology model of vWF: The β-sandwich is on the left side while the Ca^{2+}-segment is on the right side. The Cys-bridges connecting the two segments are shown as yellow sticks and arrows. The β-strands, α-helix and loops are colored blue, orange, and gray, respectively, and the positions of Ω5–7 are labeled with black arrows. The Ca^{2+}-binding residues are shown as cyan sticks, and the Ca^{2+}-ion is shown as a pink sphere. (C) Close-up of the Ca^{2+}-binding site. The coloring scheme is the same as in panel B. All Ca^{2+}-binding residues (D1455, N1600, D1602, D1604, N1607 and D1608) and one of the Cys-bridges (C1598 and C1444) are labeled, and this demonstrates how D1455 from the β-sandwich interacts with the Ca^{2+}-ion. (D) The full-length homology model compiled from several models with different templates. The subdomains and domains are colored as follows: the β-barrel subdomain (red), the polyserine linker (yellow), the α-helical subdomain (forest green), the DUF1943 domain (magenta), elongation of the DUF1943 domain (hot pink), the undetermined structural region (purple), the vWF domain (cyan), and the C-terminal region (orange). (E) A 2D illustration of the chains A to E, used when preforming rigid-body fitting of the homology model.

a sequence identity of the pairwise alignment of 24.1%, which is slightly below the suggested threshold (25%) for creating a reliable homology model [47]. In other words, a pairwise alignment may not be enough to identify gaps and robustly conserved amino acids. We therefore conducted a MSA to confirm gaps and alignment of conserved and domain-defining residues across 12 species, including representative insects, nematodes and mammals. The MSA and the final structural alignment are presented in Fig. S2.

A visual inspection of the structural alignment revealed some interesting aspects. In the almost 200 amino acid-long alignment, the first 40 residues and the last 80 residues are well conserved. In the less

conserved regions, four larger gap regions (Ω) have been introduced (Ω4–7). Ω4 is also missing in all species containing the vWF protein based on the MSA, while downstream Ω5 and Ω7 are conserved in most of the species containing the Vg protein (Fig. S2A). Ω6 seems to be included in all species but is missing in the VWD3, a cysteine-rich domain that forms four intrachain disulfide bridges and two interchain disulfide bridges. The interchain bridges stabilize dimerization of VWD domains in the human vWF protein as opposed to the intrachain bridges formed between cysteine residues inside a single VWD domain. The interchain bridging cysteine residues are not included in the target sequence, and based on the MSA, they are also not conserved in the template domain. However, the eight intrachain bridging cysteine residues are included in the template. Four of these are conserved in the target (C1444, C1466, C1598, and C1634; Fig. 1B). The VWD3 domain also contains a Ca^{2+}-binding site experimentally known from the structural template with key residues also present in the target sequence [26]. We recognize this as a class II calcium binding site because the coordinating residues, as well as the neighboring residues, make up two short regions [48] (r. 1453–1456 and r. 1596–1609; Fig. S2) that are well conserved among all species in the MSA. This indicates an essential site for function and/or stability of the domain. We conclude that the significant regions for domain function or stability, the intrachain disulfide bonds, as well as the Ca^{2+}-binding residues, are conserved and correctly aligned. We also conclude that the MSA was able to identify robustly conserved features of Vg, and we therefore proceeded with interactive homology modeling using the structural alignment provided by the MSA (Fig. S2B).

The amino acid sequence of the target was fitted onto the three-dimensional coordinates of the template using the structural alignment. Breaks in the backbone were ligated using loop building, and the side chains of nonconserved residues were rearranged to the most optimal rotamer orientation, reducing the number of steric clashes. Finally, we performed energy minimization to release local backbone strain and electron density clashes. The overall quality of the target model was validated using several software tools. To account for sequential errors, we also included the quality scores of the template (Figs S3 and S4). Based on the results, the backbone phi and psi angles of 14 residues, detected as outliers by Ramachandran analysis (Fig. S3C) [49], and rotamers of 19 residues, detected by PROCHECK, were manually edited (Table S4). The main limiting factor for the quality metrics of the model were the errors already listed as well as the presence of the longer gap regions. It was not possible to include Ω5–7 in the model because this creates a region with too many unfavorable interactions and torsion angles. However, these regions exhibit low conservation (Fig. S2). The local quality estimate by SWISS-MODEL (Fig. S3B) shows that the middle region is of lower quality relative to the first and last missing regions. The Ca^{2+}-binding residues and intrachain disulfide bonds are in higher-quality regions. The PROCHECK summary shows that the main difference between the target and template models originates from the calculated stereochemical parameters (geometry, bad contacts and bond length and angles; Fig. S3A). The residue-by-residue list produced by PROCHECK (Fig. S4E) identified residues deviating from the ideal values. However, these residues were altered during loop building, often resulting in an unfavorable orientation for the chosen residues [50]. We conclude that key structural features of the target are modeled correctly except for the low-quality middle region that contains residues with stereochemical parameters deviating from the ideal values. The homology modeling approach used has a proven track record of producing models of sufficient quality when facing similar challenges [51]. We demonstrated this by comparing our model to an automatically produced model by MODELLER. We find that in our model, the local quality is better in the regions of low conservation (Fig. S3B), and the global quality is higher (Fig. S4A–C). For the conserved region, our interactive modeling approach achieves a better result by including C1634, which creates an intrachain disulfide bond, two additional β-strands and a more appropriate rotamer option for the Ca^{2+}-binding residue N1607 (Fig. S5).

We are thus for the first time able to present a detailed structural model of the vWF domain of honey bee Vg. The structure can be understood as two segments: one consisting of 11 antiparallel β-strands organized into a β-sandwich while the other is comprised of the Ca^{2+}-binding site, a short α-helix, and three short β-strands (Fig. 1B,C). Connecting the two segments are the two intrachain disulfide bonds. The two segments are also connected through the Ca^{2+}-binding site via the interaction of residue D1455 (Fig. 1C). The Ca^{2+}-binding residues are in loop regions (i.e., normally flexible regions), but we suggest that binding of a Ca^{2+}-ion might confer stability to this region. The Ca^{2+}-binding segment of the domain exhibits higher quality than the antiparallel β-sandwich. Despite the lower quality, the residues in the secondary structure elements exhibit a higher local quality score compared to the residues in the loop regions (Fig. S3B). We conclude that the β-strands are organized in a sterically

reasonable manner, while the loop regions are most likely not described accurately.

Full-length structure prediction of honey bee vitellogenin

We performed template-based prediction of the remaining domains of honey bee Vg using the integrated MODELLER software in HHpred. We generated eight models using different sections of the honey bee Vg amino acid sequence as input (Table 1). By aligning the predicted models covering the same domains (Fig. S6), we observed that the general fold is the same except for models describing DUF1943 (Models 1, 7, and 8; Fig. S6B). Using human MTP as a template returned a straight β-sheet with fewer and longer β-strands. In addition, we also used the deep learning modeling method RAPTORX to generate a full-length and complete prediction (Fig. S7). The model is mostly based on nine different templates with sequence identity ranging from 5% to 21% but also includes regions resulting from deep learning predictions. The total model assembles all predicted domains like pearls on a string and cannot predict how they are organized relative to each other. However, the general fold of each model is consistent with the results from MODELLER (Fig. S6A–E). We built the final structure using Model 1 for residues 21–1059, Model 9 for residues 1060–1140, Model 2 for residues 1190–1408, the vWF homology model from Quality control of the von Willebrand factor homology model for residues 1440–1634 and Model 9 for residues 1635–1770. We selected these models based on whether their fold were consensus folds and removed the long, extending loop regions. The final model has 93.1% sequence coverage of honey bee Vg and includes the conserved domains (ND, DUF1943 and vWF) in addition to undetermined regions now structurally described for the first time for an invertebrate Vg (two β-sheets downstream of DUF1943 and the C-terminal region; Fig. 1D). Based on the compilation of models, the final prediction was divided into chains A (the ND), B (the β-sheet from Model 9), C (the β-sheet from Model 2), D (the vWF domain) and E (the C-terminal region) as presented in Fig. 1E.

The very recent publication and code availability for AlphaFold v2.0 [27] enabled us to produce a structure prediction of honey bee Vg. The first step of the pipeline is to produce an MSA, and the resulting number of hits can indicate the prediction accuracy. The developers observe a decrease in prediction accuracy when the alignment depth falls below 30 sequences and an increase of accuracy until 100 sequences, where they observe a threshold effect [27]. The honey bee Vg MSA have an average of 1988 hits per residue (Fig. S8A), suggesting a high prediction quality. The resulting AlphaFold models had an average predicted local distance difference test (plDDT) ranging from 81.7692 to 84.5747 (Fig. S8B), which is a per-residue estimate of confidence [27,52]. The highest-ranking model colored by the plDDT confidence scale (Fig. 2A) shows a generally confident backbone prediction of honey bee Vg. Some regions fall below 70, which the developers of AlphaFold state should be treated with caution, and these residues map to short loops in domains or longer flexible segments in-between domains (Fig. 2B). The developers state that plDDT residue scores below 50 strongly indicate disorder which in our case is consistent with our knowledge of the protein. The very low scoring residues 341–380 (average pIDDT: 33.1242) map to the polyserine linker, which is known to be flexible and disordered [19]. Similar disorder is predicted for the N-terminal signal peptide residues 1–17 (average plDDT: 47.8064) and the segments upstream and downstream of the vWF domain, residue 1425–1437 and 1674–1684 (average plDDT: 44.5930 and 42.9336), respectively. Aligning the top ranking AlphaFold predictions demonstrates a consistent fold for the confident regions and some inconsistency of the low-confidence regions (Fig. S8C). The predicted disorder of residues 1674–1684 results in a variable positioning of the downstream C-terminal region between the predictions, suggesting flexibility of the domain position.

The final homology model and the AlphaFold prediction agree on the fold of the stable domain (Fig. S8D). AlphaFold produces 3D coordinates for every atom in the protein, so the prediction takes up more space, compared to the homology model where there are missing atoms, particularly downstream of the DUF1943 domain (Fig. S8D). However, the overall consistency in both of our predictions confirms that our structural prediction is strong.

Using PowerFit, ADP_EM, and Chimera to determine the domain assembly of full-length vitellogenin

The full-length models of Vg indicate the general fold of each domain. However, the domain assembly in the final homology model is speculative and derived from lamprey Vg and the deep learning method along with strong biases. To reduce these biases and provide some validation of the structural assembly, we performed rigid-body fitting of our model to a low-resolution EM map (Fig. S9, EMDB-22113, deposited) of *in vivo*-

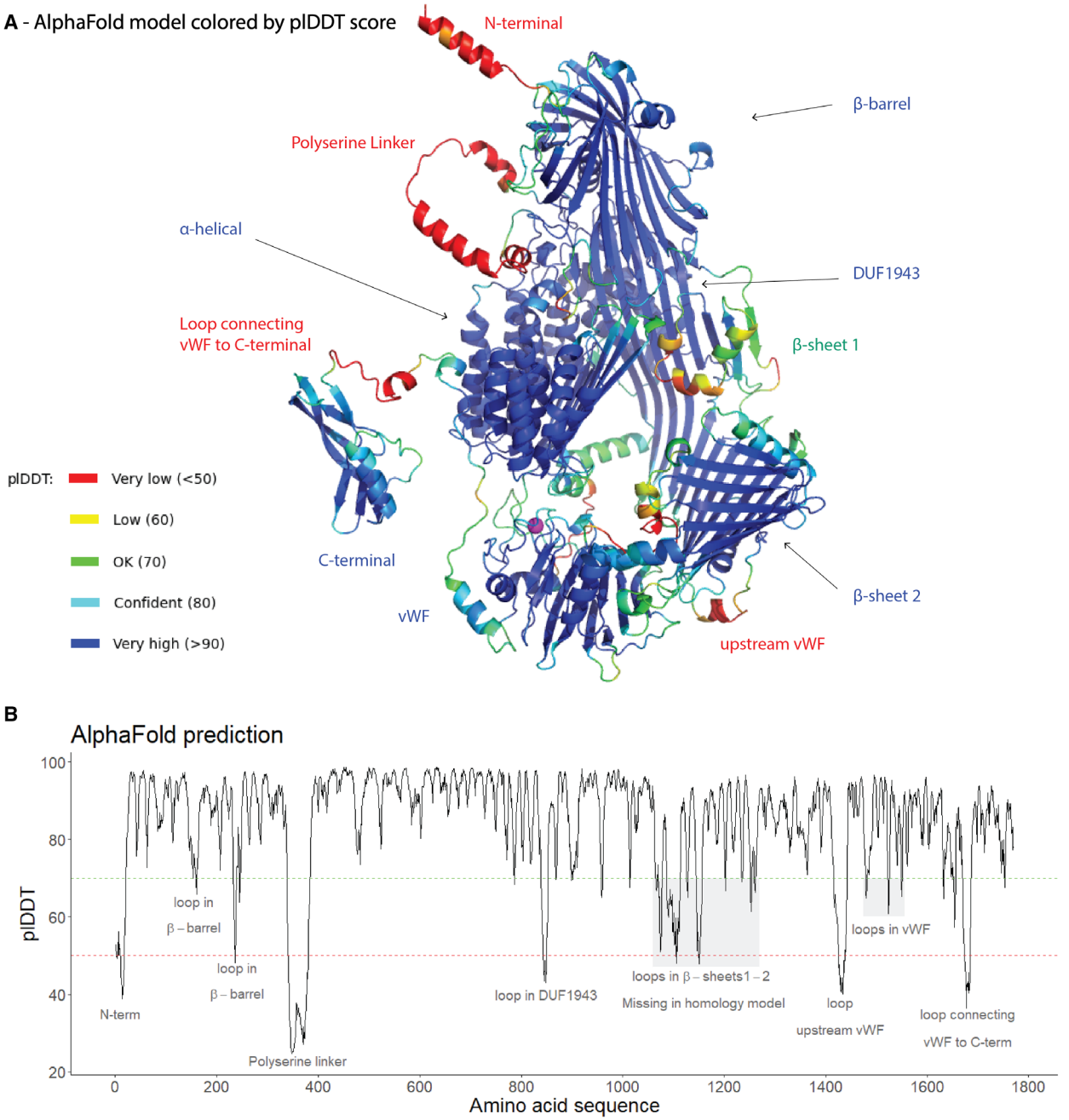


Fig. 2. AlphaFold prediction. (A) The top ranked AlphaFold model is shown as cartoon, colored by the plDDT scale. The high scoring domains (β-barrel, α-helical, DUF1943, β-sheet 2, vWF domain, and C-terminal domain) are labeled in blue, while the medium confident region (β-sheet 1) is labeled in green, and the low confident regions (N-terminal, polyserine linker, the segment upstream and downstream of vWF domain) are labeled in red. The Ca^{2+}-ion is shown as a magenta sphere. (B) The plDDT score is plotted per residue for the top ranked AlphaFold model. Each region that scores below 70 (green dotted line) is labeled. The very low plDDT (< 50) is indicated with a red dotted line.

obtained honey bee Vg. The EM map reveals a rough overview of the surface and two distinct cavities, hereafter named top, base, left and right, upper cavity (UC) and lower cavity (LC) in reference to this specific orientation (Fig. S10A). Fitting of the complete homology model placed chains D and E consistently outside the contour map, while chains A to C did not take up all the available space inside it (Fig. S10C).

This indicates incorrect domain assembly of chains D and E. Fitting of the RAPTORX structure gave similar results leaving chains C, D, and E outside the contour map, clearly demonstrating improper domain assembly (Fig. S10D). To avoid problems related to template-based assembly, we fitted the chains individually. Chains A and D occupy somewhat separate parts of the contour, but chain A overlaps with chain C and partly chain B and E (Fig. S10E,F). These individual domain fits support the assembly of chain A to C in the predicted model and further suggest improper assembly of chains D and E. Keeping chains A to C united but chains D and E separate resulted in two alternative orientations (Fig. S11B,C) leaving out chain E, which is not compatible with either alternative (Fig. S10F). The first 68 residues of chain E were built using a template-free method, while the last 58 residues were compiled from a multiple alignment of the last five templates (Table S7) ranging from 5% to 14% sequence identity. HHpred recognizes none of these templates. Faced with a speculative prediction and its incompatibility with the EM map, we removed the C-terminal domain from the domain assembly. The resulting fits from two independent rigid-body fitting methods (PowerFit [40,41] and ADP_EM [42]) was optimized using CHIMERA fit-in-map [43], producing correlation scores that could be compared directly (Figs S11A, S10B, and S12A). The highest scoring fit of chain A to C from ADP_EM is overlapping perfectly with the second-best fit from PowerFit (Fig. S11B1), while the highest scoring fit of the same chains from PowerFit is agreeing with the relative orientation of the domains. The best fit from Powerfit is not overlapping, however, with the second-best fit from ADP_EM (Fig. S11B2). The correlation score for the second ADP_EM fit is lower, and more atoms are outside the contour, compared to the other fits. Both alternatives are compatible with the ADP_EM and the PowerFit orientation of chain D (Fig. S11C). Secondary structure elements from the α-helical subdomain and DUF1943 are protruding outside the contour for both alternatives. For alternative 2, the DUF1943 and additionally the β-barrel subdomain are seemingly restricting access to both cavities (Fig. S11D).

To further investigate the two alternatives, we fitted previously generated homology models of the β-barrel and α-helical domains of honey bee Vg [3,25] and the X-ray structure of lamprey Vg (PDB ID: 1LSH [24]) to the EM map. The respective or homologous domains consistently fit in the two relative orientations and scored high values for both alternatives (Fig. S12). The β-barrel and α-helical domain supported alternative 1, while lamprey Vg favored alternative 2 according to the scores. The EM map is an *in vivo* representation of honey bee Vg, while the 1LSH structure is a distant homologue with 24% of the sequence missing in the crystal structure. The AlphaFold prediction with 100% sequence coverage serves as a far better representation of honey bee Vg. Fitting the top ranked AlphaFold prediction resulted in two different orientations by selecting the highest scoring fit from PowerFit and ADP_EM, respectively (Fig. 3A). The best fit from PowerFit has fewer atoms outside the contour and a higher correlation score, compared to the best fit from ADP_EM (Fig. 3B). The very low-confidence fold of the N-terminal signal peptide and the polyserine linker is protruding in both alternatives (Fig. 3C,D). In addition, smaller loops with a fold confidence ranging from low to intermediate are also protruding in both fits but these mismatches between model and contour map are more pronounced in the ADP_EM fit (Fig. 3D). The model cavities are restricted in the ADP_EM fit by the β-barrel and a long β-sheet which is the AlphaFold prediction of a more complete chain C, and these domains are confidently modeled. Both cavities in the PowerFit fit are also somewhat restricted by in-between domains segments, which have a lower confidence fold. Taken together, the orientation represented by PowerFit is the best fit of the AlphaFold prediction. This orientation also conforms to the best fits of individual domains: chain A to C (Fig. S11B2), chain D (Fig. S11C, PF1), β-barrel (Fig. S12B, ADP2), α-helical (Fig. S12C, PF2 and ADP1) and lamprey Vg (Fig. S12D, PF1 and ADP1). This further supports the PowerFit orientation of the AlphaFold prediction, but now with a more optimized fit. Using the full-length sequence representation results in a structure which fills more of the density space while keeping the percentage of protruding atoms low and the correlation score high. This suggests that the domain assembly in the AlphaFold prediction is an accurate representation of honey bee Vg.

The final model is presented in Fig. 4. The LC serves as the better-known lipid-binding site. It is easily accessible, while the hydrophobic core is buried in the EM map (Fig. 4A). The UC is partly built up by the β-barrel. The vWF domain is placed close to the LC bringing the Ca^{2+}-ion into close proximity to the cavity (Fig. 4B). This is supported by the results produced by the Volume, Area, Dihedral Angle Reporter (VADAR; Fig. S4D). The fractional accessible surface area report shows that the two short β-strands downstream of the Ca^{2+}-binding site are reported as exposed (r. 145–156 in plot 1, Fig. S4D).

FEBS Open Bio **12** (2022) 51–70 © 2021 The Authors. *FEBS Open Bio* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies

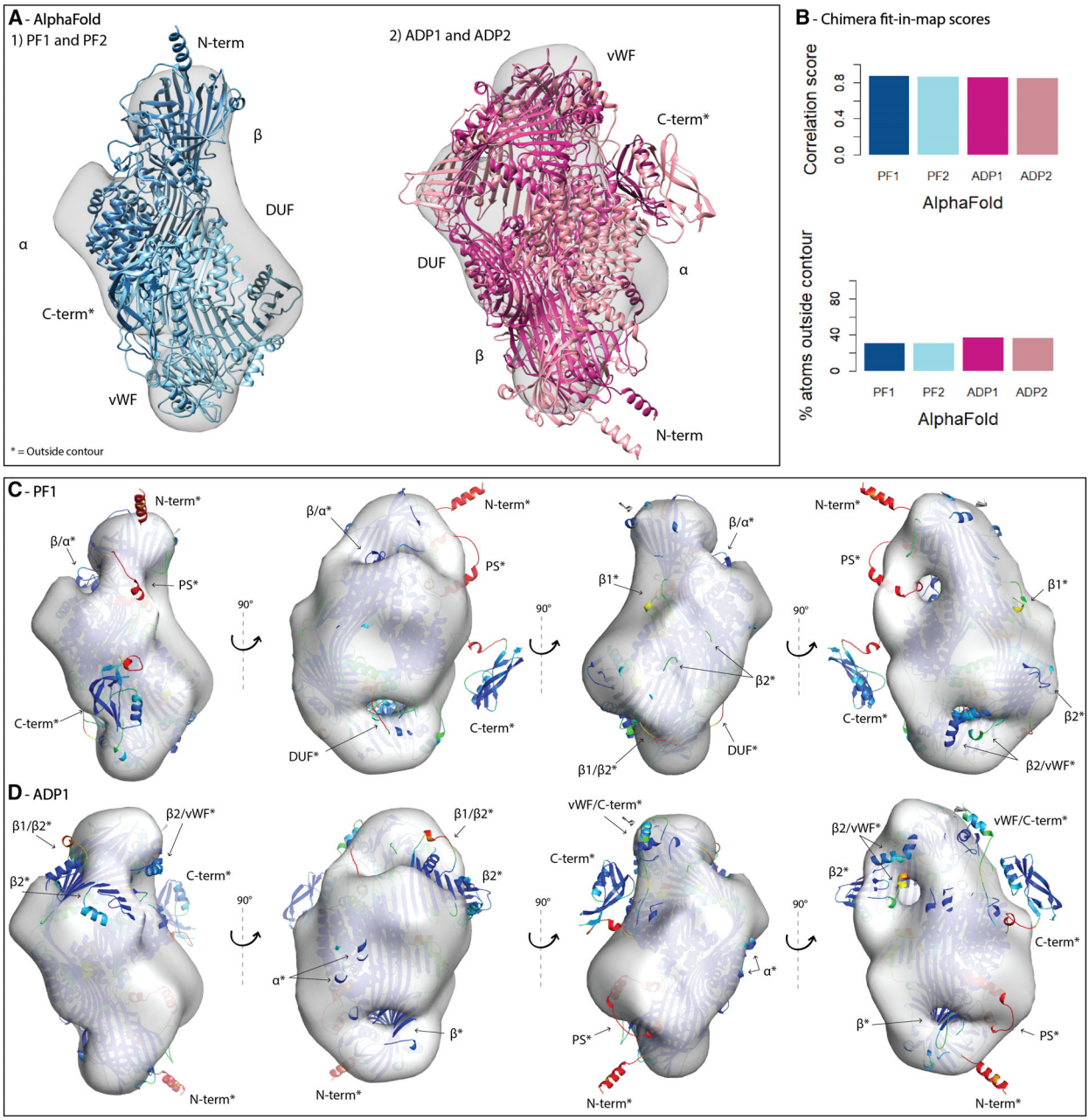


Fig. 3. Rigid-body fitting of AlphaFold. (A) The EM map are shown as a transparent surface, and the fits of AlphaFold from PowerFit (PF) and ADP_EM (ADP) are shown as cartoons and colored by method and scores (dark blue: PF1, light blue: PF2, dark pink: ADP1, light pink: ADP2). The N-terminal (N-term), β-barrel (β), α-helical (α), DUF1943 (DUF), vWF domain (vWF), and C-terminal (C-term) domains are labeled. (B) The correlation score and percent of atoms outside the contour calculated by CHIMERA were plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). (C) The EM map and the highest ranked PowerFit fit of AlphaFold is shown in at four different angels, colored by plDDT score. The label is marked with '*' if residues are outside the contour of the EM map and '/' between domain labels indicate that the pointed to segment is in-between domains. The polyserine linker and the two β-sheets downstream of the DUF1943 domain are labeled PS, β1, and β2, respectively. (D) The EM map and the highest ranked ADP_EM fit of AlphaFold. The same coloring and labeling are used as in panel C.

The fractional residue volume plot reports a potential cavity in the vicinity of the Ca^{2+}-binding site. In addition to the hydrophobic regions of Vg to be buried in the two cavities, the previously established hydrophilic and positively charged side of the α-helical domain [3] faces the surface in our model,

providing further support for a correct assembly. The polyserine region is also very exposed, favoring the reported dephosphorylation and cleavage events [19]. In the final model, we also mapped out residue positions of interest (Fig. 4C,D). The five functional polymorphisms are in association with a cavity (three in the lipid-binding site and two in the vWF domain). Anderson et al. [53] specified 12 polar interactions among nine residues on each monomer of lamprey Vg. Seven of these residues are conserved in honey bee Vg, and mapping these to the final model shows them to be accessible to solvent. Simulating the dimerization in PYMOL with the final model confirms dimerization to be a feasible oligomeric arrangement for honey bee Vg (Fig. 4E). However, re-fitting the Vg dimer in the EM map results in 33–39% of the atoms inside the contour (Tables S5 and S6). Taken together, this further supports the predicted assembly and demonstrates the EM map to be a representation of monomeric honey bee Vg.

Vitellogenin oligomerization state

While lamprey Vg forms a dimer with a modest 245 Å^2 hydrophobic interface in the crystal structure [24], mixed evidence exists for the oligomerization status of honey bee Vg. As described above, the negative-stain EM map with a resolution of 27 Å supports Vg to be monomeric since only one Vg molecule can be placed in the EM map, even at low contouring level. However, the sole known experimentally solved structure suggests that Vg can appear as a dimer [53], at least under some conditions. To further investigate this, we obtained purified Vg from honey bees and evaluated two different amounts using BN-PAGE (Fig. 5A). The lower molecular weight band (151 kDa) constitutes most of the material in the sample and is assumed to be monomeric Vg. The additional weaker band with higher molecular weight (345 kDa) is assumed to be a minor fraction of dimeric Vg. Contamination by other proteins in the sample seems

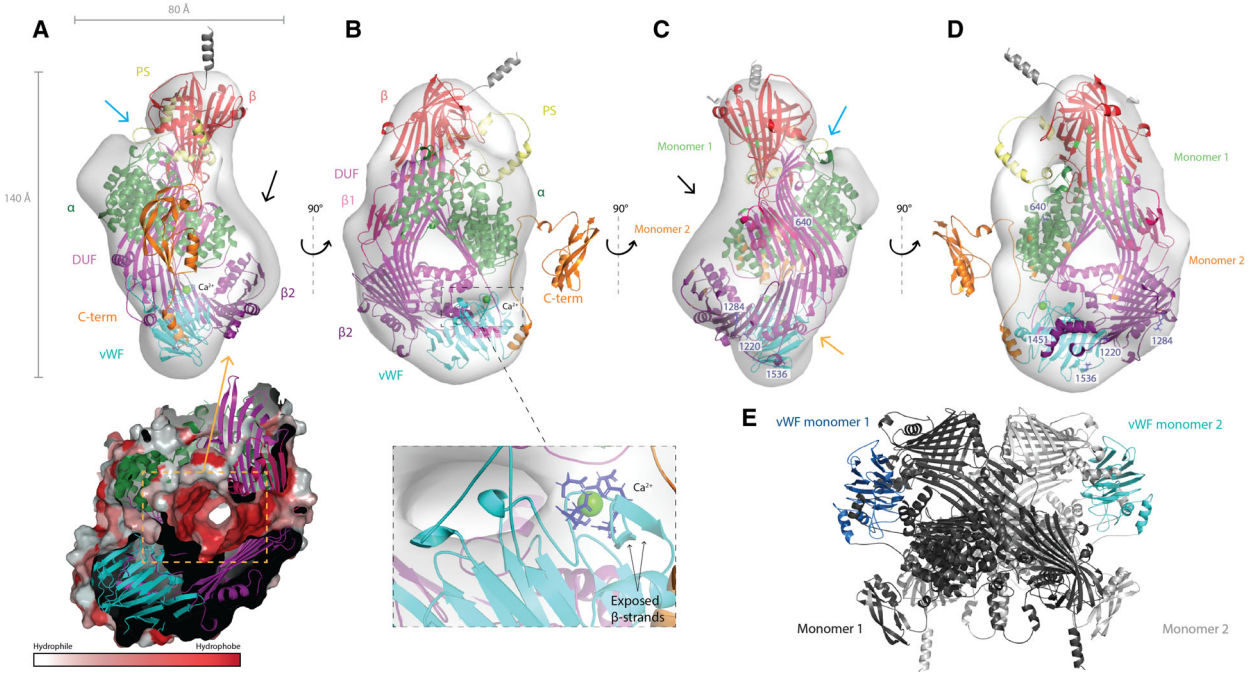


Fig. 4. Honey bee vitellogenin final assembly. The EM map is shown as a transparent surface from four different angles and have the AlphaFold model fitted inside. The polyserine linker (PS, yellow), β-barrel (β, red), α-helical (α, green), DUF1943 (DUF, magenta), β-sheet 1 (β1, hot pink), β-sheet 2 (β2, purple), vWF domain (vWF, cyan), and C-terminal (C-term, orange) domains are labeled, as well as the UC (blue arrow), LC (orange arrow), and empty density (black arrow). (A) The measurements of the EM map are shown along the x- and y-axis. The surface of the LC, colored by Eisenberg hydrophobicity scale [44], is shown inside the orange dashed box surrounded by the domains building up the cavity. (B) Here, we zoom in on the Ca^{2+}-binding sites, and show the two exposed β-strands (black arrows) and their proximity to the LC. (C, D) The five residue positions (640, 1220, 1284, 1451, and 1536) identified as candidates of functional polymorphisms are colored blue and labeled. The conserved residues in honey bee Vg that make polar contacts during dimerization are colored green (monomer 1) and orange (monomer 2). (E) The simulated Vg dimer is shown with monomer 1 (dark gray) and 2 (light gray). The vWF domain is colored in each monomer (monomer 1, dark blue and monomer 2, cyan).

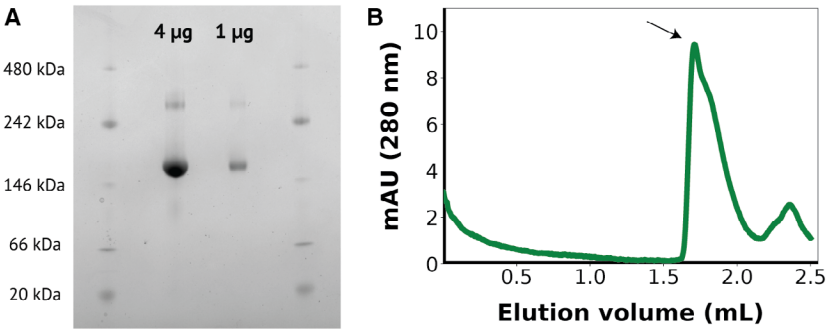


Fig. 5. *In vitro* oligomerization state analysis of vitellogenin. (A) BN-PAGE gel results. Both the bands corresponding to the monomer and the dimer can be observed for Vg loaded in different amounts. (B) SEC elution profile for purified Vg. The peak containing Vg is labeled with an arrow corresponding to an elution volume consistent with monomeric Vg.

unlikely since only one band for Vg can be observed from the sample in a denaturing PAGE (not shown). Next, we performed SEC (Fig. 5B), and the content of the concentrated fractions was analyzed with BN-PAGE (not shown). The main peak obtained corresponded to monomeric Vg, and its apparent molecular weight was estimated to be 178 kDa based on the elution volume. No peak corresponding to the dimeric form was obtained, although when the fraction from the main peak was concentrated, it showed on a native blue PAGE both as a monomer and a dimer in similar proportion to that observed in Fig. 5A. Together, these results suggest that Vg can dimerize at higher protein concentrations *in vitro*.

Discussion

With this study, we aimed to gain more insight into the structure of honey bee Vg and to attempt a full-length model of the protein. Our results reveal structural features that have not yet been described for Vg in invertebrates.

First, we presented a detailed structural prediction of the vWF domain. Through homology modeling, we identified a potential class II Ca^{2+}-binding site, which appears to be highly conserved across Vg and vWF-containing species. The Ca^{2+}-ion coordinates with 4 Asp and 2 Asn residues, through their OD1 or OD2 atoms, respectively, except for D1604, which coordinates through its main chain carbonyl O-atom. In the human WD3 domain, the residue corresponding to D1604 is I1002 (Fig. S2). The side chain of isoleucine is unable to interact meaningfully with calcium [54]. We speculate that the introduction of a sixth calcium-coordinating residue, aspartate, creates an additional bond to the Ca^{2+}-ion, increasing the interaction and strengthening the coordination. Identifying a total of six coordinating residues and a loop structure in the binding site enabled us to categorize this as a class II site [48].

We were able to present a full-length structure prediction of an invertebrate Vg. However, our concern about the remaining domains is that the use of distant homologues with low sequence identity can create predictions influenced by the template used. Studies show that general protein folds are well conserved across great phylogenetic distances despite low conservation of the amino acid sequence [55]. Focusing mainly on the general fold and creating several models by using different query sequence lengths, we increased our confidence in the prediction for each domain. The striking similarity between the AlphaFold prediction and the predicted homology model chains validates our modeling results. In addition, AlphaFold provides a confident domain fold of the C-terminal region, and predicts folds for loop regions missing in the homology model, enabling us to present a 100% complete structure representation of honey bee Vg, with considerable confidence within each domain. Using PowerFit, ADP_EM and CHIMERA, we were able to present a domain assembly of the full-length structure prediction. The negative-stain EM map has a low resolution (27 Å), which increases the margin of error. To limit the number of possible orientations, we fitted the homology models according to size, beginning with the largest. We also fitted the previously predicted domains, the crystal structure of lamprey Vg and the AlphaFold prediction to validate our modeled fold and its placement in the EM map. We evaluated each fit based on the scoring, protruding atoms and overlapping fits of separate domains. We concluded that the AlphaFold PowerFit orientation, with the DUF1943 domain, the two downstream β-sheets and vWF domain oriented around the LC and the β-barrel

and α-helical subdomain toward the UC (Fig. 3A1), was the most probable representation for honey bee Vg. The energetics for the full-length model and the separate domains (e.g., whether polar surfaces or hydrophobic surfaces were exposed to the solvent) are logical, as demonstrated for the lipid binding site (Fig. 4A). The final model does not occupy all available density while the C-terminal region is outside the contour, which represents about 4.6% of the atoms. The position of this domain is not clear as the Alphafold results indicate a flexibility in the connecting loop. The unassigned density in the low-resolution EM map above the UC could potentially be where the C-terminal region is positioned (Fig. 4A,C). Honey bee Vg is also found to be phosphorylated and glycosylated, [25] which is not represented in the protein structure and could explain the excess of density.

Both cavities identified in the EM map are compatible with the assembly, and the LC is identified as the lipid-binding site, which recognizes lipids, possible fragments of gram-negative and gram-positive bacteria [24]. The UC, built up partly by the β-barrel subdomain, has not been described earlier, and whether the UC has similar recognition potential, to the LC is not known. The *in vitro* mutagenesis experiments performed for the human vWF protein [56] illustrate the importance of the Ca^{2+}-binding site for recognition of factor VIII in a blood-clotting cascade. A study from 2013 shows fbVg to be membrane associated and speculates the receptor binding site to be in the 150 kDa subunit and not in the β-barrel domain as previously believed [3]. Insect Vg receptors belong to a subfamily of the low-density lipoprotein receptor family, and calcium interaction has been shown to be essential for ligand association [57,58]. Our findings support these results and suggest the vWF domain as the potential Vg receptor binding site. Additionally, the vWF domain has been implicated in having adhesive and lubricant properties [59,60] as seen for vWF and mucin proteins in humans. The structure of the WD3 domain, used as template here, was recently functionally compared to the MUC2 in humans. Since the two proteins shows high structural similarity, Javitt et al. [61] suggest that WD3 has a similar polymerization function and is essential for macromolecular assemblies in the epithelial mucosa and vasculature. Our study shows that the interchain disulfide bonds, essential for oligomerization in the human vWF [26,56], are not conserved in honey bees. In addition, residues in the β-barrel and α-helical domain are interacting in the Vg dimer, and not the vWF domains (Fig. 4E), thereby ruling out this kind of polymerization activity for the vWF domain in honey bees. However, the Ca^{2+}-binding site, the intrachain disulfide bonds and the β-sandwich are highly conserved, suggesting a similar function in mucosal immunity, as seen for mucins and vWF proteins in humans.

Insects, which have an open circulation system, have developed an efficient coagulation mechanism that is an essential part of their innate immune system [62]. When exposed to invading microbes, a clotting cascade is initiated, trapping and eventually killing the invaders [63]. The hemolymph clot was recently characterized in a Brazilian whiteknee tarantula, showing the main content to be proteins encompassing vWF-like domains. Sanggaard et al. [64] results also indicate that the clot functional and structural overlaps with such clots observed in insects. We propose that honey bee Vg can initiate or aid in this clotting mechanism, interacting through the vWF domain, and protect honey bees from pathogens and mechanical damage, like in zebrafish Vg [4]. Our identification of three residue positions exhibiting high genetic differentiation in the LC could be a result of adaption to binding substrates present in specific environments. Our results work well with this theory since we also identified the last two functional polymorphisms close to the LC. This suggests that the vWF domain recognizes environmental factors such as pathogens. Specifically, site 1451 (Fig. 4C,D) is in a small hydrophobic pocket close to the Ca^{2+}-binding site. Our MSA shows conservation of hydrophobicity in this position, which is often seen for binding sites. Based on our collected data, this speculation cannot be confirmed, but could form the basis of new experimental work in which this is explored.

Our results suggest that honey bee Vg is predominantly monomeric *in vitro*. First, only one copy of the Vg model could fit into the low-resolution EM map. Second, SEC analysis showed only one peak, and this corresponded to monomeric Vg. Third, native gel results also showed a higher tendency toward a monomeric state determined by the much weaker 345 kDa band (presumably a dimer). On the contrary, we demonstrated that the seven residues of each monomer that are creating polar contacts during dimerization in lamprey Vg are conserved in honey bee Vg, making it plausible that Vg dimers can form in honey bees in certain cases. We note that no reducing agent was present in the loading buffer or gel, making it possible that dimers are stabilized by disulfide bonds. Additionally, we cannot rule out that high salt concentration in the SEC prevented the formation of the Vg dimer. Taken together, it is difficult to determine whether dimerization occurs *in vivo* or is an artifact of the *in vitro* conditions, as dimerization occurs frequently in a high concentration sample containing just one

type of protein [65]. We speculate that dimerization can be dose-dependent and thus become more prevalent at elevated Vg concentration. The concentration of Vg in honey bee hemolymph has been reported as high as 100 $\mu g \cdot \mu L^{-1}$, illustrating that the protein is highly soluble [66]. More efforts are needed to conclude the oligomeric state of Vg in honey bees and to evaluate earlier evidence describing honey bee Vg to be monomeric [57,67].

To summarize, our study presents new evidence of the full-length protein and domain assembly for honey bee Vg. We are thus able to identify properties and describe the structural landscape of the large and versatile protein. Our results verify a second cavity of honey bee Vg in addition to the well described lipid-binding cavity and describe the structural units potentially forming this cavity. As a result, we are able to suggest the possibility that the vWF domain contributes to the immune system of honey bees, which is currently of global concern due to declining pollinator numbers. Efforts are being made to generate a higher resolution and up-to-date EM map, which could be used to preform molecular dynamic flexible fitting and enable studies of Vg protein–protein interactions and ligand binding. Our findings encourage future initiatives in investigating this domain together with the full-length protein to unravel some of the questions asked here.

Acknowledgements

We thank Eivind Fjeldstad for his valuable guidance for running AlphaFold. The authors acknowledge The Research Council of Norway grant number 262137 for funding toward running costs and positions. MM-C is supported by an H2020 MSCA International Training Network, ESC by an H2020 MSCA Individual Fellowship, HL by NCMM core funding. The FP7 WeNMR (project# 261572), H2020 West-Life (project# 675858), and the EOSC-hub (project# 777536) European e-Infrastructure projects are acknowledged for the use of their web portals, which make use of the EGI infrastructure with the dedicated support of CESNET-MetaCloud, INFN-PADOVA, NCG-INGRID-PT, TW-NCHC, SURFsara, and NIKHEF, and the additional support of the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands, Poland, Portugal, Spain, UK, Taiwan, and the US Open Science Grid. Molecular graphics and analyses performed with UCSF CHIMERA, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. The authors acknowledge BioCat (RCN grant number 249023) for travel grants and conferences support.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

VL executed homology modeling, structure predictions, rigid-body fitting, and purification of honey bee Vg and ØH and GVA supervised the research. EH-G done the negative staining and generation of the EM map. MM-C and ESC performed native gel and SEC and HL supervised the research. VL wrote the manuscript with assistance from ØH and GVA. All authors contributed to the manuscript.

Data accessibility

The data that support the findings of this study are available in the supplementary material of this article (Tables S1–S7, Figs S1–S12, and the EM map validation report [Appendix S1]). The structural data from homology modeling of the vWF domain are openly available at ModelArchive https://modelarchive.org/doi/10.5452/ma-sfueo (access code: okHs98Pcl2), and the structural data from AlphaFold are available in the supplementary material of this article.

References

1 Hayward A, Takahashi T, Bendena WG, Tobe SS, Hui JH. Comparative genomic and phylogenetic analysis of vitellogenin and other large lipid transfer proteins in metazoans. *FEBS Lett*. 2010;**584**(6):1273–8.
2 Corona M, Velarde RA, Remolina S, Moran-Lauter A, Wang Y, Hughes KA, et al. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc Natl Acad Sci USA*. 2007;**104**(17):7128–33.
3 Havukainen H, Munch D, Baumann A, Zhong S, Halskau O, Krogsgaard M, et al. Vitellogenin recognizes cell damage through membrane binding and shields living cells from reactive oxygen species. *J Biol Chem*. 2013;**288**(39):28369–81.
4 Zhang S, Dong Y, Cui P. Vitellogenin is an immunocompetent molecule for mother and offspring in fish. *Fish Shellfish Immunol*. 2015;**46**(2):710–5.
5 Sun C, Hu L, Liu S, Gao Z, Zhang S. Functional analysis of domain of unknown function (DUF) 1943, DUF1944 and von Willebrand factor type D domain (VWD) in vitellogenin2 in zebrafish. *Dev Comp Immunol*. 2013;**41**(4):469–76.

6 Du X, Wang X, Wang S, Zhou Y, Zhang Y, Zhang S. Functional characterization of vitellogenin_n domain, domain of unknown function 1943, and von Willebrand factor type D domain in vitellogenin of the non-bilaterian coral *Euphyllia ancora*: implications for emergence of immune activity of vitellogenin in basal metazoan. *Dev Comp Immunol*. 2017;**67**:485–94.

7 Salmela H, Amdam GV, Freitak D. Transfer of immunity from mother to offspring is mediated via egg-yolk protein vitellogenin. *PLoS Pathog*. 2015;**11**(7): e1005015.

8 Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc Natl Acad Sci USA*. 2006;**103**(4):962–7.

9 Nakamura A, Yasuda K, Adachi H, Sakurai Y, Ishii N, Goto S. Vitellogenin-6 is a major carbonylated protein in aged nematode, *Caenorhabditis elegans*. *Biochem Biophys Res Comm*. 1999;**264**(2):580–3.

10 Ando S, Yanagida K. Susceptibility to oxidation of copper-induced plasma lipoproteins from Japanese eel: protective effect of vitellogenin on the oxidation of very low density lipoprotein. *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol*. 1999;**123**(1):1–7.

11 Amdam GV, Norberg K, Hagen A, Omholt SW. Social exploitation of vitellogenin. *Proc Natl Acad Sci USA*. 2003;**100**(4):1799–802.

12 Havukainen H, Halskau O, Amdam GV. Social pleiotropy and the molecular evolution of honey bee vitellogenin. *Mol Ecol*. 2011;**20**(24):5111–3.

13 Hernandez Lopez J, Schuehly W, Crailsheim K, Riessberger-Galle U. Trans-generational immune priming in honeybees. *Proc Biol Sci*. 2014;**281** (1785):20140454.

14 Sadd BM, Kleinlogel Y, Schmid-Hempel R, Schmid-Hempel P. Trans-generational immune priming in a social insect. *Biol Lett*. 2005;**1**(4):386–8.

15 Nelson CM, Ihle KE, Fondrk MK, Page RE, Amdam GV. The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol*. 2007;**5**(3):e62.

16 Kohlmeier P, Feldmeyer B, Foitzik S. Vitellogenin-like a-associated shifts in social cue responsiveness regulate behavioral task specialization in an ant. *PLoS Biol*. 2018;**16**(6):e2005747.

17 Suren-Castillo S, Abrisqueta M, Maestro JL. Foxo inhibits juvenile hormone biosynthesis and vitellogenin production in the German cockroach. *Insect Biochem Mol Biol*. 2012;**42**(7):491–8.

18 Dittmer J, Alafndi A, Gabrieli P. Fat body-specific vitellogenin expression regulates host-seeking behaviour in the mosquito *Aedes albopictus*. *PLoS Biol*. 2019;**17** (5):e3000238.

19 Havukainen H, Underhaug J, Wolschin F, Amdam G, Halskau O. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. *J Exp Biol*. 2012;**215**(Pt 11):1837–46.

20 Kent CF, Issa A, Bunting AC, Zayed A. Adaptive evolution of a key gene affecting queen and worker traits in the honey bee, *Apis mellifera*. *Mol Ecol*. 2011;**20**(24):5226–35.

21 Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. Genetic integrity of the dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *J Apic Res*. 2014;**53**(2):269–78.

22 Munoz I, Henriques D, Jara L, Johnston JS, Chavez-Galarza J, De La Rua P, et al. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (*Apis mellifera mellifera*). *Mol Ecol Resour*. 2017;**17**(4):783–95.

23 Henriques D, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, et al. High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and cost-effective SNP-based tool. *Sci Rep*. 2018;**8**(1):8552.

24 Thompson JR, Banaszak LJ. Lipid-protein interactions in lipovitellin. *Biochemistry*. 2002;**41**(30):9398–409.

25 Havukainen H, Halskau O, Skjaerven L, Smedal B, Amdam GV. Deconstructing honeybee vitellogenin: novel 40 kDa fragment assigned to its n terminus. *J Exp Biol*. 2011;**214**(Pt 4):582–92.

26 Dong X, Leksa NC, Chhabra ES, Arndt JW, Lu Q, Knockenhauer KE, et al. The von Willebrand factor D'D3 assembly and structural principles for factor VIII binding and concatemer biogenesis. *Blood*. 2019;**133** (14):1523–33.

27 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;**596**(7873):583–9.

28 Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol*. 2018;**430**(15):2237–43.

29 Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;**16**(6):276–7.

30 Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;**48** (3):443–53.

31 Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*. 1997;**18**(15):2714–23.

32 Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-

MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis*. 2009;**30**(S1):S162–73.
33 van Gunsteren WF. *Biomolecular simulations: the GROMOS96 manual and user guide*. Zürich: VDF Hochschulverlag AG an der ETH Zürich; 1996. p. 1–1042.
34 Laskowski RA, MacArthur MW, Moss DS, Thornton JM. Procheck: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*. 1993;**26**(2):283–91.
35 Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 2011;**27**(3):343–50.
36 Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, et al. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*. 2003;**31**(13):3316–9.
37 Schrodinger L. The pymol molecular graphics system, version 1.8; 2015.
38 Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. *Proteins*. 1995;**23**(3):318–26.
39 Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *bioRxiv*. 2020. 2020.2010.2012.336859.
40 van Zundert GC, Trellet M, Schaarschmidt J, Kurkcuoglu Z, David M, Verlato M, et al. The DisVis and PowerFit web servers: explorative and integrative modeling of biomolecular complexes. *J Mol Biol*. 2017;**429**(3):399–407.
41 Zundert GCP, Bonvin AMJJ. Fast and sensitive rigid-body fitting into cryo-em density maps with powerfit. *AIMS Biophys*. 2015;**2**(2):73–87.
42 Garzón JI, Kovacs J, Abagyan R, Chacón P. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*. 2007;**23**(4):427–33.
43 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;**25**(13):1605–12.
44 Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*. 1984;**179**(1):125–42.
45 Aase ALTO, Amdam GV, Hagen A, Omholt SW. A new method for rearing genetically manipulated honey bee workers. *Apidologie*. 2005;**36**(3):293–9.
46 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;**215**(3):403–10.
47 Venclovas C. Methods for sequence-structure alignment. *Methods Mol Biol*. 2012;**857**:55–82.
48 Pidcock E, Moore GR. Structural characteristics of protein binding sites for calcium and lanthanide ions. *J Biol Inorg Chem*. 2001;**6**(5–6):479–89.
49 Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 1963;**7**:95–9.
50 Bordoli L, Schwede T. Automated protein structure modeling with SWISS-MODEL workspace and the protein model portal. *Methods Mol Biol*. 2012;**857**:107–36.
51 Dalton JAR, Jackson RM. An evaluation of automated homology modelling methods at low target–template sequence similarity. *Bioinformatics*. 2007;**23**(15):1901–8.
52 Mariani V, Biasini M, Barbato A, Schwede T. LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013;**29**(21):2722–8.
53 Anderson TA, Levitt DG, Banaszak LJ. The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein. *Structure*. 1998;**6**(7):895–909.
54 Lu C-H, Lin Y-F, Lin J-J, Yu C-S. Prediction of metal ion-binding sites in proteins using the fragment transformation method. *PLoS One*. 2012;**7**(6):e39252.
55 Friedberg I, Margalit H. Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. *Protein Sci*. 2002;**11**(2):350–60.
56 Springer TA. von Willebrand factor, Jedi knight of the bloodstream. *Blood*. 2014;**124**(9):1412–25.
57 Sappington TW, Raikhel AS. Molecular characteristics of insect vitellogenins and vitellogenin receptors. *Insect Biochem Mol Biol*. 1998;**28**(5):277–300.
58 Atkins AR, Brereton IM, Kroon PA, Lee HT, Smith R. Calcium is essential for the structural integrity of the cysteine-rich, ligand-binding repeat of the low-density lipoprotein receptor. *Biochemistry*. 1998;**37**(6):1662–70.
59 Faiz ZM, Mardhiyyah MP, Mohamad A, Hidir A, Nurul-Hidayah A, Wong L, et al. Identification and relative abundances of mRNA for a gene encoding the vWD domain and three Kazal-type domains in the ovary of giant freshwater prawns, *Macrobrachium rosenbergii*. *Anim Reprod Sci*. 2019;**209**:106143.
60 Finn RN. Vertebrate yolk complexes and the functional implications of phosvitins and other subdomains in vitellogenins. *Biol Reprod*. 2007;**76**(6):926–35.
61 Javitt G, Khmelnitsky L, Albert L, Bigman LS, Elad N, Morgenstern D, et al. Assembly mechanism of mucin and von Willebrand factor polymers. *Cell*. 2020;**183**(3):717–29.e716.
62 Loof TG, Schmidt O, Herwald H, Theopold U. Coagulation systems of invertebrates and vertebrates and their roles in innate immunity: the same side of two coins? *J Innate Immun*. 2011;**3**(1):34–40.

63 Eleftherianos I, Revenis C. Role and importance of phenoloxidase in insect hemostasis. *J Innate Immun*. 2011;**3**(1):28–33.
64 Sanggaard KW, Dyrlund TF, Bechsgaard JS, Scavenius C, Wang T, Bilde T, et al. The spider hemolymph clot proteome reveals high concentrations of hemocyanin and von Willebrand factor-like proteins. *Biochim Biophys Acta*. 2016;**1864**(2):233–41.
65 Wang W, Xu W-X, Levy Y, Trizac E, Wolynes PG. Confinement effects on the kinetics and thermodynamics of protein dimerization. *Proc Natl Acad Sci USA*. 2009;**106**(14):5517–22.
66 Amdam GV, Hartfelder K, Norberg K, Hagen A, Omholt SW. Altered physiology in worker honey bees (Hymenoptera: Apidae) infested with the mite Varroa destructor (Acari: Varroidae): a factor in colony loss during overwintering? *J Econ Entomol*. 2004;**97**(3):741–7.
67 Tufail M, Takeda M. Molecular characteristics of insect vitellogenins. *J Insect Physiol*. 2008;**54**(12):1447–58.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Domain architecture of honey bee and lamprey vitellogenin. The N-term (green), DUF1943 (pink) and vWF (blue) domains are conserved in both species, as well as the two structural subdomains, β-barrel (red arrow) and α-helical domain (dark green curved line). A) Honey bee Vg contains a proteolytic cleavage site, polyserine region (yellow S) linking the two subdomains. The five residue-positions (640, 1220, 1284, 1451 and 1536) identified to be candidates of functional polymorphisms are marked (brown stars). B) Lamprey Vg contains an addition domain, DUF1943 (purple). The yolk protein organization of IuVg is shown as gray boxes; lipovitellin heavy chain (LvH), Phosvitin (Pv), lipovitellin light chain (LvL), β-Component (β-C) and C-terminal coding region (CT). The dotted lines indicate that these regions (Pv, β-C and CT) are missing from the crystallographic structure (PDB ID: 1LSH).

Fig. S2. Multiple sequence and structural alignment. The coloring for the conserved residues/regions, gaps and secondary structure annotations are explained in the green box. The conserved Ca2+-binding region are colored in two shades of pink, dark pink is more conserved compared to the lighter pink. A) Extraction of the MSA. The original residue numbering for honey bee Vg is included on top. B) The final structural alignment with the original residue numbering included above each sequence. The annotations are retrieved from the template (PDB ID: 6N29). Both figures are created in Geneious Prime (v. 2019.0.3) and Adobe illustrator (v. 24.0.02).

Fig. S3. ProCheck summary, local quality estimate and Ramachandran plots. A) The ProCheck quality evaluations summarized and categorized by calculation results. The ideal residue values and standard deviation for any given model are derived from Morris et al. 1992.[1] The max deviation, in residues properties, is calculated from the mean value of the residue-by-residue listing values (Fig. S4E) of the full-length structure. The number of bad contacts is defined as the non-bonded atoms at a distance of <= 2.6 Å. The bond length and angles are calculated in similar manner as the max deviation, but the ideal values are based on Engh and Huber 1991.[2] The Morris et al. (1992) class summarizes the three above stereochemical parameters by assigning a number between 1 (best) to 4 (worst), indicating the overall quality of the model. B) Local QMEAN results are presented. The first plot is analysis of the template (green), while the second is analysis of the target modeled interactively (cyan) and automatically (red). The Ca^{2+}-binding region (magenta Ca), the Cys residues forming the intra-chain disulfide bridges (orange, C) are in the higher quality region, while Ω5-7 (black) are in the lower quality region. The local score is calculated for each residue in the model and the average local score for the template is 0.93 ± 0.07, while the target average score is 0.40 ± 0.07 (cyan) and 0.44 ± 0.06 (red). C) The Ramachandran plot produced by ProCheck. The plot on the left is the template (PDB ID: 6N29), while the target (honey bee vWF domain) is on the right. Below each plot, the statistic is presented.

Fig. S4. Global quality estimate, VADAR plots and ProCheck residue listing. A-C) The plots of the global QMEAN have the QMEAN4 scores for a set PDB structures plotted (gray dots) with the QMEAN4 score along the x-axis and the number of residues in the structures as long the y-axis. The global scores value QMEAN4 range from 0 to 1, where 1 is good. A) Analysis of the template (red star) and the QMEAN4 value is written on the plot. B) Analysis of the interactively homology modeled (red star) structure and C) The automatically homology modeled (red star) structure from MODELLER. D) Four different analyses were performed by VADAR, presented in one plot each, with the template (gray) compared to the target (green). Plot 1: a low fractional ASA score indicates a buried residue, while a score above 0.5 (dotted black line) indicates an exposed residue. A score above 1.0 (red line) indicates a problem in the structure. Plot 2: When a protein structure is efficiently packed the score

should be around 1.0 ± 0.1. A score above 1.2 (blue line) or below 0.8 (red line) could indicate a poor refinement or identify cavities. Plot 3: Each residue is assigned a score between 0-3 (high is good quality) for three different measurements (torsion angle, omega angle and fractional volume). The total quality score for each residue can be from 0-9 and the threshold for a good quality is set to 6 (red line). Plot 4: Calculates the 3D quality of each residue based on its environment and gives a score between 0-9 (high is good quality), and the threshold for a good quality is set to 4 (red line). E) The Residue-by-Residue listing for ProCheck lists all residues in a structure and present all calculations for each. A short example is shown here for the first six residues in the target structure. Each value is compared to the ideal values which is noted on top. The deviating values are marked with * (one standard deviation) and + (half a standard deviation) sign. For example, the omega dihedral angle of residue S1443 is 16.9 standard deviation away from the ideal value, which is a result from the loop building of Ω1.

Fig. S5. Comparison of vWF homology models. A) The sort region around the Ca^{2+}-binding site (Ca^{2+}-ion, green) is shown from the interactively modeled (cyan) structure and the automatically modeled (gray) structure. The Cys-residues (C1444, C1466, C1598 and C1634) and Ca^{2+}-binding residues are shown as yellow/cyan (interactively) and orange/magenta (automatically) sticks. The missing C1634 and β-strands in the automatically modeled structure are shown (gray arrows). B) All the Ca^{2+}-binding residues are in the same orientation in both models (light blue: interactively and light pink: automatically), except N1607. The interactions to the Ca^{2+}-ion is shown as yellow dotted lines and measured (Å) for N1607.

Fig. S6. Comparison of homology models from MODELLER and RaptorX. A) The N-terminal domain: Model 1 (green) aligned with Model 4 (red), 5 (yellow) and 6 (forest green). B) The DUF1943 domain: Model 1 (magenta) aligned with Model 8 (cyan), Model 7 (orange) and Model 9 (blue). The identified curve in the longer β-sheet in Model 1, 8 and 9 and the missing curve in Model 7 is marked with arrows. C) The DUF1943 domain Model 1 (magenta), the downstream region residue 1060 to 1140 of Model 9 (hot pink) and the loop region (gray). D) The undetermined domain: Model 2 (purple) aligned with Model 9 (blue), with the long loop region (gray). E) The interactively homology model of vWF domain (cyan) with the C-terminal region from Model 9 (orange).

Fig. S7. RaptorX structural prediction of full-length honey bee vitellogenin. A) The β-barrel subdomain (red), the polyserine linker (yellow), the α-helical subdomain (forest green), the DUF1943 domain (magenta), elongation of the DUF1943 domain (hot pink arrow), the undetermined structural region (purple), the vWF domain (cyan) and the C-terminal region (orange) are generated as one full-length model. The two loop regions (gray arrows) are also predicted. B) Domain 1 to 6 from Table S7 are colored red, cyan, purple, blue, green and orange, respectively, and if templates was used, the PDB ID is written in parenthesis.

Fig. S8. AlphaFold output. A) The number of sequence hits in the MSA produced by AlphaFold, is plotted per residue. The average number of hits per residue (gray dotted line), and the threshold at 100 sequence per residue (red dotted line) is marked. B) The plDDT score for the five outputted models by AlphaFold is plotted per residue, and the average plDDT score per model is listed to the right, which produces the rank from 0 (best) to 4 (worst). C) The ranked models are aligned, colored by the same coloring scheme in panel B, and the consistently folded domains (β-barrel (β), α-helical (α), DUF1943 (DUF), β-sheet 1 (β1), β-sheet 2 (β2) and vWF domain (vWF)) are labeled in bold letters, while the more variable domains (N-terminal, polyserine linker (PS) and C-terminal) are labeled in grey letters. D) The final homology model domains (β-barrel (red), polyserine linker (yellow), α-helical (green), DUF1943 (magenta), β-sheet 1 (hotpink), β-sheet 2 (purple), vWF (cyan, Ca^{2+}-ion shown as green sphere) and C-terminal domain (orange) is aligned to their respective domains in the top ranked AlphaFold prediction (grey). The grey brackets to the lower right indicate the region where AlphaFold have predicted a fold for the main missing atoms in the homology model.

Fig. S9. EM map validation. A) Map visualization to allow visual inspection of the internal detail of the map and identification of artifacts. The primary map, central slices of the map and largest variance of the map is shown in three orthogonal directions. The 3D surface view of the primary map at recommended contour level 0.07. B) Statistical analysis of the map. In the first graph the map-value distributions is plotted in 128 intervals along the x-axis, and the y-axis is logarithmic. The spike around 0 indicate that the volume has been masked. The second graph shows how the enclosed volume varies with the contour level. The volume at the recommended contour (red line) is 289 nm3; this corresponds to an approximate mass of 261 kDa. C) The provided Fourier-Shell Correlation (blue) is plotted together with the reported resolution, (black line, *Reported resolution corresponds to spatial frequency of 0.037 $Å^{-1}$). A curve is displayed for the half-

bit criterion (dashed red), in addition to lines showing the 0.143 gold standard cut-off (dashed orange line) and 0.5 cut-off (green dotted line). All the graphs are assembled from the EmDataBank map validation report (copy included).

Fig. S10. Rigid-body fitting for honey bee vitellogenin homology models. A) The EM map is shown as a gray surface. The distinct cavity creases are marked with stars and arrows, upper cavity (blue) and lower cavity (yellow). The four curves in the surface are labeled (top, base, left and right). B) The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). C-E) The fits from the full-length homology model, RaptorX and chain A is presented inside the EM map, with the same coloring scheme as in panel B. The β-barrel (β), α-helical (α), DUF1943 (DUF), vWF and C-terminal (C-t) domains are labeled. If the domain is outside of the contour it is noted by a "*"-mark. F) The fits of chain B to E separately with the same coloring scheme as in panel B, but they are labeled according to chains and not domains.

Fig. S11. Rigid-body fitting of chain A to C and D. A) The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). B) The EM map are shown as a transparent surface, and the fits of chain A to C from PF and ADP are shown as cartoons and colored by method and scores (dark blue: PF1, light blue: PF2, dark pink: ADP1, light pink: ADP2). The β-barrel (β), α-helical (α) and DUF1943 (DUF) domains are labeled. C) The EM map and the fits of chain D is shown in same coloring scheme as in panel B. The label is marked with "*" if the fit is outside the contour of the EM map. D) The EM map are shown as a surface, less transparent than in panel B, with the fits of chain A to C (1: PF2 and ADP1, 2: PF1 and ADP2) in the same coloring scheme as in panel B. The EM map is shown at four different angels, and arrows points to secondary structure elements from β, α or DUF domain which are outside the contour of the EM map.

Fig. S12. Rigid-body fitting for previously published homology models and a distant homologue. A) The same plot as in Fig. S10 for the β-barrel and α-helical subdomains, and the crystal structure of lamprey Vg (1LSH). B-D) Same presentation and coloring scheme as in Fig. S10C-S10F.

Table S1. Alignment parameters.

Table S2. List of species used in the multiple sequence alignment.

Table S3. Loop building based on gaps in the structural alignment.

Table S4. Edited residues during quality control.

Table S5. Rigid-body fitting scores from PowerFit and Chimera.

Table S6. Rigid-body fitting scores from ADP_EM and Chimera.

Table S7. RaptorX structure prediction.

Appendix S1. wwPDB EM Validation Summary Report.

Appendix S2. Top ranked Vitellogenin model by AlphaFold.

Supplementary material

Table S1. Alignment parameters

The pairwise alignment was performed using EMBOSS Needle with default settings. The multiple alignment was performed using BLAST with default settings. The structural alignment was performed and altered in spdbv. The sequence identify is higher (30.6 %) in spdbv due to different default penalty scores.

Alignment	Matrix	Gap Open	Gap extend	End Gap penalty	End Gap open	End gap Extend
Emboss	BLOSUM62	10	0.5	false	10	0.5
BLAST	Automatically selected	11	1	N/A	N/A	N/A
Spdbv	PAM200	6	4	N/A	N/A	N/A

Table S2. List of species used in the multiple sequence alignment

UniProt ID	Species
sp|Q868N5.1|VIT_APIME	*Apis mellifera*
sp|Q2VQM6.1|VIT2_SOLIN	*Solenopsis invicta*
sp|Q7Z1M0.1|VIT1_SOLIN	*Solenopsis invicta*
sp|Q2VQM5.1|VIT3_SOLIN	*Solenopsis invicta*
sp|Q9U8M0.1|VIT1_PERAM	*Periplaneta americana*
sp|Q9BPS0.1|VIT2_PERAM	*Periplaneta americana*
sp|Q16927.2|VIT1_AEDAE	*Aedes aegypti*
sp|Q05808.1|VIT_ANTGR	*Anthonomus grandis*
sp|Q27309.1|VIT_BOMMO	*Bombyx mori*
sp|P55155.2|VIT1_CAEEL	*Caenorhabditis elegans*
sp|P05690.5|VIT2_CAEEL	*Caenorhabditis elegans*
sp|P80012.2|VWF_BOVIN	*Bos taurus*
sp|Q28833.2|VWF_PIG	*Sus scrofa*
sp|P04275|VWF_HUMAN	*Homo sapiens*
sp|Q8CIZ8|VWF_MOUSE	*Mus musculus*
sp|Q28295.2|VWF_CANLF	*Canis lupus*

Table S3. Loop building based on gaps in the structural alignment

The table shows the loop building preformed in spdbv. The gaps are numbered according to the structural alignment, and the specific residues, how many and type of gap is noted. The last three columns list the parameters given by spdbv for the selected loop (except gap 5-7, were loop building was unsuccessful).

Gap	Target vWF	Number of residues	Type	*Ab initio* Loop		
				Clash Score	Pair potential	Force field energy
1	S1443	1	Insertion	4	-3.15	1563.8
2	D1447-K1448	1	Deletion	-3	-2.27	117.3
3	P1460	1	Insertion	-6	-1.90	674.8
4	H1482-N1483	5	Deletion	-4	2.80	2138.5
5	V1494-G1504	11	Insertion	Removed res. 1494-1504 from sequence		
6	I1517-Y1526	10	Insertion	Removed res. 1515-1522 from sequence		
7	V1537-Y1544	8	Insertion	Removed res. 1537-1541 from sequence		
8	D1561	1	Insertion	-8	-0.81	-49.6
9	K1570-F1571	1	Deletion	-3	0.26	102.1
10	L1575-A1576	2	Insertion	-6	0.53	16872.0
11	D1589-Y1590	1	Deletion	-7	0.64	30.3
12	I1630	1	Insertion	-8	0.33	46059.2

Table S4. Edited residues during quality control

Based on the Ramachandran plot and bad contacts detected by ProCheck, the listed residue rotamer option were edited to the most optimal rotamer. Regions in the Ramachandran Plot can be defined as: A - Core alpha, a - Allowed alpha, ~a - Generous alpha, B - Core beta, b - Allowed beta, ~b - Generous beta, L - Core left-handed alpha, l - Allowed left-handed alpha, ~l - Generous left-handed alpha, p - Allowed epsilon, ~p - Generous epsilon, XX - Outside major areas.

Detected by Ramachandran Plot	Detected region	Region edited to	Detected by ProCheck "Bad Contacts" Edited to a more optimal rotamer option to avoid clashes
R1450	XX	p	K1448
H1482	~b	~a	K1457
E1484	XX	b	Y1459
K1485	XX	L	L1463
L1486	~l	~l	M1471
Q1526	XX	~a	N1478
F1563	XX	A	I1477
K1570	XX	~a	E1507
F1571	XX	~a	T1524
L1577	XX	~b	V1529
D1578	XX	l	F1531
D1589	XX	~b	I1536
S1631	~l	~l	V1547
S1632	~l	~l	D1572
			L1575
			M1583
			Y1592
			V1610
			Y1627

Table S5. Rigid-body fitting scores from PowerFit and Chimera.

The fits presented generated in PowerFit are ranked (Fit), which is also used when fitted the models in Chimera. Scores from PowerFit presented is the Cross Correlation score (CSS), Fisher z-score (Fish-z), the z-score as factor of standard deviations (rel-z) and the sigma difference to the best fit ($(z_1-z_N)/\sigma$). Chimera also generates Correlation score (C), in addition to the average map value (AVM). A count of the number of atoms outside the contour, from the total atoms in the model is generated. The percentage of atoms outside is also included.

	PowerFit					Chimera fit-in-map				
Model	Fit	CCS	Fish-z	rel-z (z/σ)	$(z_1-z_N)/\sigma$	C	AVM	Outside	Total	%
Full-length Vg	1	0.415	0.442	32.2	0.00	0.7575	0.06003	6250	13277	47
	2	0.391	0.413	30.0	2.14	0.7417	0.05491	6984	13277	53
	3	0.389	0.411	29.9	2.26	0.8114	0.06224	6015	13277	45
RaptorX	1	0.371	0.390	28.2	0.00	0.6961	0.04817	8882	14381	62
	2	0.356	0.372	26.9	1.28	0.7265	0.05336	7981	14381	55
	3	0.352	0.368	26.6	1.58	0.6971	0.05010	8135	14381	57
Chain A	1	0.460	0.498	28.0	0.00	0.8877	0.08099	2064	8301	25
	2	0.458	0.495	27.9	0.15	0.8465	0.07797	2558	8301	31
	3	0.448	0.482	27.1	0.89	0.8685	0.08124	1902	8301	23
Chain B	1	0.758	0.992	14.9	0.00	0.7051	0.01949	662	662	100
	2	0.700	0.867	13.0	1.87	0.6035	0.01896	662	662	100
	3	0.696	0.859	12.9	1.99	0.7116	0.01910	662	662	100
Chain C	1	0.602	0.696	18.3	0.00	0.8163	0.09602	32	1714	2
	2	0.598	0.690	18.1	0.16	0.8167	0.09401	62	1714	4
	3	0.558	0.629	16.6	1.74	0.7800	0.09085	344	1714	20
Chain D	1	0.695	0.858	20.2	0.00	0.1898	0.00622	1376	1376	100
	2	0.690	0.848	20.0	0.24	0.9022	0.09620	311	1376	23
	3	0.677	0.824	19.4	0.81	0.4387	0.00542	1376	1376	100
Chain E	1	0.538	0.602	12.1	0.00	0.8309	0.09487	138	1224	11
	2	0.536	0.598	12.1	0.07	0.6326	0.01540	1224	1224	100
	3	0.527	0.586	11.8	0.32	0.6617	0.01798	1224	1224	100
Chain A to C	1	0.449	0.483	32.0	0.00	0.8776	0.07667	3362	10677	31
	2	0.433	0.464	30.7	1.29	0.8778	0.07665	3362	10677	31
	3	0.428	0.457	30.3	1.70	0.8474	0.07520	3624	10677	34
β-barrel	1	0.619	0.723	23.4	0.00	0.8235	0.09408	166	4824	3
	2	0.602	0.696	22.5	0.88	0.8137	0.09315	208	4824	4
α-helical	1	0.667	0.806	25.8	0.00	0.8277	0.0927	252	3678	7

	2	0.666	0.803	25.7	0.09	0.8277	0.0951	68	3678	2
1LSH	1	0.466	0.505	33.7	0.00	0.8987	0.08254	2688	10935	25
	2	0.461	0.499	33.3	0.42	0.8820	0.07918	3185	10935	29
AlphaFold	1	0.451	0.485	39.7	0.00	0.8552	0.07000	10724	28204	38
	2	0.427	0.457	37.3	2.34	0.8767	0.07414	8834	28204	31
	3	0.423	0.452	36.9	2.76	0.8747	0.07413	8821	28204	31
Vg dimer	1	0.326	0.338	38.8	0.00	0.7386	0.04643	35252	56408	62
AlphaFold	2	0.324	0.336	38.5	0.37	0.6609	0.04007	37296	56408	66
	3	0.310	0.321	36.8	2.03	0.6989	0.04340	36184	56408	64

Table S6. Rigid-body fitting scores from ADP_EM and Chimera.

The fits generated in ADP_EM are presented as in Table S5. The fits are ranked (Fit) and given a correlation score (C). The Chimera Fit-in-map scores are presented as in Table S5.

	ADP_EM		Chimera fit-in-map				
Model	Fit	C	C	AVM	Outside	Total	%
Full-length Vg	1	0.626	0.8003	0.06121	6025	13264	45
	2	0.615	0.8101	0.06224	6005	13231	45
	3	0.595	0.7877	0.06366	6195	12545	49
RaptorX	1	0.623	0.7126	0.05419	7800	13958	56
	2	0.622	0.6864	0.05103	7846	13768	57
	3	0.603	0.7153	0.05426	7838	13916	56
Chain A	1	0.624	0.8657	0.08148	2062	8301	25
	2	0.522	0.8438	0.07940	2479	8301	30
	3	0.474	0.8842	0.08103	2032	8301	24
Chain B	1	0.700	0.7851	0.09304	139	662	21
	2	0.559	0.7892	0.09314	135	662	20
	3	0.529	0.7813	0.10300	0	662	0
Chain C	1	0.573	0.8190	0.09597	33	1714	2
	2	0.558	0.8186	0.09603	8	1714	0
	3	0.522	0.8140	0.09557	25	1714	1
Chain D	1	0.681	0.9204	0.09540	132	1376	10
	2	0.498	0.9189	0.09542	132	1376	10
	3	0.489	0.8255	0.10040	5	1376	0
Chain E	1	0.694	0.7703	0.06953	495	1224	40
	2	0.533	0.8118	0.09954	32	1224	3
	3	0.517	0.8030	0.09483	55	1224	4
Chain A to C	1	0.64	0.8536	0.07537	3734	10677	35
	2	0.628	0.8776	0.07666	3353	10677	31
	3	0.578	0.8503	0.07322	3751	10677	35
β-barrel	1	0.621	0.8179	0.08273	1410	4824	29
	2	0.537	0.8323	0.96010	75	4824	2
α-helical	1	0.616	0.8177	0.09211	109	3678	3
	2	0.581	0.8318	0.09778	111	3678	3
1LSH	1	0.638	0.8984	0.08254	2693	10935	25
	2	0.567	0.8916	0.08058	3057	10935	28
AlphaFold	1	0.640	0.8603	0.06998	7168	24015	30
	2	0.603	0.8525	0.07730	8768	24015	37
	3	0.589	0.8283	0.06922	8504	24015	35

Vg dimer	1	0.620	0.6330	0.03619	37636	56408	67
AlphaFold	2	0.612	0.732	0.04550	34199	56408	61
	3	0.595	0.705	0.04344	35558	56408	63

Table S7. RaptorX structure prediction.

The resulting full-length structure prediction was compiled of six domains, which in again was based on homology modeling or template-free modeling. The amino acid sequence of honey bee Vg used in each domain are listed, as well as the alignment score, P-value, sequence identity and the PDB IDs of the templates.

Domain	Vg sequence	Model	Score*	P-value	Seq. ident.	Template
1	1-1142	1/1	801	1.5e-16	12	1LSH_A
2	1440-1643	1/1	133	1.9e-08	21	6RBF_A
3	1202-1439	1/2	141	1.6e-08	8	1LSH_B
		2/2	92	1.0e-05	6	3WJB_A
4	1143-1202	1/5	3	N/A	0	Template-free
		2/5	3	N/A	0	Template-free
		3/5	3	N/A	0	Template-free
		4/5	3	N/A	0	Template-free
		5/5	3	N/A	0	Template-free
5	1712-1770	1/5	54	1.2e-03	5	4YU8_A
		2/5	54	1.3e-03	9	4JPH_A
		3/5	53	1.5e-03	10	5BPU_A
		4/5	52	1.7e-03	14	4NT5_A
		5/5	52	1.7e-03	7	2KD3_A
6	1644-1712	1/5	5	N/A	0	Template-free
		2/5	5	N/A	0	Template-free
		3/5	5	N/A	0	Template-free
		4/5	5	N/A	0	Template-free
		5/5	5	N/A	0	Template-free

*Score: The alignment score which can go from 0 to the length of the domain sequence, with 0 indicating the lowest score.

Supplementary Figures

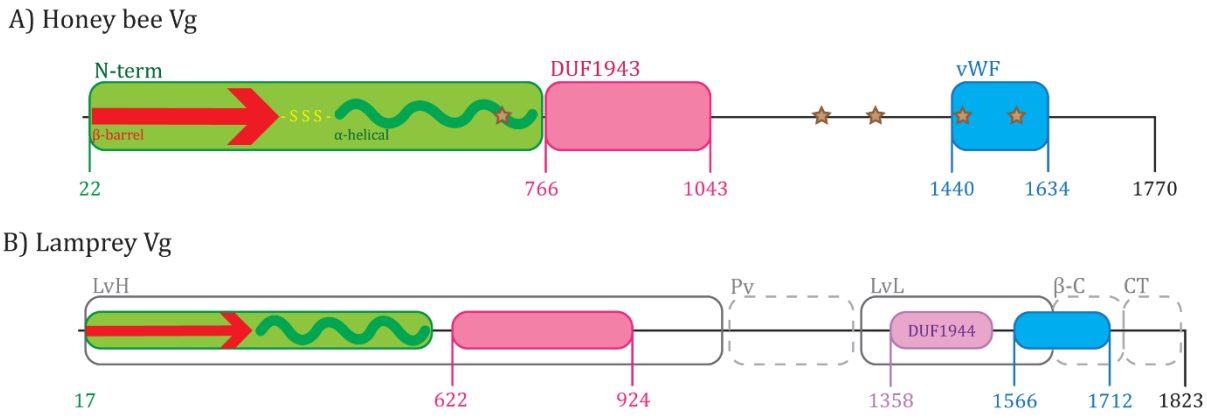


Figure S1. Domain architecture of honey bee and lamprey vitellogenin. The N-term (green), DUF1943 (pink) and vWF (blue) domains are conserved in both species, as well as the two structural subdomains, β-barrel (red arrow) and α-helical domain (dark green curved line). **A)** Honey bee Vg contains a proteolytic cleavage site, polyserine region (yellow S) linking the two subdomains. The five residue-positions (640, 1220, 1284, 1451 and 1536) identified to be candidates of functional polymorphisms are marked (brown stars). **B)** Lamprey Vg contains an addition domain, DUF1943 (purple). The yolk protein organization of IuVg is shown as gray boxes; lipovitellin heavy chain (LvH), Phosvitin (Pv), lipovitellin light chain (LvL), β-Component (β-C) and C-terminal coding region (CT). The dotted lines indicate that these regions (Pv, β-C and CT) are missing from the crystallographic structure (PDB ID: 1LSH).

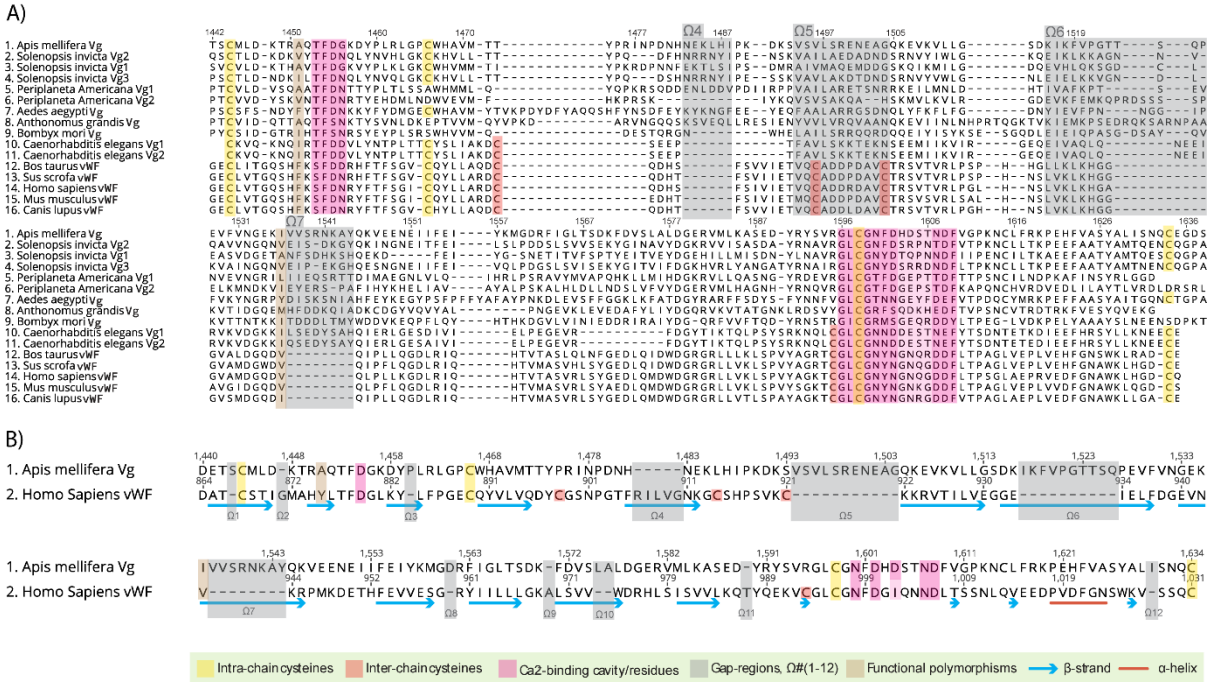


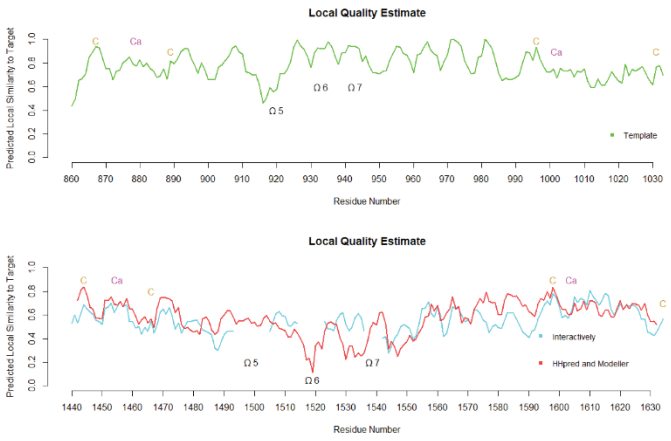
Figure S2. Multiple sequence and structural alignment. The coloring for the conserved residues/regions, gaps and secondary structure annotations are explained in the green box. The conserved Ca2+-binding region are colored in two shades of pink, dark pink is more conserved compared to the lighter pink. **A)** Extraction of the MSA. The original residue numbering for honey bee Vg is included on top. **B)** The final structural alignment with the original residue numbering included above each sequence. The annotations are retrieved from the template (PDB ID: 6N29). Both figures are created in Geneious Prime (v. 2019.0.3) and Adobe illustrator (v. 24.0.02).

A)

ProCheck summary	Template (6N29)	Target (Am vWF)
Number of residues	175	171
Ramachandran Plot		
– Core	86.8%	70.8%
– Allowed	12.6%	21.5%
– Generous	0.7%	7.6%
– Disallowed	0.0%	0.0%
All Ramachandran's	8 (out of 172)	23 (out of 161)
Chi1-chi2 plots	1 (out of 107)	3 (out of 114)
Side-chain parameters		
– Better	5	5
– Inside	0	0
– Worse	0	0
Residue properties		
– Max deviation	3.0	19.9
– Bad contacts	6	42
– Bond length/angle	2.5	35.3
– Morries et al class	1 2 2	2 2 3
Cis peptides	0	2
G-factors		
– Dihedrals	-0.20	-0.26
– Covalent	0.54	-1.11
– Overall	0.09	-0.57
Planar groups within limits	100.0%	100.0%

Accepted
May be worth investigating further
Worth investigating further

B)



C)

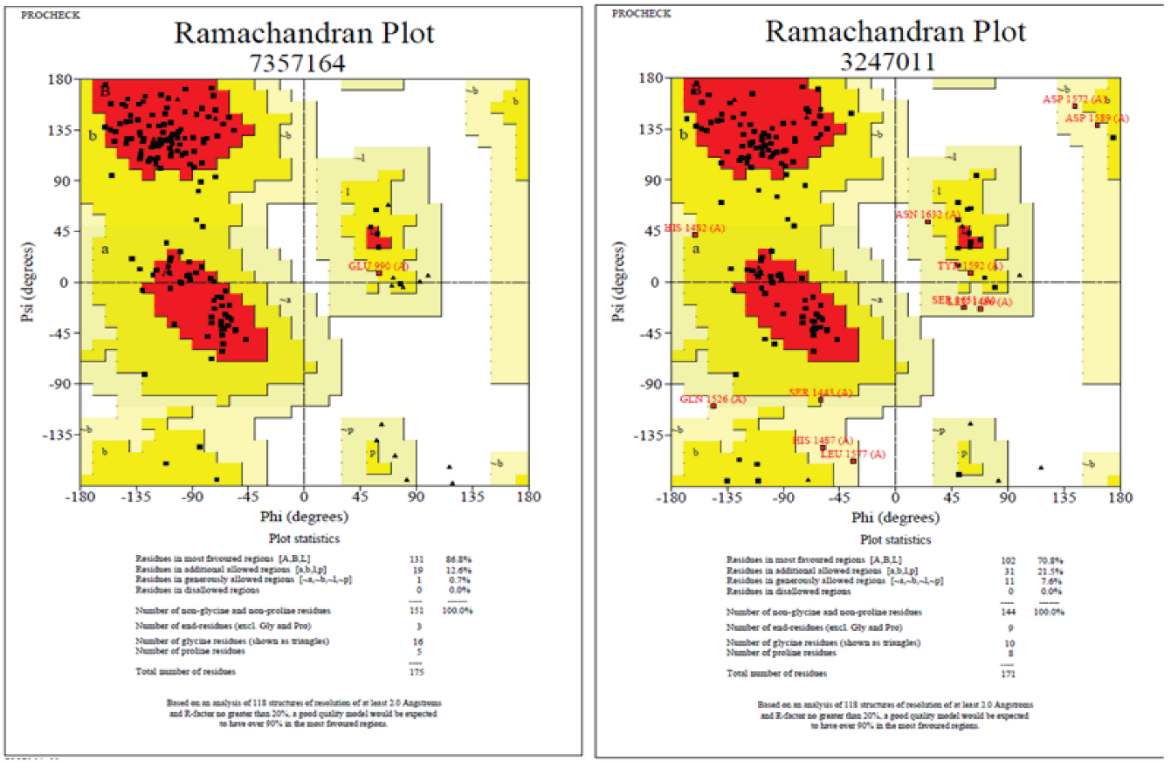


Figure S3. ProCheck summary, local quality estimate and Ramachandran plots. A) The ProCheck quality evaluations summarized and categorized by calculation results. The ideal residue values and standard deviation for any given model are derived from Morris et al. 1992.[1] The max deviation, in residues properties, is calculated from the mean value of the residue-by-residue listing values (Fig. S4E) of the full-length structure. The number of bad contacts is defined as the non-bonded atoms at a distance of <= 2.6 Å. The bond length and angles are calculated in similar manner as the max deviation, but the ideal values are based

on Engh and Huber 1991.[2] The Morris et al. (1992) class summarizes the three above stereochemical parameters by assigning a number between 1 (best) to 4 (worst), indicating the overall quality of the model. **B)** Local QMEAN results are presented. The first plot is analysis of the template (green), while the second is analysis of the target modeled interactively (cyan) and automatically (red). The Ca^{2+}-binding region (magenta Ca), the Cys residues forming the intra-chain disulfide bridges (orange, C) are in the higher quality region, while Ω5-7 (black) are in the lower quality region. The local score is calculated for each residue in the model and the average local score for the template is 0.93 ± 0.07, while the target average score is 0.40 ± 0.07 (cyan) and 0.44 ± 0.06 (red). **C)** The Ramachandran plot produced by ProCheck. The plot on the left is the template (PDB ID: 6N29), while the target (honey bee vWF domain) is on the right. Below each plot, the statistic is presented.

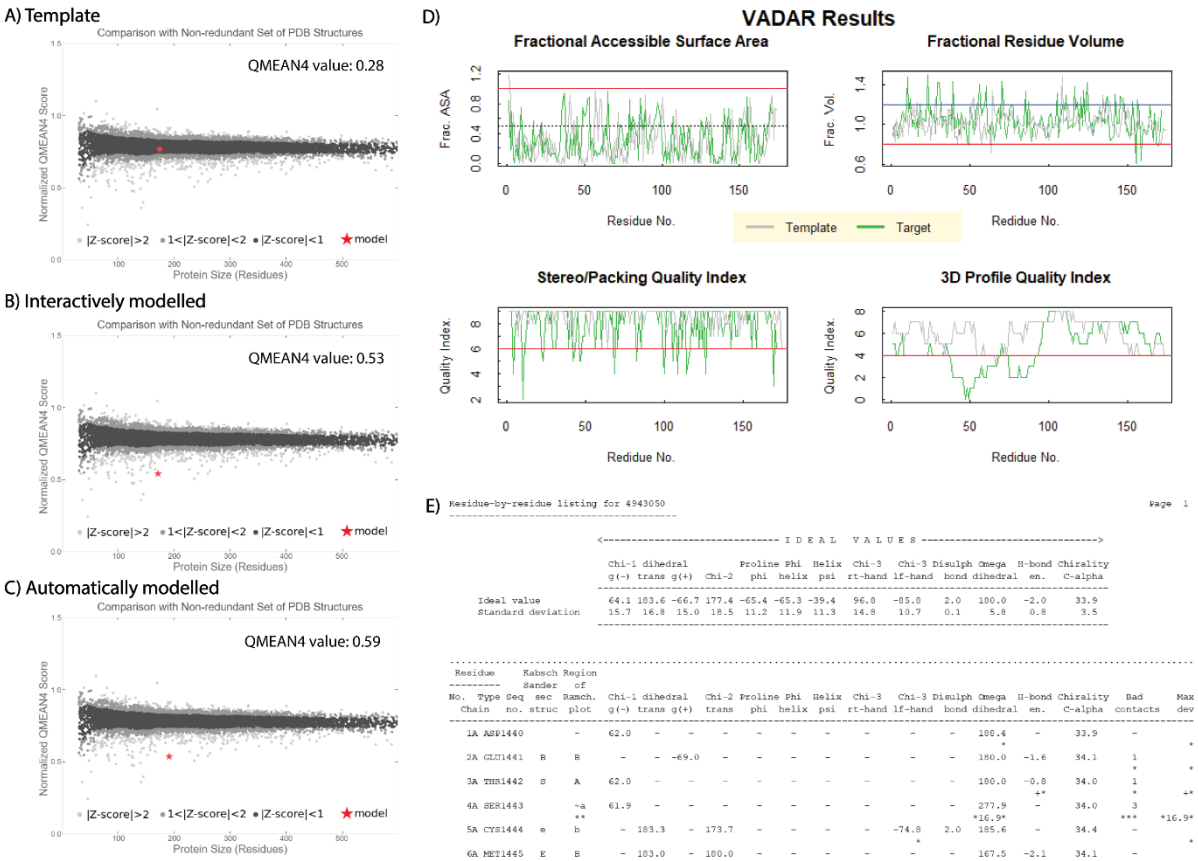


Figure S4. Global quality estimate, VADAR plots and ProCheck residue listing. A-C) The plots of the global QMEAN have the QMEAN4 scores for a set PDB structures plotted (gray dots) with the QMEAN4 score along the x-axis and the number of residues in the structures as long the y-axis. The global scores value QMEAN4 range from 0 to 1, where 1 is good. **A)** Analysis of the template (red star) and the QMEAN4 value is written on the plot. **B)** Analysis of the interactively homology modeled (red star) structure and **C)** The automatically homology modeled (red star) structure from MODELLER. **D)** Four different analyses were performed by VADAR, presented in one plot each, with the template (gray) compared to the target (green). **Plot 1**: a low fractional ASA score indicates a buried residue, while a score above 0.5 (dotted black line) indicates an exposed residue. A score above 1.0 (red line) indicates a problem in the structure. **Plot 2**: When a protein structure is efficiently packed the score should be around 1.0 ± 0.1. A score above 1.2 (blue line) or below 0.8 (red line) could indicate a poor

refinement or identify cavities. **Plot 3**: Each residue is assigned a score between 0-3 (high is good quality) for three different measurements (torsion angle, omega angle and fractional volume). The total quality score for each residue can be from 0-9 and the threshold for a good quality is set to 6 (red line). **Plot 4**: Calculates the 3D quality of each residue based on its environment and gives a score between 0-9 (high is good quality), and the threshold for a good quality is set to 4 (red line). **E)** The Residue-by-Residue listing for ProCheck lists all residues in a structure and present all calculations for each. A short example is shown here for the first six residues in the target structure. Each value is compared to the ideal values which is noted on top. The deviating values are marked with * (one standard deviation) and + (half a standard deviation) sign. For example, the omega dihedral angle of residue S1443 is 16.9 standard deviation away from the ideal value, which is a result from the loop building of Ω1.

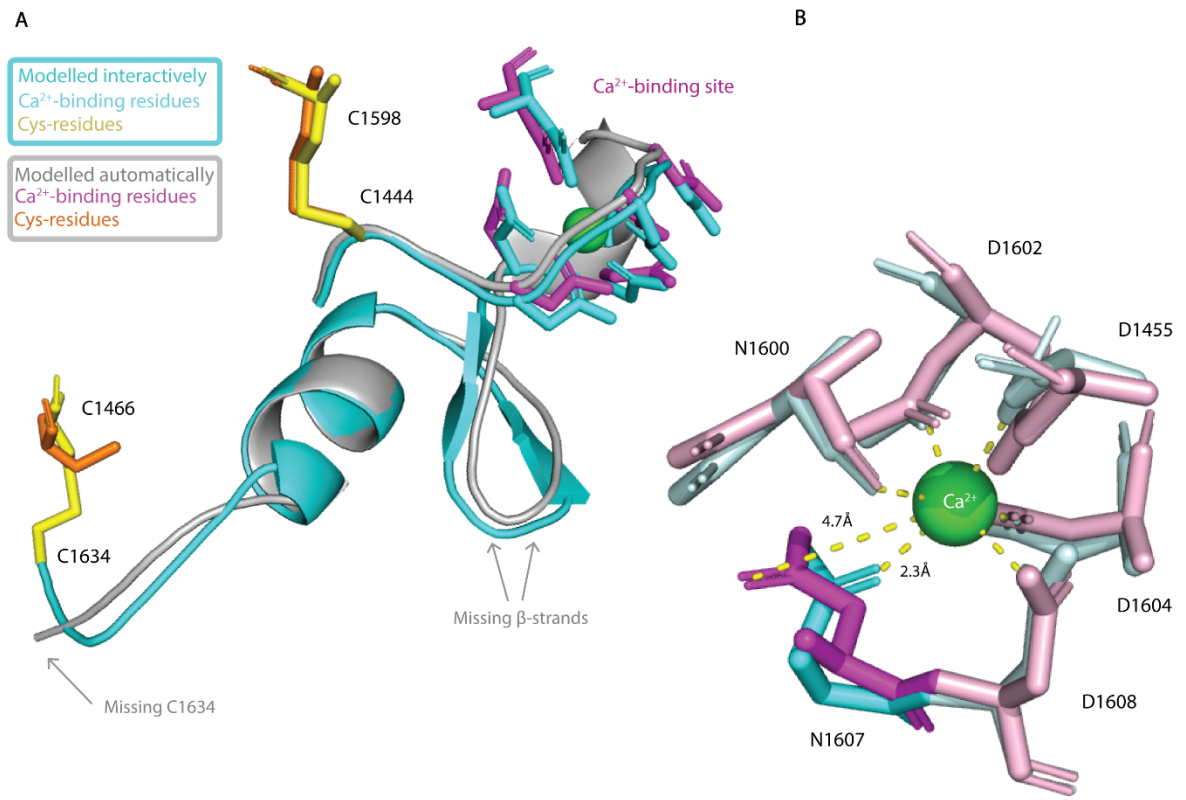


Figure S5. Comparison of vWF homology models. A) The sort region around the Ca^{2+}-binding site (Ca^{2+}-ion, green) is shown from the interactively modeled (cyan) structure and the automatically modeled (gray) structure. The Cys-residues (C1444, C1466, C1598 and C1634) and Ca^{2+}-binding residues are shown as yellow/cyan (interactively) and orange/magenta (automatically) sticks. The missing C1634 and β-strands in the automatically modeled structure are shown (gray arrows). **B)** All the Ca^{2+}-binding residues are in the same orientation in both models (light blue: interactively and light pink: automatically), except N1607. The interactions to the Ca^{2+}-ion is shown as yellow dotted lines and measured (Å) for N1607.

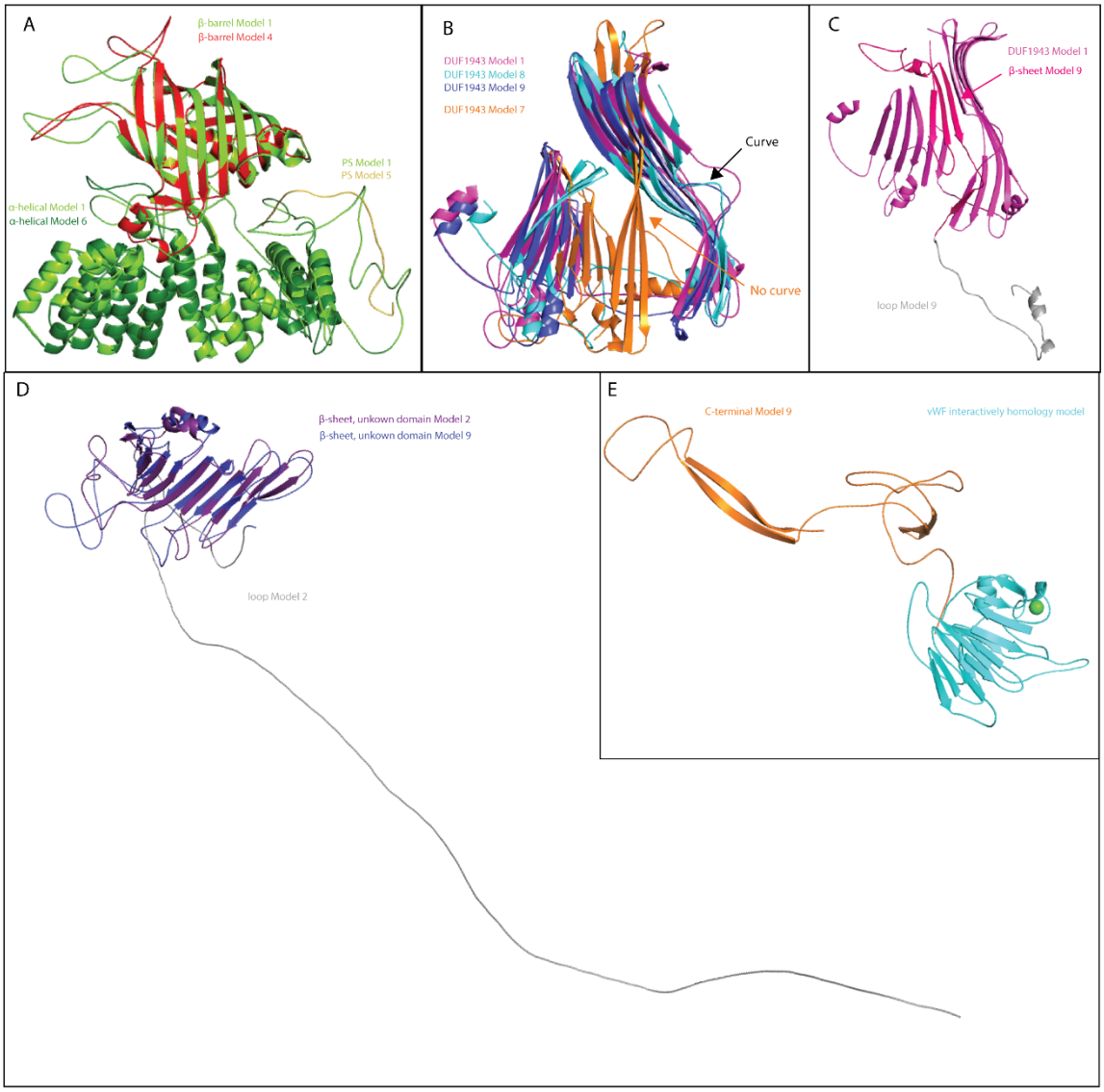


Figure S6. Comparison of homology models from MODELLER and RaptorX. A) The N-terminal domain: Model 1 (green) aligned with Model 4 (red), 5 (yellow) and 6 (forest green). **B) T**he DUF1943 domain: Model 1 (magenta) aligned with Model 8 (cyan), Model 7 (orange) and Model 9 (blue). The identified curve in the longer β-sheet in Model 1, 8 and 9 and the missing curve in Model 7 is marked with arrows. **C)** The DUF1943 domain Model 1 (magenta), the downstream region residue 1060 to 1140 of Model 9 (hot pink) and the loop region (gray). **D)** The undetermined domain: Model 2 (purple) aligned with Model 9 (blue), with the long loop

region (gray). **E)** The interactively homology model of vWF domain (cyan) with the C-terminal region from Model 9 (orange).

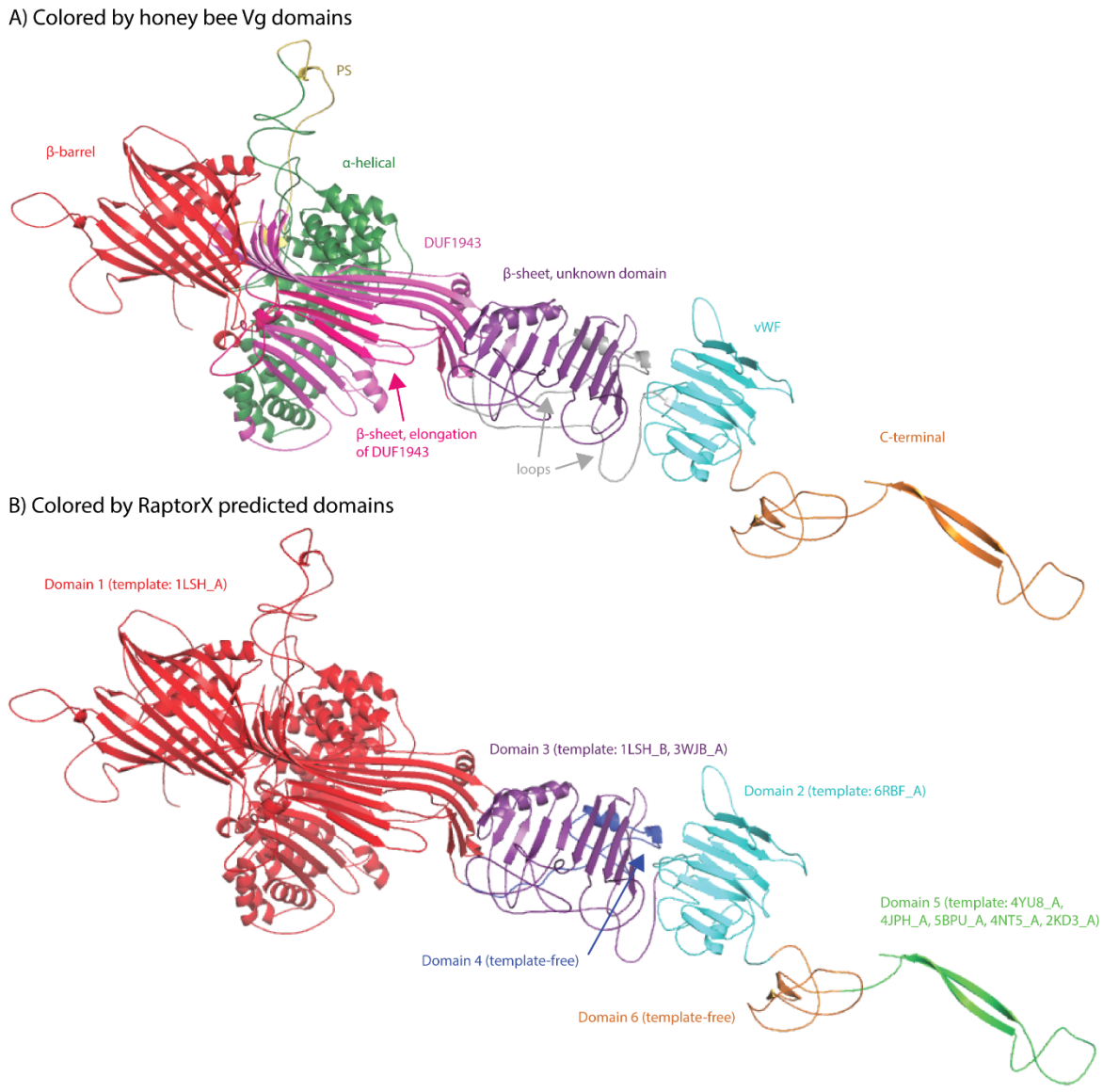


Figure S7. RaptorX structural prediction of full-length honey bee vitellogenin. A) The β-barrel subdomain (red), the polyserine linker (yellow), the α-helical subdomain (forest green), the DUF1943 domain (magenta), elongation of the DUF1943 domain (hot pink arrow), the undetermined structural region (purple), the vWF domain (cyan) and the C-terminal region (orange) are generated as one full-length model. The two loop regions (gray arrows) are also

predicted. **B)** Domain 1 to 6 from Table S7 are colored red, cyan, purple, blue, green and orange, respectively, and if templates was used, the PDB ID is written in parenthesis.

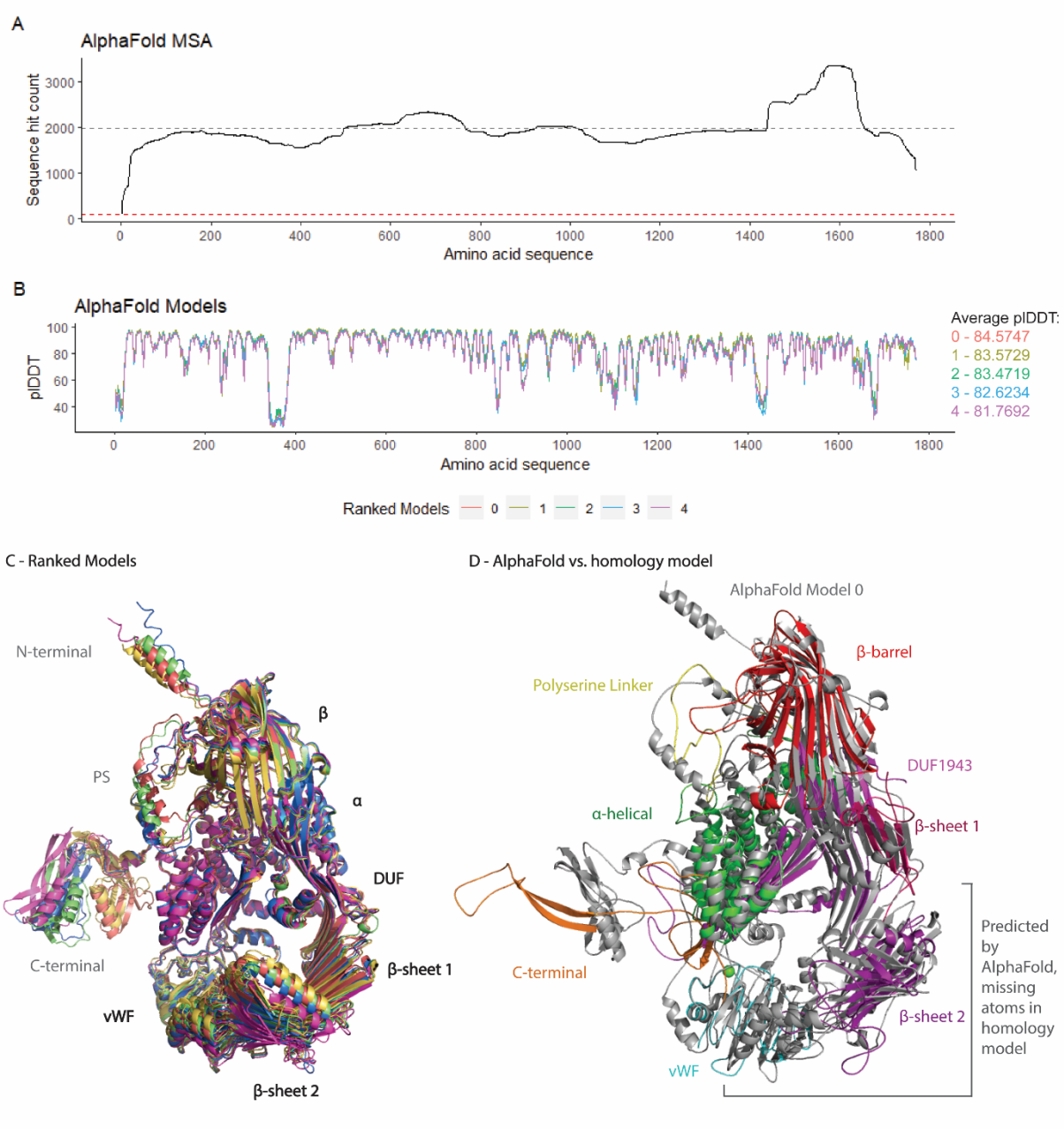


Figure S8. AlphaFold output. A) The number of sequence hits in the MSA produced by AlphaFold, is plotted per residue. The average number of hits per residue (gray dotted line), and the threshold at 100 sequence per residue (red dotted line) is marked. **B)** The plDDT score

for the five outputted models by AlphaFold is plotted per residue, and the average plDDT score per model is listed to the right, which produces the rank from 0 (best) to 4 (worst). **C)** The ranked models are aligned, colored by the same coloring scheme in panel B, and the consistently folded domains (β-barrel (β), α-helical (α), DUF1943 (DUF), β-sheet 1 (β1), β-sheet 2 (β2) and vWF domain (vWF)) are labeled in bold letters, while the more variable domains (N-terminal, polyserine linker (PS) and C-terminal) are labeled in grey letters. **D)** The final homology model domains (β-barrel (red), polyserine linker (yellow), α-helical (green), DUF1943 (magenta), β-sheet 1 (hotpink), β-sheet 2 (purple), vWF (cyan, Ca^{2+}-ion shown as green sphere) and C-terminal domain (orange) is aligned to their respective domains in the top ranked AlphaFold prediction (grey). The grey brackets to the lower right indicate the region where AlphaFold have predicted a fold for the main missing atoms in the homology model.

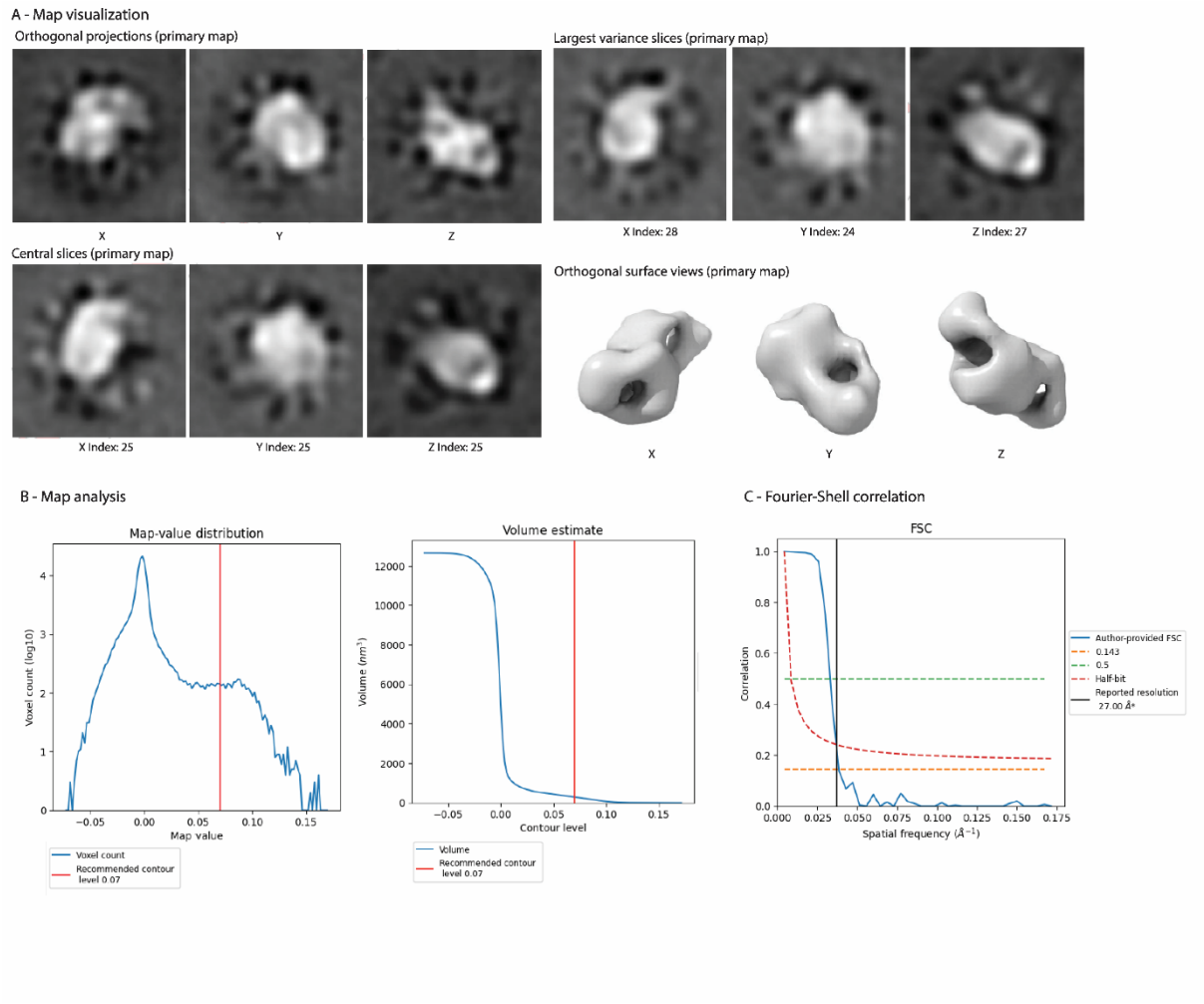


Figure S9. EM map validation. A) Map visualization to allow visual inspection of the internal detail of the map and identification of artifacts. The primary map, central slices of the map and largest variance of the map is shown in three orthogonal directions. The 3D surface view of the primary map at recommended contour level 0.07. **B)** Statistical analysis of the map. In the first graph the map-value distributions is plotted in 128 intervals along the x-axis, and the y-axis is logarithmic. The spike around 0 indicate that the volume has been masked. The second graph shows how the enclosed volume varies with the contour level. The volume at the recommended contour (red line) is 289 nm3; this corresponds to an approximate mass of 261 kDa. **C)** The provided Fourier-Shell Correlation (blue) is plotted together with the reported resolution, (black line, *Reported resolution corresponds to spatial frequency of 0.037 $Å^{-1}$). A

curve is displayed for the half-bit criterion (dashed red), in addition to lines showing the 0.143 gold standard cut-off (dashed orange line) and 0.5 cut-off (green dotted line). All the graphs are assembled from the EmDataBank map validation report (copy included).

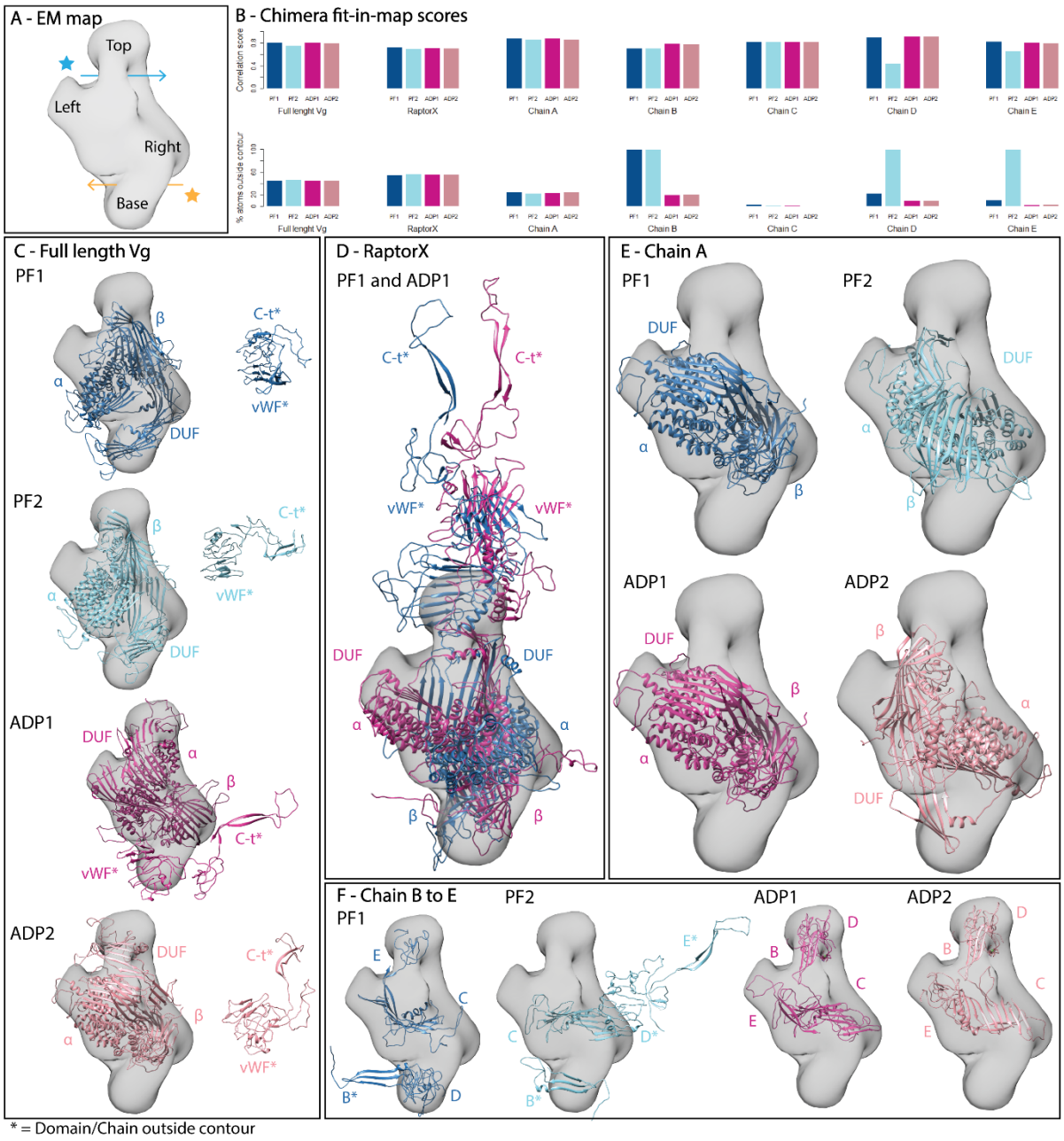


Figure S10. Rigid-body fitting for honey bee vitellogenin homology models. **A)** The EM map is shown as a gray surface. The distinct cavity creases are marked with stars and arrows, upper cavity (blue) and lower cavity (yellow). The four curves in the surface are labeled (top, base, left and right). **B)** The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). **C-E)** The fits from the full-length homology model, RaptorX and chain A is

presented inside the EM map, with the same coloring scheme as in panel B. The β-barrel (β), α-helical (α), DUF1943 (DUF), vWF and C-terminal (C-t) domains are labeled. If the domain is outside of the contour it is noted by a "*"-mark. **F)** The fits of chain B to E separately with the same coloring scheme as in panel B, but they are labeled according to chains and not domains.

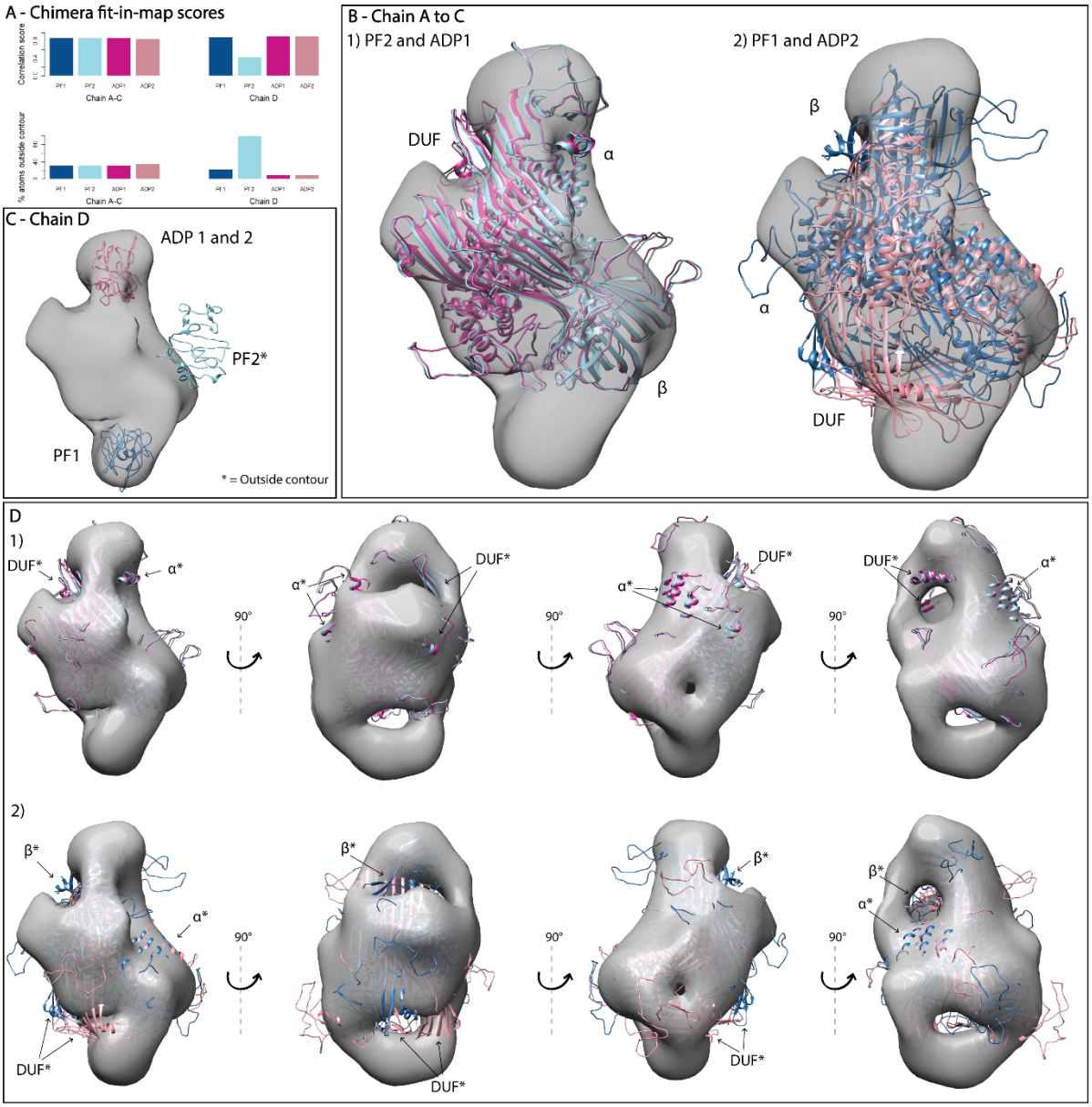


Figure S11. Rigid-body fitting of chain A to C and D. A) The correlation score and precent of atoms outside the contour calculated by Chimera was plotted for each fit from PowerFit (PF, blue) and ADP_EM (ADP, pink), and ranked according to the correlation score (dark color: highest score, light color: second highest score). **B)** The EM map are shown as a transparent surface, and the fits of chain A to C from PF and ADP are shown as cartoons and colored by method and scores (dark blue: PF1, light blue: PF2, dark pink: ADP1, light pink: ADP2). The β-barrel (β), α-helical (α) and DUF1943 (DUF) domains are labeled. **C)** The EM map and the fits

of chain D is shown in same coloring scheme as in panel B. The label is marked with "*" if the fit is outside the contour of the EM map. **D)** The EM map are shown as a surface, less transparent than in panel B, with the fits of chain A to C (1: PF2 and ADP1, 2: PF1 and ADP2) in the same coloring scheme as in panel B. The EM map is shown at four different angels, and arrows points to secondary structure elements from β, α or DUF domain which are outside the contour of the EM map.

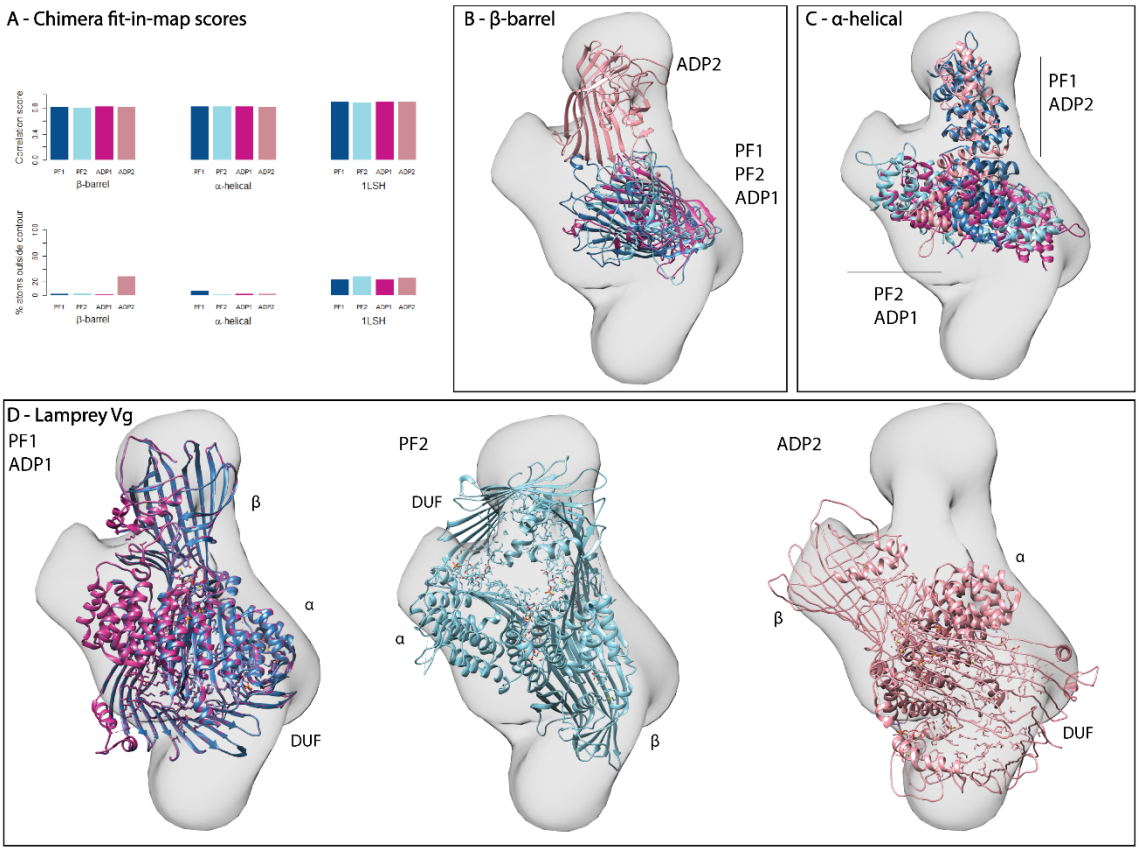


Figure S12. Rigid-body fitting for previously published homology models and a distant homologue. A) The same plot as in Fig. S10 for the β-barrel and α-helical subdomains, and the crystal structure of lamprey Vg (1LSH). **B-D)** Same presentation and coloring scheme as in Fig. S10C-S10F.

References

1. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. Proteins. 12(4):345-364.
2. Engh RA, Huber R. 1991. Accurate bond and angle parameters for x-ray protein structure refinement. Acta Crystallographica Section A. 47(4):392-400.

VI