A structural investigation of novel fungal polyglycine hydrolases

by

Nicole Dowling

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Biology

Waterloo, Ontario, Canada 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor	Dr. David Rose Professor Emeritus – University of Waterloo (Biology)
Internal Member	Dr. Todd Holyoak Associate Professor – University of Waterloo (Biology)
Internal Member (2)	Dr. Moira Glerum Professor – University of Waterloo (Biology)
Internal-External Member	Dr. Brian Ingalls Professor - University of Waterloo (Applied Mathematics)
External Examiner	Dr. Lynne Howell Professor - University of Toronto (Biochemistry)

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The statement of contributions is in accordance with the Contributor Roles Taxonomy (CRediT) for each chapter of this thesis.

Chapter 1

All writing for this chapter is solely generated by Nicole Dowling.

Chapter 2

This chapter is a manuscript submission in collaboration with the USDA Agricultural Research Service in Peoria, IL. Nicole Dowling: Conceptualization, formal analysis (structure), investigation, methodology, visualization, writing - original draft preparation. Dr. Todd Naumann: Conceptualization, formal analysis (enzyme activity), funding acquisition, investigation, visualization, writing - review and editing. Neil Price: Conceptualization, funding acquisition, investigation, writing - review and editing. David Rose: Conceptualization, funding acquisition, supervision, writing - review and editing.

Chapter 3

All data, analysis and writing for this chapter is solely generated by Nicole Dowling.

Chapter 4

All analysis and writing for this chapter are solely generated by Nicole Dowling.

Abstract

Polyglycine hydrolases (PGH) are a family of fungal proteases that are known to cleave the polyglycine linker of *Zea mays* chitinase, ChitA, thwarting one mechanism of plant defense against fungal infection. Previously, little was known at the atomic level about the interaction between these proteases and their target. There has been limited biochemical characterization and no structural characterization of this family of proteases. In this work, we analyze the atomic structure of one of these polyglycine hydrolases, Fvan-cmp. The structure was solved by X-ray crystallography using a *de novo* RoseTTAFold model. We report models for the other identified polyglycine hydrolases utilizing the previously determined structure, as well as insights into features likely involved in the catalytic mechanism. The PGH structural characterization identified a two-domain structure, simply named N- and C- domain. The N-domain is a novel tertiary fold found throughout all kingdoms but functionally unidentified. The C-domain shares structural similarities with Class C β-lactamases including the conserved active site motifs and catalytic residues. Utilizing a combination of *in vitro* and *in silico* methods, we propose a PGH-ChitA complex model that is supported by previous understanding of PGHs and the structural data. Throughout this work, we discuss the merits and limitations of current *in silico* methods with a focus on *de novo* protein modelling and protein-protein docking methods.

Acknowledgements

This work would not have been possible without the support of many people. I would like to acknowledge the generosity of the PlantForm Corporation (Guelph, Ontario) coupled with Mitacs. The Mitacs Accelerate program permitted the exploration of this work and provided the base for non-academic network-building.

I would like to thank my committee for your constant support throughout this entire project. You saw me through the beginning of my Master's degree, the transition to my PhD degree and throughout the many pivots we took in this project. Dr. Todd Holyoak and Dr. Moira Glerum - you both offered incredible advice, found the 'holes' in my experiments and questioned everything. You are both responsible for the scientist I have become.

I would like to thank the graduate students that I befriended over these years that offered a space for venting, experimental advice, and just good friendship. A special shoutout to Michelle McKnight and Norman Tran for being there for the good, the bad and the ugly.

I would like to thank my family for their support over the last five years as I changed degree paths, changed research projects, and constantly talked about my work at every gathering. No one was more supportive of this work than my husband, Ian. You may not understand the intricacies of my work, but you understood how much I cared about it. Thank you for the late lab dinners, listening to countless presentations, and being at every graduate milestone.

I would like to thank Dr. Mungo Marsden for your time and continuous support of my work. You allowed me the opportunity to work in your lab as a BIOL 499 undergraduate student and fostered my love of research. You offered a listening ear, sage advice, and funny anecdotes on several occasions. I would not have pursued a graduate career if I hadn't spent those eight months in your lab.

Lastly, I would like to thank my supervisor, Dr. David Rose for everything over these five years. You accepted me as a graduate student with no biochemistry, no glycobiology nor structural biology experience. Honestly, I'm still a little surprised that you said yes. I cannot thank you enough for all the opportunities that you've given me throughout this degree. You supported me through many research questions, edits, and conferences. You gave me the space to become an independent scientist. You are an incredibly brilliant, patient, and thoughtful mentor and I was incredibly lucky to have you as my supervisor.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BRENDA	The comprehensive enzyme information system
Bz-cmp	*Bipolaris zeicola* chitinase-modifying protein
CAZy	Carbohydrate Active enZymes database
ChitA	*Zea mays* chitinase, ChitA alloform
CMP	Chitinase-modifying protein
E	Enzyme
Es-cmp	*Epicoccum sorghi* chitinase-modifying protein
Fvan-cmp	*Fusarium vanettenii* chitinase-modifying protein
GlcNAc	N-acetylglucosamine
HADDOCK	High Ambiguity Driven protein-protein DOCKing server
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IUB	International Union of Biochemistry
MES	2-(N-morpholino)ethanesulfonic acid
MSA	Multiple sequence alignment
PAE	Predicted aligned error
PBP	Penicillin-binding protein
PEG	Polyethylene glycol
pLDDT	Predicted local distance difference test
PGH	Polyglycine hydrolase
PTM	Post-translational modification
PDB	Protein Data Bank
RMSD	Root mean square deviation
S	Substrate
Tris	Tris(hydroxymethyl)aminomethane
WT	Wild type

Chapter 1 : Chitinases, chitinase-modifying proteins & protein modelling methods

1.1 Introduction

1.1.1 Chitinases

Classification

Chitinases are glycoside hydrolases found in many organisms across the different kingdoms[1]. These enzymes are extraordinarily diverse, which may contribute to their ubiquitous presence within the kingdoms. Owing to this diversity, chitinases are classified by a variety of characteristics: their biochemistry, amino acid sequence, and locus of activity. Biochemically, the International Union of Biochemistry (IUB) identified four classes of chitinases; these are compiled within the comprehensive enzyme information system (BRENDA)[2]. Five Glycoside Hydrolase (GH) families have been identified for chitinases categorized by the Carbohydrate Active Enzymes (CAZy) database, http://www.cazy.org/[3]. In general, chitinases can be distinguished by whether they perform endo- or exo- cleavage of the chitin polymer (Table 1)[4].

Presently, there are seven classes of chitinases distinguished by their N-terminal sequence, sub-cellular localization, isoelectric pH (pI), signal peptide presence and the inducers[1,4]. The chitinases classes have been reviewed in detail by several authors[1,4–6]. It is well-established that the chitinase classes vary extensively in mechanism and structure[5]. The classes can be sorted into two groups based on sequence similarity. Group 1 is comprised of classes I, II and IV, while Group 2 encompasses classes III and V. Class I possess cysteine-rich N-terminus sequences, a leucine- or valine-rich signal peptide, distinct conserved loop structures and are localized to the vacuole. This class consists entirely of plant chitinases with the majority having

endo-activity[4,7]. Class II chitinases are like Class I but lack the cysteine-rich N-terminal sequence and are found in plant, fungi and bacteria and mostly consisting of exo-chitinases. Class III is a distant class of chitinases with no similarity to either classes I or II and contains a conserved DXDXE motif[1,8]. Class IV chitinases share similar characteristics with Class I but are substantially smaller sized by the absence of conserved loop structures[5,9]. Class V chitinases are found in plants and share the Class III DXDXE motif but differ in their enzymatic mechanism[1,8].

Table 1. Enzymes Classification of Chitinases

EC Number	GH Family	Type of Chitinases	Originating Kingdoms
3.2.1.14	18 19 23	Endo-chitinase Exo-chitinase	All
3.2.1.52	3 18 20 84	Exo-chitinase	All
3.2.1.200	18	Exo-chitinase	Bacteria, Fungi, Plants
3.2.1.201	18 19	Exo-chitinase	Archaea, Bacteria

Chitinases are diverse enough in sequence and biochemistry to belong to several EC and GH categories. The prevalence of IUB classified chitinases within the kingdoms are noted.

Plant Chitinases

Chitinases are an interesting group of enzymes as they have organism-specific biological roles. In bacteria, chitinases play a role in nutrition and parasitism whereas in fungi their role is in morphogenesis[4]. Plant chitinases contribute to plant defense mechanisms with location-specific alloforms[4,10]. The localization of chitinases varies extensively in plants; most commonly found in the vacuole (class I), apoplasts (class II, IV) and some are extracellularly located[11].

The *Zea mays* ChitA chitinase (EC. 3.2.1.14) belongs to GH19 family and is a class IV basic chitinase identified in maize defense against ear rot infections. ChitA has two distinct domains connected by a polyglycine linker: 4 kDa amino-terminal hevein-like domain and 24

kDa carboxy-terminal chitinase domain[12]. The hevein-like domain contains the N-terminal signal sequence and is responsible for chitin binding while the chitinase domain - as the name suggests - is responsible for the catalytic activity[13]. A truncated structure of the *Zea mays* ChitA chitinase domain (PDB ID: 4MCK) was solved by X-ray crystallography [14]. In Figure 1, we highlight the crystal structure and provide a representation of the full-length ChitA enzyme.

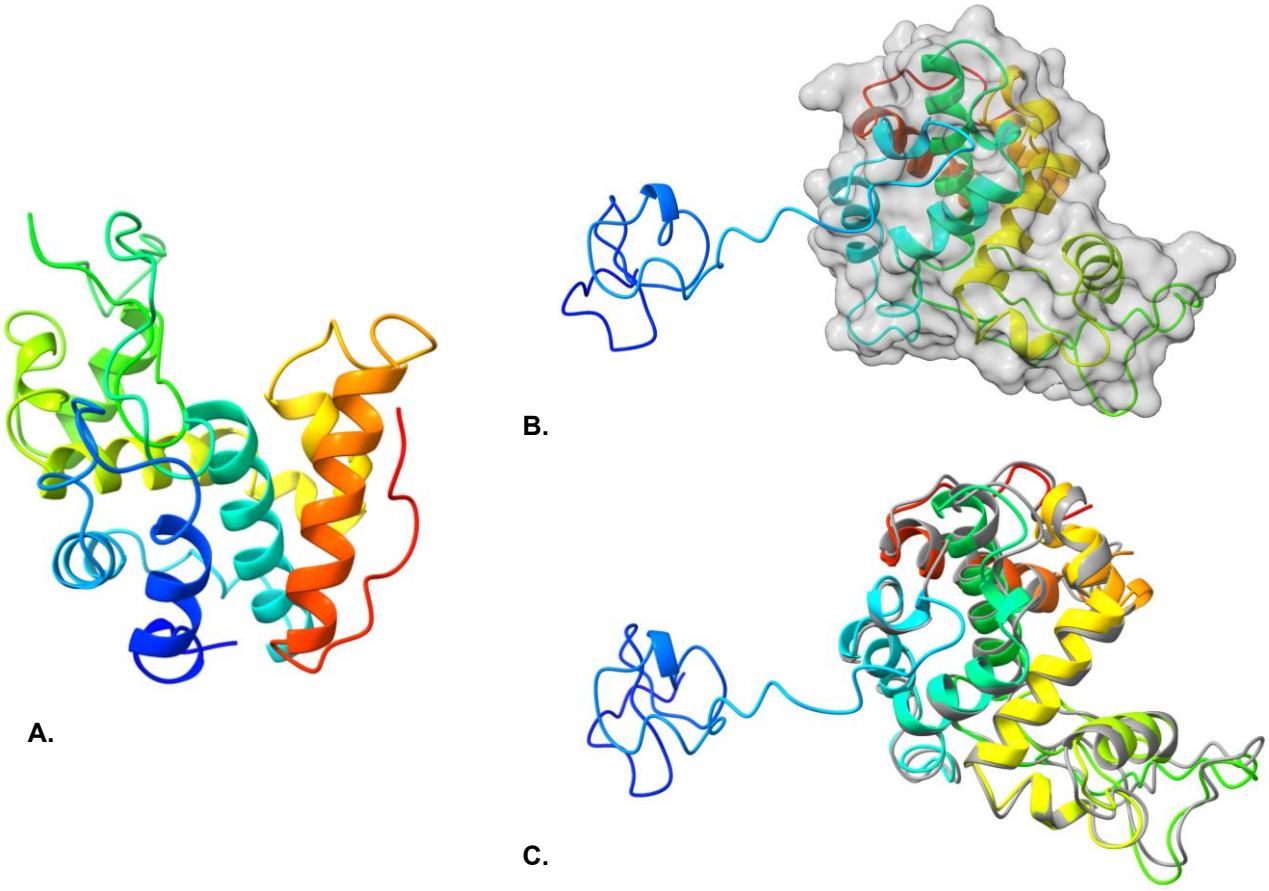


Figure 1. ChitA ΔN-EQ and full-length ChitA structure

This figure uses the rainbow spectrum colour scheme to aid in visualization of the three-dimensional structure in a two-dimensional space. ChimeraX-1.4 was used to align and visualize the proteins.
(A) The crystal structure of truncated *Zea mays* ChitA chitinase (PDB ID: 4MCK)[14]. The structure omits the hevein-like domain and polyglycine linker. The first 56 residues were omitted from the structure and represented by the naming modification of ChitA ΔN-EQ.

A full-length ChitA model (rainbow) was generated through RoseTTAFold and aligned to the crystal structure of ChitA (grey)[14,15]. The alignment included 182 atoms with a final Cα-rmsd of 1.006 Å. The surface representation of the crystal structure (B) and the ribbon representation of the crystal structure (C) provide insight to the completeness and accuracy of the model.

ChitA has been suggested to play a role in seed protection during germination and may contribute to inhibiting fungal hyphae spread through its cleavage of chitin polymers[10,13]. The activity of this enzyme is achieved by its catalytic triad: Glu62, Arg177 and Glu165[14]. This enzyme works by a single displacement mechanism, inverting the anomeric carbon, seen in Figure 2[16,17]. *In vitro* activity studies have shown: (i) preference for tetrameric substrates over dimer substrates; and (ii) cleavage of two N-acetylglucosamine (GlcNAc) residues on the reducing end of chitin polymers[18].

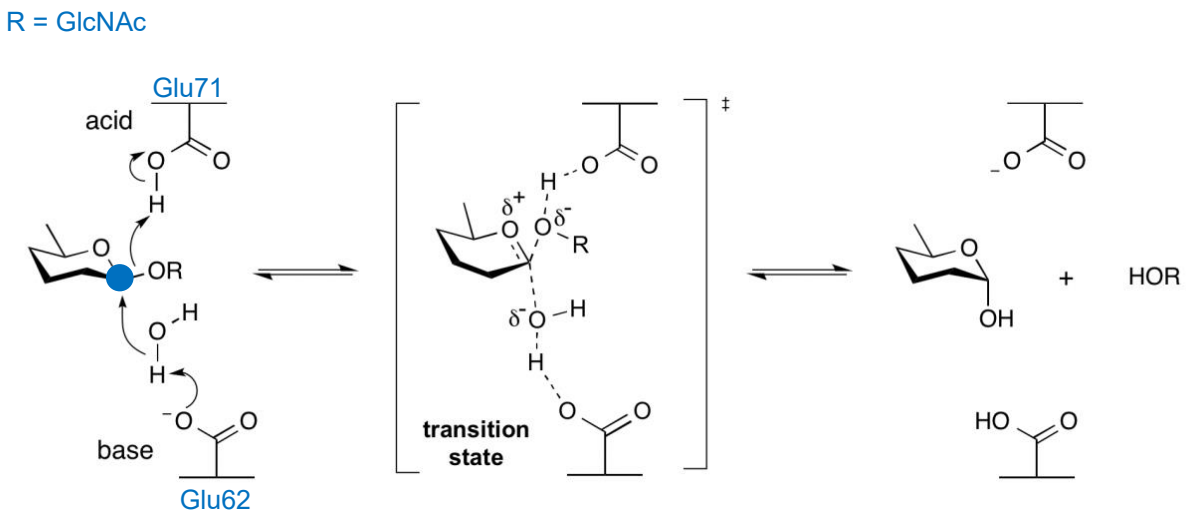


Figure 2. Single Displacement Mechanism for chitinolytic enzymes

The mechanism displayed, described a single displacement mechanism with inversion of the anomeric carbon (blue). The two residues directly involved are Glu62 and Glu71 which have been identified as the proton donor and general base, respectively[14]. This figure is adapted from the CAZy*pedia* resource on glycoside hydrolases[19].

As previously discussed, *Zea mays* ChitA and its alloform, ChitB contain a polyglycine linker that connects their hevein-like domain to their chitinase domain. This region is a

proteolytic target for a subclass of chitinase-modifying proteins which renders the affected chitinases inactive[20].

1.1.2 Chitinase-modifying proteins

Chitinase-modifying proteins (cmp) are proteases that are secreted by fungi to truncate chitinases[21]. Three types of proteases encompass chitinase-modifying proteins: fungalysin metalloproteases, kilbournases and polyglycine hydrolases[12,21,22]. Polyglycine hydrolases (PGH) are presently the only type of chitinase-modifying proteins found to truncate class IV plant chitinases by cleavage of glycine-glycine bonds[12].

Polyglycine hydrolases

Polyglycine hydrolases have been identified in the fungal classes of Dothideomycetes and Sordariomycetes and have been recombinantly expressed from *Bipolaris zeicola*, *Epicoccum sorghi*, *Fusarium vanettenii*, and *Galerina marginata*. PGHs are classified as serine proteases that belong to the S12 family[20]. This family of serine proteases is shared with D-alanyl-D-alanine carboxypeptidases, β-lactamases, and penicillin-binding proteins[23]. This family functions through an acyl-enzyme intermediate mechanism with a serine - lysine catalytic dyad[23,24]. As expected, preliminary sequence and biochemical analysis found commonalities with S12 family members and PGHs.

Present literature identifies similarities between polyglycine hydrolases and class C β-lactamases/penicillin binding proteins based on primary sequence. Polyglycine hydrolases contain a predicted β-lactamase domain based on their primary sequences and contain the conserved sequence motifs: SVSK and YSN[20,24]. Prominent β-lactamase inhibitors, clavulanic acid and ampicillin were tested for effects on PGH catalysis. These inhibitors had little effect on the PGH activity against *Zea mays* ChitA. The proteolytic activity is well established despite

little being known about the catalytic mechanism of polyglycine hydrolases apart from the identity of the nucleophilic serine[20].

Polyglycine hydrolases target the polyglycine linker on *Zea mays* ChitA chitinase through cleavage of the glycine-glycine bonds[12]. PGHs have been identified to cleave along the polyglycine linker with promiscuous glycine-glycine bond preference, seen in Figure 3. *Bipolaris zeicola* PGH (Bz-cmp) cleaves ChitA after Gly3, Gly4, Gly5, and Gly6 and Es-cmp cleaves after Gly3, Gly4, Gly5, Gly6, Gly7, and Gly8[20]. It's uncertain if the varied protease cleavage products are due to the nature of the enzyme or the nature of the substrate.

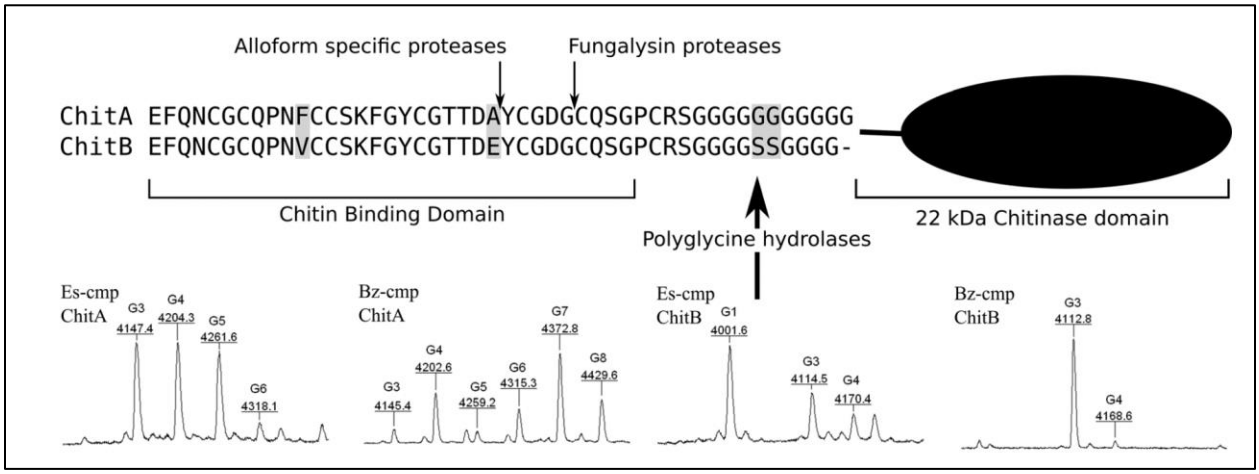


Figure 3. Cleavage activity of polyglycine hydrolases

The polyglycine linker sequence excerpt of *Zea mays* chitinase alloforms, ChitA and ChitB. MALDI-TOF peaks of ChitA and ChitB following incubation of the enzyme with substrate show different sites of cleavage. The shaded sequences indicate the sequence differences between the chitinase alloforms. Figure was adapted from a previous paper by Naumann et al. (2015).

An inhibitor peptide study made three key findings about the binding interaction between PGHs and ChitA: (i) four residues outside of the polyglycine linker are recognized on ChitA and suspected to play a role in substrate orientation, (ii) there are double phenylalanine residues important for substrate binding and (iii) removal of Ser residues from the polyglycine linker

changes the specificity of PGHs[20,25]. The four residues; Ser59, Phe63, Lys67, and Gly71 all reside on a single α-helix in the chitinase domain, adjacent to the polyglycine linker (Figure 4). Structural investigation into polyglycine hydrolases may provide more helpful insight about this group of enzymes and these findings from the peptide work will provide a starting point for the investigation of polyglycine hydrolase - ChitA interaction.

EFQNCGCQPNFCCSKFGYCGTTDDYCGDGCQSGPCRSGGGGGGGGGGG

SGGANVANVVSDAFFNGIKNQAGSGCEGKNFYTRSAFLSAVNKYPGFAHGG

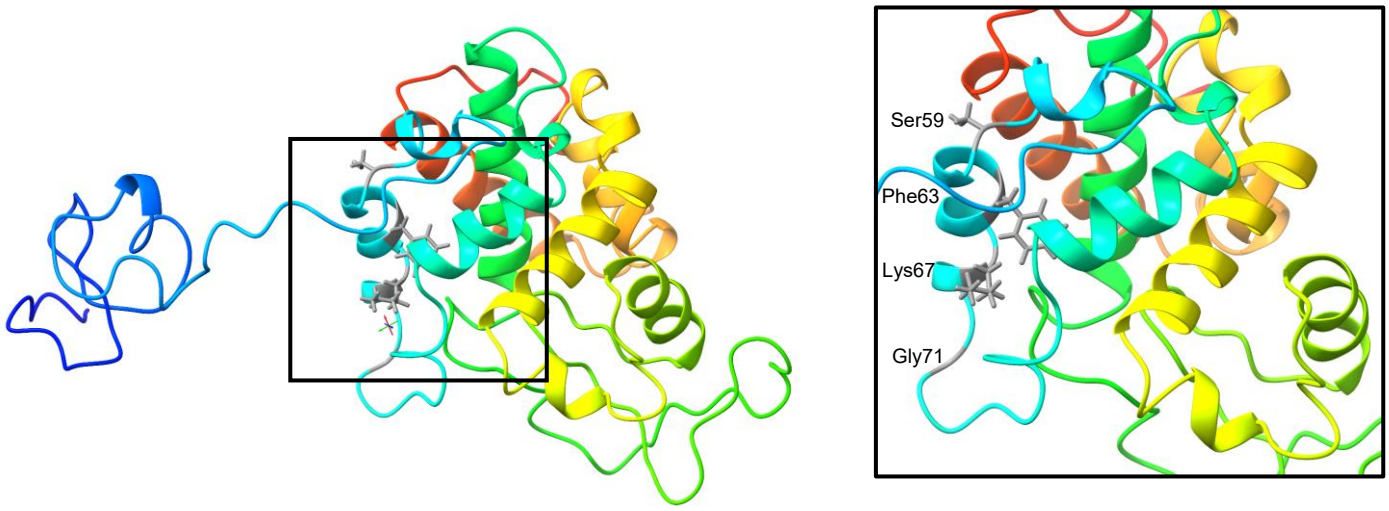
TEVEGKREIAAFFAHVTHETGHFCYISEINKSNAYCDASNRQWPCAAGQKYY

GRGPLQISWNYNYGPAGRDIGFNGLADPNRVAQDAVIAFKTALWFWMNNVH

RLMPQGFGATIRAINGALECNGNNPAQMNARVGYYKQYCQQLRVDPGPNLT

C

A.



B.

Figure 4. ChitA residues implicated in the binding interaction with PGHs

(A) The full-length sequence of recombinant ChitA. The chitinase domain sequence is highlighted by the blue box. The highlighted residues: Ser59, Phe63, Lys67, and Gly71 are implicated in the PGH-ChitA interaction.

(B) The full-length ChitA RoseTTAFold model is show in full colour. Its residues: Ser59, Phe63, Lys67, and Gly71 (grey) are found on the α-helix adjacent to the polyglycine linker and are identified as important residues for proper PGH orientation.

1.1.3 Protein modelling

Protein modelling is the process of generating a three-dimensional rendition of the protein's structure from its primary sequence. There are three main methodological categories: comparative homology modelling, threaded fold-recognition, and *de novo* modelling.

Homology modelling

Comparative homology modelling exploits evolutionary relationships between proteins to build a three-dimensional model of the target protein[26]. There are four conserved steps when generating a homology model for any protein: (i) identification of related structures (templates), (ii) alignment of target sequence onto template structure, (iii) model building and (iv) model evaluation[27]. The first step will often determine the success of the model prediction because this method relies heavily on the presence of similar structures within the Protein Data Bank (PDB). The third step contains the initial model build as well as model optimization[28]. There are many programs that model three-dimensional proteins through this methodology, to name a few: HHpred, Phyre2, Modeller, SWISS-MODEL, Robetta[15,26,29–32].

Like most established methods, the benefits and drawbacks are well-noted. The benefits of this methodology are the model quality, computational efficiency, and the reliance on atomic protein structures. Homology models are only as accurate as their templates are similar. With the plethora of atomic protein structures, homology models are often high quality[33,34]. The use of a template for model generation greatly reduces the computational power compared to non-template-based methods. The major drawback of this methodology is the requirement for similar empirical structures. The reliance on atomic structures is a double-edged sword for this

methodology. It lends sophistication, lessening computational demand and favours biology but reduces success in novel protein structures and discovery-based projects.

Fold-recognition modelling

Fold recognition or commonly referred to as threaded fold recognition modelling is a methodology that utilizes previously solved structures to determine an unknown three-dimensional structure[35]. Similar to comparative homology modelling, this methodology takes advantage of atomic structures previously deposited in the PDB but dissimilarly focuses on identifying known folds rather than a specific evolutionarily related template[35,36]. Fold-recognition works under the assumption that there a limited number of protein folds that comprise all proteins[37]. This methodology has three distinct steps in the process: (i) compilation of structures for search (folds), (ii) model building and quality calculations, and (iii) best model choice. The primary sequence is "threaded" onto the template structures to identify the best-fit for model building[35,38]. There are a few programs that utilize this method in their protein model predictions: IntFOLD, RaptorX, HHpred, Phyre2, and I-TASSER[29,30,36,37,39].

As fold-recognition by threading methodology is a templated-based method, it shares similar benefits and drawbacks with comparative homology modelling. The benefits are relatively accurate model generation, and use of templates with low sequence similarity. Unlike other template-based methods, fold-recognition can predict a protein model from local structure similarities without reliance on global similarity to a template. This ability to exploit local elements expands the scope of templates. The major drawback for fold-recognition modelling is that it will be unable to model novel folds as with any template-based methodology[40].

De novo modelling

De novo protein modelling is a general term defining template-free protein modelling methods. Before *de novo,* the standard - and still common place - term was *ab initio* protein modelling. This predecessor translates to "from the beginning" and generates a three-dimensional protein model based in physicochemical theories. *De novo* methods rely on interatomic interactions that govern protein folding based in energy and entropy principles[41]. In the simplest approach, this modelling method builds a model based on achieving the minimum global free energy [40]. Recognizing that protein domains are continuous, this method begins with identification of protein domain boundaries through sequence-derived secondary structures and physical factors such as solvent accessibility[42–44]. Presently, there are a few programs that model three-dimensional proteins through *de novo* methods: C-QUARK, Rosetta, Touchstone II, AlphaFold, and I-TASSER[43,45–48].

The benefits and drawbacks to this modelling method have been analyzed for decades and continuously evolve with the method. The present benefits of this modelling method are that it has the capacity to predict novel protein folds and doesn't require a structurally similar protein for model generation[40]. *De novo* methods that are advanced through deep- artificial intelligence do rely on previous structures to sophisticate their algorithms but do not rely on a similar structure for a given query[49]. The drawbacks to this method remain as accurate global conformations of proteins and its ability to handle large protein predictions[42,43]. The introduction of machine-learning *de novo* methods have made these drawbacks less evident, but the computational power required for large *de novo* protein predications is still significant.

1.2 Summary

In this thesis, I aim to provide a better understanding of polyglycine hydrolases and their interaction with *Zea mays* ChitA chitinase. Through the integration of new and old modelling and biophysical methodologies I have proceeded to solve and analyze the structure of a novel polyglycine hydrolase by X-ray crystallography facilitated by *de novo* protein modelling (Chapter II), define an interaction model between polyglycine hydrolases and ChitA, and identify and discuss the merits and limitations of newer *in silico* methods within the field of macromolecular structural biology (Chapter III).

Chapter 2 : The crystal structure of a polyglycine hydrolase determined using a RoseTTAFold model

This chapter has been accepted to the journal, International Union of Crystallography (IuCr) Acta D. It is included here with modifications.

2.1 Introduction

The phase problem has traditionally been a major bottleneck during structure solution by X-ray crystallography. In recent years, however, there has been a disruptive advance in available tools within structural biology. Previously, phases were either determined experimentally with multiple diffraction experiments or, more commonly, by molecular replacement of a highly similar experimental structure. Without experimental phases or an adequate structural model, researchers were forced to turn to protein modelling. Prior to the release of RoseTTAFold and AlphaFold, sequence-based protein modelling was quite limited[15,48]. Such modelling relied heavily on sequence similarity of experimentally determined structures. Recent advances in modelling methods have introduced a powerful new option for structural biologists. Novel protein structural analyses with limited similarity to current experimental structures are often no longer stalled by experimental phasing.

Polyglycine hydrolases are secreted fungal proteases that selectively cleave the polyglycine linker that connects the two functional domains of *Zea mays* chitinase, ChitA. Their ability to cleave ChitA was first observed when protein extracts from corn ears rotted by the fungus *Cochliobolus carbonum* (syn. *Bipolaris zeicola*) were found to have altered chitinase activity profiles[50]. Based on the observed activity, the altered chitinase was purified and identified as ChitA[50]. The polyglycine cleaving activity of the fungal protease, named Bz-cmp, was later described[12] and the identity of Bz-cmp was determined, facilitated by development of

next generation sequencing technologies[20]. Analysis of the primary structure of Bz-cmp shows that it consists of an amino-terminal domain of novel sequence and a carboxy-terminal domain that resembles bacterial β-lactamases. Polyglycine hydrolases are part of a larger group of fungal proteases that separate the domains of ChitA and homologous chitinases called chitinase-modifying proteins (cmp). Two other types of cmps, fungalysin metalloproteases[51] and PA domain-containing subtilases named kilbournases[22] have been identified but they do not cleave polyglycine targets.

To date, there are very few examples in nature that describe a polyglycine proteolytic target. In addition to *C. carbonum*, polyglycine hydrolase-encoding genes are present in the genomes of many fungi in the class Dothideomycetes. Es-cmp, from *Epicoccum sorghi*, is the most well-characterized polyglycine hydrolase due to its high level of expression in both fungal cultures and when expressed recombinantly in the yeast *Komagataella phaffii* (syn. *Pichia pastoris*)[12,20]. Polyglycine hydrolase-encoding genes are also present in the genomes of some fungi of the related order Sordariomycetes including *Fusarium vanettenii* (syn. *Fusarium solani* f. sp. *pisi*; syn. *Nectria haematococca*), a plant pathogen that does not infect corn[52]. Interestingly, a few examples of polyglycine hydrolases are also present in the fungal division Basidiomycota, including the mushroom producing wood rot fungus *Galerina marginata*[53]. Despite a preliminary biochemical characterization of Bz-cmp and Es-cmp relatively little is known about these enzymes[12]. The focus of our work is to investigate these novel proteases by structural and biochemical means, to better understand their proteolytic mechanism and other characteristics.

In the present chapter, we discuss the structure of one of these polyglycine hydrolases, from *F. vanettenii.* The structure was solved by molecular replacement, using a RoseTTAFold model[15]. The preliminary structure was determined through MolRep and Buccaneer before being

refined through Refmac[54–56]. The structure solution depicts two distinct domains, referred to throughout as the N- and C- domain. The N-domain exhibits a tertiary fold, previously structurally uncharacterized, with predicted fungal ties. Our analysis shows that this tertiary fold is the first to be reported in an experimentally determined structure. The C-domain resembles a fungal beta-lactamase domain fold, although with proteolytic rather than β-lactamase activity.

2.2 Materials & Methods

2.2.1 Cloning of expression plasmids and integration into *K. phaffii*

Cloning of the Fvan-cmp expression plasmid pTAN163 and integration of linearized plasmid into the genome of *K. phaffii* to create expression strain TAN563 was described previously[57]. The Gm-cmp expression plasmid pTAN170 was cloned in a similar way and integrated into the *K. phaffii* genome to create expression strain TAN423. For cloning, genomic DNA was isolated from *Galerina marginata* CBS 339.88 and used as a PCR template and the two exons of Galma1_254471 were amplified using oligoes KS242 (GAGAGGCTGAAGCTGAATTCTCTCCCACTGACCTTTCTCTCAAAC) and KS243 (CCCCAGACCGCATGCGTATGAATGAAATTCGCCAG), first exon, or KS244 (CATACGCATGCGGTCTGGGGAATAGGTCCTCGTCC) and KS245 (AGATGAGTTTTTGTTCTAGATCAAACAGTGGGATATGCATTCAAG), second exon. Expression plasmids pTAN259, pTAN260, and pTAN261 encoding expression of Fvan-cmp(F543G), Fvan-cmp(R563K/D564T), and Fvan-cmp(F543G/R563K/D564T) were cloned using synthetic DNAs (Integrated DNA technologies, Coralvillle Iowa) to create *K. phaffii* expression strains TAN617, TAN618, and TAN619.

2.2.2 Fvan-cmp purification

Recombinant Fvan-cmp protein was produced by heterologous strains of *K. phaffii* and purified from expression cultures as described previously for Bz-cmp and Es-cmp[20].

2.2.3 Polyglycine hydrolase enzymatic activity

Fvan-cmp and Gm-cmp activity on corn ChitA was tested as detailed previously by adding protease to solutions containing 1 mM ChitA in buffer (10 mM sodium acetate, pH 5.2) followed by incubation at 30 °C for 1 hour prior to analysis by SDS-PAGE or MALDI-TOF MS[20]. The *N*-terminal peptides released by the polyglycine hydrolase proteolytic activity were assayed by matrix-assisted laser desorption/ionization time-of-flight-mass spectrometry, essentially as described previously[20]. The instrument used was a Bruker-Daltonics Microflex LRF (Bruker-Daltonics, Billerica, MA), with a pulsed N2 laser (337 Hz, 60 Hz pulse, 3000 shots), and with reflectron acquisition. The matrix used was 2,5-dihydrobenzoic acid (2,5-DHB). Mass analysis was done using Peptide Mass Calculator v3.2 (http://rna.rega.kuleuven.be/masspec/pepcalc.htm).

Beta-lactamase activity was tested using the colorimetric substrate nitrocefin as described previously[58]. For purified Fvan-cmp, 200 nM enzyme was incubated with substrate for 24 h at 30 °C. For mutants, cell-free media was concentrated 10-fold by ultrafiltration and added at 10% of assay volume.

2.2.4 Crystallization

Fvan-cmp protein was stored in 20 mM Tris-HCl at pH 7.5. Crystals were obtained at 14 °C by the hanging drop vapour diffusion method. The drops were set up using 1 µL of reservoir and 1 µL of Fvan-cmp at 21 mg/mL equilibrated against 500 µL of reservoir solution. Fvan-cmp crystallized in the presence of 0.6 M sodium chloride, 0.1 M MES pH 6.5, and 20% w/v PEG-4000. The protein crystallized in a thick plate morphology clustered from a single nucleation

point after 2-3 weeks. Crystals were cryoprotected in 10% w/v PEG-400, sodium chloride, MES pH 6.5, and PEG-4000 at the previously indicated concentrations.

2.2.5 Data Collection

We were able to collect a few datasets for Fvan-cmp at the University of Waterloo home source diffractometer and remotely at the Canadian Light Source (CLS) CMCF-BM beamline, reported in Table 2. For the structure solution, we used the most complete data set collected at the home source. Data were collected at the home source diffractometer at the University of Waterloo using the Rigaku RUH3R rotating anode and Rigaku RAXIS IV++ detector. Collection occurred at a temperature of 93 K and wavelength of 1.54 Å. Diffraction data were processed with Structure Studio and HKL2000 software[59]. Fvan-cmp protein crystals diffracted to a resolution of 2.2 Å and appear to belong to the $P2_12_12_1$ space group. There was no evidence of oligomerization in solution or in the crystal. Data collection statistics are reported in Table 3.

Table 2. Fvan-cmp Data Sets

Collection Date	Source		
2019-07-12	University of Waterloo	Images	1-361
		Wavelength	1.54
		Distance	135
		Theta	5
2019-08-06	University of Waterloo	Images	1-361
		Wavelength	1.54
		Distance	145
2019-09-10	CMCF-BM	Images	1-720
		Wavelength	0.9190
		Distance	300.00

Table 3. Data Collection, Refinement, and Validation Statistics for Fvan-cmp (PDB ID: 7TPU)

Data collection statistics	
Wavelength (Å)	1.54178
Space group	$P2_12_12_1$
Unit cell (Å)	a= 80.80 b= 94.65 c= 110.48
Unit cell (°)	α= 90.00 β= 90.00 γ= 90.00
Resolution range (Å)	53.76 – 2.19
Completeness (%)	99.1 (99.2)*
R_{merge}	0.36 (1.2)*
Mean I/Sigma (I)	4.63 (1.94)*
CC ½	0.874 (0.567)*
Redundancy	6.1
Wilson B-factor (Å^2)	31.7
Refinement statistics	
No. reflections	43664
R_{work}/R_{free}	0.197, 0.254
Average B-factor (Å^2)	36.0
No. atoms	
Protein	9266
Water	475
RMS Bonds (°)	0.008
RMS Angles (Å)	1.479
Validation statistics	
Ramachandran favored (%)	100
Ramachandran outliers (%)	0
Clashscore	4

*Overall data value (value at the highest resolution shell, 2.24 - 2.19 Å)

2.2.6 RoseTTAFold Model Generation

The full sequence for Fvan-cmp was submitted to the Robetta server for model generation only selecting for the RoseTTAFold modelling method. RoseTTAFold is a fully automated process that combines *de novo* modelling with comparative protein modelling[15]. The output of the server gave five models of the structure. All models ranged from residue 13 to 616, with the first 12 residues remaining unmodelled. We chose to use the first model based off the metrics presented within the interface. The model was truncated including coordinates with less than 3 Å error estimation.

2.2.7 Structure determination and refinement

Phases were not able be obtained experimentally so molecular replacement was conducted on data for Fvan-cmp using the RoseTTAFold model. Molecular replacement was done in MolRep within the CCP4i suite[54,60]. The N-glycans were manually built using the carbohydrate module within Coot[61,62]. The Fvan-cmp structure was refined through successive rounds of Refmac and Coot and its glycans were validated through CCP4 suite's Privateer[56,60,62,63].

2.3 Results

2.3.1 Activity of polyglycine hydrolase homologs

Polyglycine hydrolase cleavage of corn ChitA has previously been demonstrated for Bz-cmp from *C. carbonum* and Es-cmp from *E. sorghi*, two corn pathogens of the fungal class Dothideomycetes[12,20]. To determine if homologous proteins encoded by more distantly related fungi would also cleave the ChitA polyglycine linker, we chose two additional homologs and expressed them recombinantly. We chose Fvan-cmp from *F. vanettenii*, a plant pathogen in the

class Sordariomyctes that does not infect corn, and Gm-cmp from *G. marginata*, a wood rot fungus from the Division Basidiomycota. The level of sequence similarity for each mature protease, compared to Bz-cmp, was determined (Fig. 5A). As expected, proteins from more distantly related fungi had lower identity (ID), lower similarity (Sim), and more gaps (Gap).

Cell-free media from yeast liquid cultures expressing Fvan-cmp and Gm-cmp were observed to truncate ChitA by SDS-PAGE-based protease assays (personal communication with T. Naumann). Fvan-cmp accumulated in the media and was purified following the same procedure used for Bz-cmp and Es-cmp[20]. The amount of Fvan-cmp necessary to convert half of ChitA to the truncated form under standard conditions (E½) was determined to be 8,000 pM, 112-fold and 276-fold greater than reported for Bz-cmp and Es-cmp, respectively[20]. Although activity was observed for Gm-cmp, the protease did not accumulate in the media to a level that could be observed by SDS-PAGE followed by Coomassie staining and we were not able to purify the protease or determine the E½.

To compare the peptide bond selectivity of the different PGHs, we performed MALDI-TOF MS-based protease assays, which allow for visualization of the smaller amino-terminal domain that is released from the larger enzymatic domain upon cleavage of the ChitA polyglycine linker (Fig. 5B). For Bz-cmp, Es-cmp, and Fvan-cmp, reactions were performed with purified proteins under standard conditions and at PGH concentrations matching their respective E½; 71 pM, 29 pM, and 8,000 pM. For Gm-cmp, 1 μL of cell-free media was added per 10 μL of reaction mix, and the incubation time was increased from 1 h to 16 h. MALDI-TOF MS analysis of reaction products confirmed that both Fvan-cmp and Gm-cmp cleave Gly-Gly bonds in the ChitA polyglycine linker (Fig. 5B). Fvan-cmp cleaves preferentially after G1, though products cleaved after G2, G3, G4, G5, and G6 were evident. This selectivity differs from

that of both Bz-cmp and Es-cmp (Fig. 5B)[20]. Gm-cmp cleaved three different peptide bonds with similar frequency, after G3, G4, and G5, similar to the selectivity of Es-cmp.

A.

Polyglycine hydrolases

					Pairwise alignment[1]		
Fungus	Division	Class	Accession[2]	Protein	ID	Sim	Gap
Cochliobolus carbonum	Ascomycota	Dothidiomycetes	XP_007709513.1	Bz-cmp	100	100	0
Epiccocum sorghi	Ascomycota	Dothideomycetes	AIZ65945.1	Es-cmp	58	75	1
Fusarium vanettenii	Ascomycota	Sordariomycetes	XP_003050894.1	Fvan-cmp	46	66	4
Galerina marginata	Basidiomycota	Agaricomycetes	KDR70855.1	Gm-cmp	37	53	6

[1] full-length mature protein compared to Bz-cmp.
[2] NCBI GenBank

B.

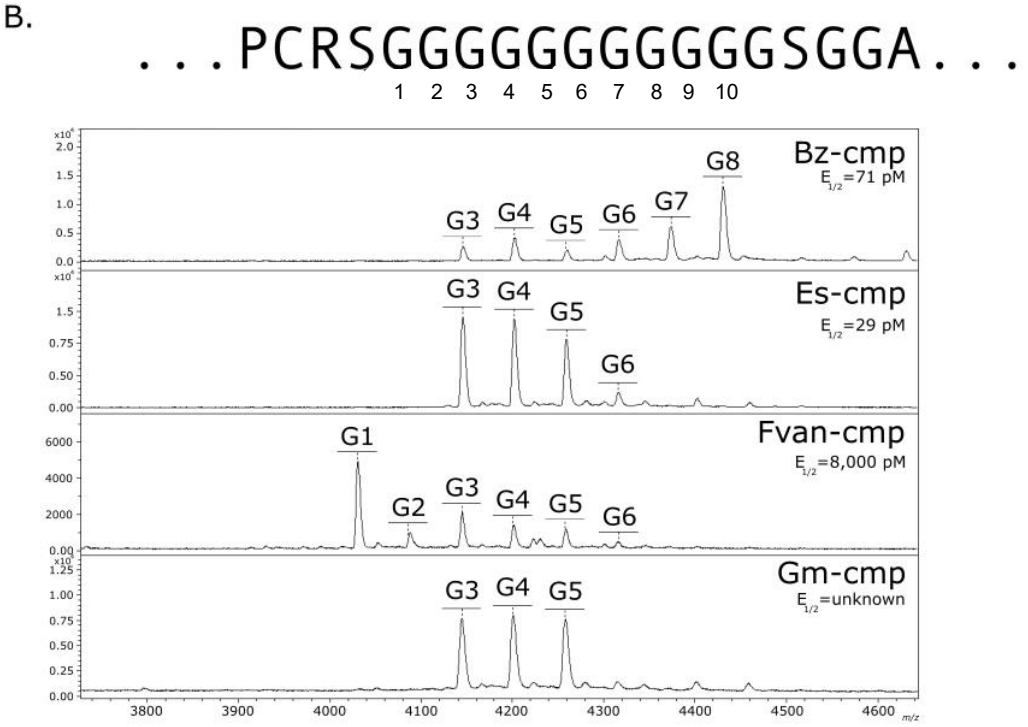


Figure 5. Polyglycine hydrolase homologs.

(A) Comparison of primary structure. The sequence of each mature PGH was compared to that of Bz-cmp. The identity (ID) Similarity (Sim) and Gap percentages (GAP) are summarized. (B) Peptide bond selectivity. Each PGH was incubated with ChitA, followed by MALDI-TOF MS analysis of the amino-terminal reaction products. All products resulted from cleavage of Gly-Gly bonds in the ChitA linker. The sequence of the ChitA polyglycine linker, plus four additional amino acids on each side, is shown above.

2.3.2 Fvan-cmp structure

Of the polyglycine hydrolases discussed above, only Fvan-cmp produced crystals suitable for analysis. The structure of Fvan-cmp was solved to 2.19 Å (PDB ID: 7TPU) by molecular replacement of a RoseTTAFold generated model, as discussed below. Figure 6 illustrates the overall structure of the protein, representing 603 of the 616 amino acid residues in the sequence and two glycosylation sites. The first 12 residues were omitted due to a lack of electron density present in the 2*Fo-Fc* and *Fo-Fc* maps. Fvan-cmp consists of two distinct domains, N- and C-domain that are connected by a linker. Each of these domains will be discussed independently in the following sections.

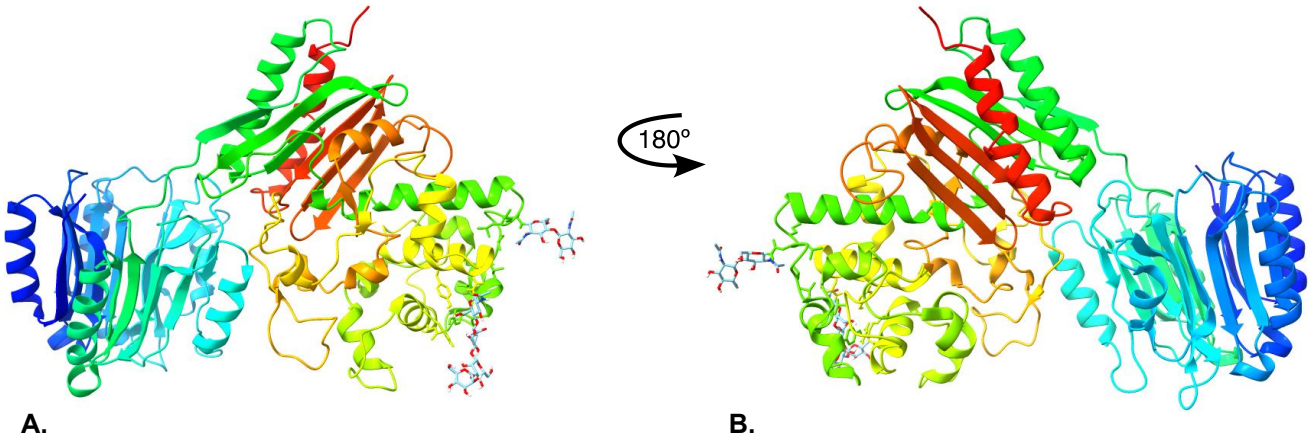


Figure 6. Fvan-cmp structure

(A) Fvan-cmp structure viewed with the N-domain on the left and the C-domain on the right. (B) Fvan-cmp structure rotated 180 degrees compared to (A), which is the orientation to view the active site of the structure. The N-glycosylation sites are represented in stick form, both on the C-domain.

2.3.3 Undefined electron density surrounding catalytic serine

The Fvan-cmp structure solution presented several challenges from solving the phase problem to discerning the structure from the imperfect crystal data. One of the anomalies from

Fvan-cmp crystals was the presence of additional electron density surround the catalytic serine residue. Upon the first inspection, we evaluated if this electron density could be explained from (i) expression and purification methods (ii) crystallization conditions.

The expression and purification methods previously described do not indicate anything that would cause modification of the catalytic serine residue. Table 4 summarizes the purification process for polyglycine hydrolases[20]. These reagents do not have the capacity to covalently modify the serine residue to account for the additional electron density. Acetate was a contender for solving the unknown electron density as it 'matched' one of the data density maps however when reflecting on the expression and purification process is not available to interact with the serine in question.

Table 4. Standard purification interventions for PGHs

Process	Chemical Reagents
Centrifugation • Removed cells	
Precipitation • Remove media from secreted protein	608 g/L ratio ammonium sulfate
Resuspension & Dialysis • Remove residual ammonium sulfate	50 mM sodium acetate pH 4.7, 100 mM NaCl
Mixed-mode cation exchange chromatography	Cation exchange buffer: 100 mM MES pH 6.0, 1.0 M NaCl
Precipitation • Concentrate protein for analysis or storage	50% v/v acetone
Resuspension	20 mM Tris-Cl pH 7.5, 500 mM NaCl

The standard protocol for purification of polyglycine hydrolases after expression. The protocol steps (left column) have a brief justification and any chemical reagents used within the step are detailed (right column).

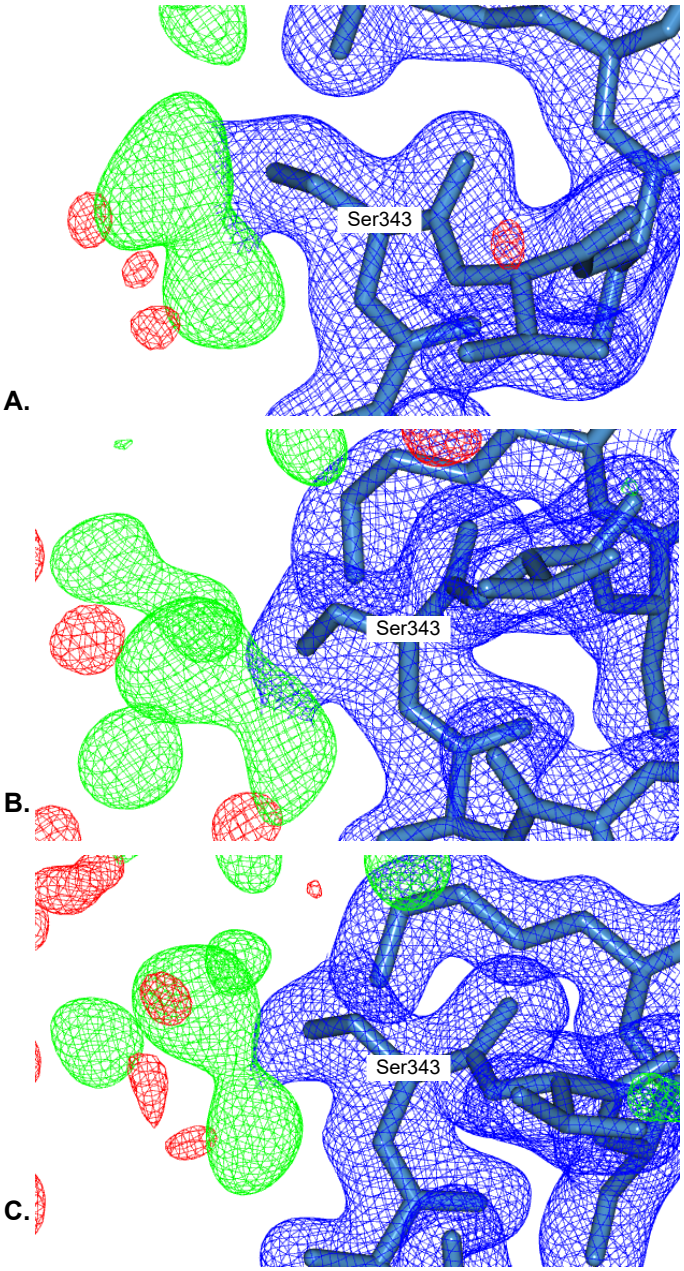


Figure 7. Unexpected electron density surrounding the catalytic serine

The images were generated in CCP4i suite's CCP4MG program with contour levels of 1.00 sigma for 2*Fo-Fc* and +/- 3.00 sigma for *Fo-Fc* maps. Green represents positive density and red represents negative density in the *Fo-Fc* map. The data sets were (A) Fvan-cmp data from the deposited structure (B) 2019-08-06 data (C) 2019-09-10 data. Orientation of the serine varied to better visualize the density surrounding the residue.

The crystallization process was relatively straightforward with this protein consisting only of the crystallization reagents and a homemade cryoprotectant cocktail. The crystallization and mounting processes are detailed in the Materials & Methods Section 2.2.4, but it is known that the only reagents used were MES, NaCl, PEG-400, and PEG-4000. Systematically, we reviewed each reagent for its potential to occupy this electron density. First, MES was ruled out owing to the size of MES compared to the available electron density. Second, it would be facile to assume the presence of a Na^+ ion but unfortunately the coordination (6) needs of this ion would be unfulfilled in this position. Lastly, polyethylene glycol is a contender as it is a polydisperse reagent consisting of a range of molecular weight. It is reasonable to suggest that the bound nature of the electron density in Figure 7A, is PEG of a low molecular weight. PEG is a dynamic reagent as it is weight-dependently hydrophilic, flexible and has the capacity to interact covalently and non-covalently[64]. These characteristics, some of which are common with the polyglycine substrate, make it a reasonable candidate for occupancy proximal to the serine residue. The nature of the substrate and the characteristics of the active site are discussed at length in Chapter 3.

2.3.4 N-domain

The Fvan-cmp N-domain (residues 13-262) consists of 4 loops, 5 α-helices, 15 β-strands assembled into a distinct tertiary fold. This distinct structure, shown in Figure 8 is comprised of five quasi-identical structural repeats (Fig. 8B) consisting of 3 β-sheets and an α-helix arranged as EHEE with beta-strands in an antiparallel assembly. Each repeat spans 44 amino acid residues with a 5-6 residue loop connecting them, provided in Figure 9. These repeats are defined as structural repeats as there appears to be limited sequence conservation between the regions. When in the tertiary structure, these five regions arrange into a barrel-like structure.

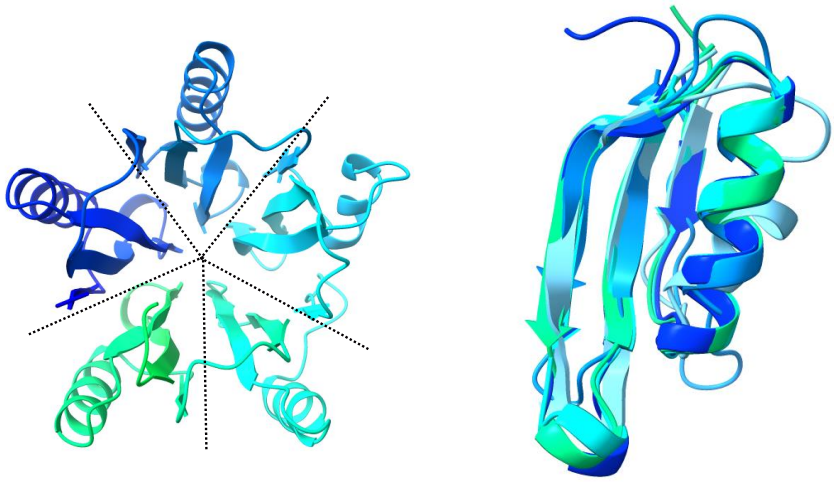


Figure 8. Fvan-cmp structural repeats

(A) Fvan-cmp structure in a top-down view of the N-domain. The five quasi-identical structural repeats that compose this domain are segregated visually by dotted lines.
(B) The structural superposition of the repeats aligned by their αCarbons.

When the structure was first solved, we found that the tertiary structure did not coincide with any known αβ barrel folds but was identified as a novel superfamily in an analysis of AlphaFold's database[65]. To investigate this, we conducted a search within two web servers, DALI protein structure comparison server and FoldSeek[66,67]. Within the DALI search, we evaluated tertiary fold likeness by the assigned Z-score metric. The Z-score is the similarity score between the query structure and its matches; strong matches have Z-scores higher than the assigned cut-off[68]. The Z-score cut-off is calculated based on the number of residues for the input query[68]. For the N-domain, the assigned Z-score cut-off was 24 and the closest match within the server had a Z-score of 4.5. Further investigation of the top hits revealed that there was no full match for the structural repeat, nor the tertiary structure described. We ran a search through the FoldSeek webserver against all currently available databases and found a similar but interesting

result. As with DALI, there was no experimentally determined structure resembling the N-domain tertiary fold. However, FoldSeek did identify similar predicted structures within the AlphaFold Protein Structure Database. To date, none of these identified proteins have been functionally characterized.

A

EFLPNQRRSNVTSHVETYYSVDGATHAEKSKALKADGYRIVSLSSYGSPDSANYAAIWVQEEGPSFEII
HDADEATYNSWLQTWKSRGYVSTQVSATGPAENAVFAGVMENINVANWFQSCELENPWAFSNTTG
NVDVVVKGFRMFGTPEERRYCILGHENVGNEQTTIQYSTPSFTVNFASTFEAETTKRFWRPSRLFLSE
DHIITPSFADTSVGKWSHAVDLTKAELKEKIETERAKGLYPIDIQGGGSGSSERFTVVFA**ERTSPKPR**Q
WNVRGEITGFEDNKAAEEEVDSIMRRFMEKNGVRQAQFAVALEGKTIAERSYTWAEDDRAIVEPDDIF
LLASVSKMFLHASIDWLVSHDMLNFSTPVYDLLGYKPADSRANDINVQHLLDHSAGYDRSMSGDPSF
MFREIAQSLPTKGAKAATLRDVIEYVVAKPLDFTPGDYSAYSNYCPMLLSYVVTNITGVPYLDFLEKNIL
DGLNVRLYETAASKHTEDRIVQESKNTGQDPVHPQSAKLVPGPHGGDGAVKEECAGTFAMAASASSSL
AKFIGSHAVWGTGGRVSSNRDGSLSGARAYVESRGTIDWALTLNTREYISETEFDELRWYSLPDFLSA
FPIAG

B

EFLPNQRRSN
V15TSHVETYYSVDGATHAEKSKALKADGYRIVSLSSYGSPDSANYAAIWVQ60
EEGPS
F66EIIHDADEATYNSWLQTWKSRGYVSTQVSATGPAENAVFAGVME110
NINVA
N116WFQSCELENPWAFSNTTGNVDVVVKGFRMFGTPEERRYCILGHE160
NVGNEQ
T167TIQYSTPSFTVNFASTFEAETTKRFWRPSRLFLSEDHIITPSFA211
DTSVGK
W218SHAVDLTKAELKEKIETERAKGLYPIDIQGGGSGSSERFTVVFA262
ERTSPKPRQWNVRGEITGFEDNKAAEEEVDSIMRRFMEKNGVRQAQFAVALEGKTIAERSYTWAEDD
RAIVEPDDIFLLASVSKMFLHASIDWLVSHDMLNFSTPVYDLLGYKPADSRANDINVQHLLDHSAGYDR
SMSGDPSFMFREIAQSLPTKGAKAATLRDVIEYVVAKPLDFTPGDYSAYSNYCPMLLSYVVTNITGVPY
LDFLEKNILDGLNVRLYETAASKHTEDRIVQESKNGQDPVHPQSAKLVPGPHGGDGAVKEECAGTFA
MAASASSLAKFIGSHAVWGTGGRVSSNRDGSLSGARAYVESRGTIDWALTLNTREYISETEFDELRW
YSLPDFLSAFPIAG

Figure 9. Full-length sequence Fvan-cmp

(A) The full-length sequence of Fvan-cmp protease. The N-domain spans residues 1-282, C-domain spans 271-616 with the linker region spanning residues 263-270. The linker region is bolded for visualization within the sequence.
(B) The N-domain consists of five structural repeats connected by a small loop. The EHEE repeats are indicated by colour VAL15 - GLN60 (1), PHE66 - GLU110 (2), ASN116 -GLU160 (3), THR167 - ALA211 (4) and TRP218 - ALA262 (5).

The novelty of the N-domain explains the difficulties during the structure solution process. The sequence search within the protein data bank (rcsb.org) did not identify an adequate model for Molecular Replacement[69]. Traditional automated modelling servers all failed to generate a full-length model. The partial-coverage models failed in the Molecular Replacement pipeline.

Recently, with the release of RoseTTAFold from the Baker lab, we were able to obtain a full-length sequence model owing to the sophistication of RoseTTAFold's deep-learning processing[15]. The Robetta server (https://robetta.bakerlab.org) outputs the top five models from the run. Observing the per-residue error plot, we trimmed our model coordinates to those residues with a predicted error of less than 3 Å. We used the trimmed RoseTTAFold model to solve the structure by Molecular Replacement.

The accuracy of the secondary structures within the Fvan-cmp structure from RoseTTAFold is remarkable. A simple backbone alignment of the error truncated RoseTTAFold model and the final structure had a final Cα-rmsd of 2.76 Å. This method is not reliable for determining side-chain orientation nor capable of determining post-translational modifications but can be used as a powerful tool in conjunction with experimental data.

2.3.5 C-domain

The Fvan-cmp C-domain (residues 271-616) consists of 7 α-helices, and 1 antiparallel β-sheet, and resembles a beta-lactamase fold. A DALI search against all structures within the protein data bank yielded high Z-scores with penicillin-binding proteins and β-lactamases. A structural alignment of the Fvan-cmp C- domain against a penicillin-binding protein (PDB ID: 2QMI) and a AmpC β-lactamase (PDB ID: 4GZB) showed a strong similarity to the β-lactamase fold[70,71]. We quantified similarity by observing and comparing rmsd values provided through

Chimera Matchmaker[72]. The structural alignment of Fvan-cmp C- domain and penicillin-binding protein (PDB ID: 2QMI) had a Cα-rmsd of 0.957 Å between the pruned atom pairs and 9.741 Å across all atom pairs. The structural alignment of Fvan-cmp C- domain and AmpC β-lactamase (PDB ID: 4GZB) had a Cα-rmsd of 0.995 Å between the pruned atom pairs and 6.978 Å across all atom pairs. In Figure 10A, we show the structural alignments of the β-lactamase domains from the three proteins: Fvan-cmp, penicillin-binding protein and AmpC β-lactamase. The αβα folds are conserved in global positioning between the three proteins.

Within the β-lactamase fold, there are three conserved sequence motifs seen within penicillin-binding proteins and multiple classes of β-lactamases. Two of the three sequences (Table 5) are observed in Fvan-cmp, aligning with Class C beta-lactamases. β-Lactamases inactivate β-lactam antibiotics, such as penicillins, cephalosporins, and carbapenems, rendering them inactive, and are an important mechanism of bacterial antibiotic resistance. The class C β-lactamases are found solely in Gram-negative bacteria and the mechanism by which they hydrolyze β-lactam antibiotics is still incompletely understood[73].

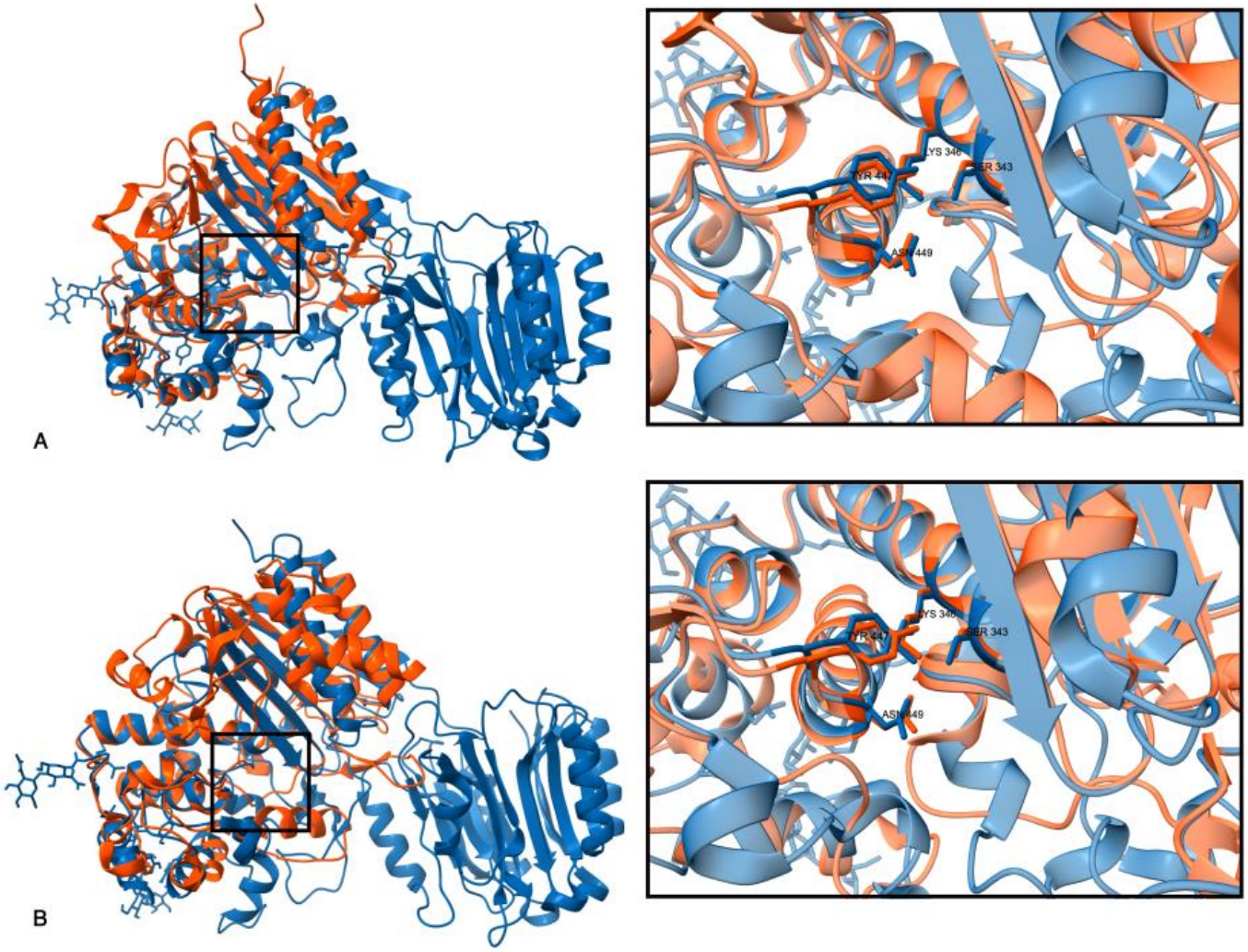


Figure 10. Structural alignment of a penicillin-binding protein (PDB: 2QMI), a β-lactamase (PDB: 4GZB) and a polyglycine hydrolase, Fvan-cmp (PDB: 7TPU).

The global structural alignment of 7TPU and 2QMI (A) and 4GZB (B). For each alignment, there is a focused view of the active site secondary structures and their global arrangement. Residues 276-297 and 365-447 were omitted for better visualization of the active site. The two conserved β-lactamase motifs within Fvan-cmp are labelled and in stick representation: S343, K346, Y447, N449. The corresponding residues are represented in stick form for the penicillin-binding protein and β-lactamase.

Table 5. Penicillin-binding protein and beta-lactamase conserved sequence motifs

	Conserved sequences	Location within fold	Fvan-cmp sequence
1	S - X - X - K	α-helix (H11)	S_{343} - V - S - K_{346}
2	Y/S - X - N*	active site facing loop, before α-helix (H20)	Y_{447} - S - N_{449}
3	K - T - G	Terminal β-strand on β-sheet (E6)	n/a

*Class A beta-lactamases and penicillin-binding proteins have a serine while Class C beta-lactamases have a tyrosine within the first position of this motif.

The conserved sequence motifs found within the αβα beta-lactamase fold occur in penicillin-binding proteins and multiple classes of beta-lactamases. The motifs occur in different secondary structures in the same relative positioning across different proteins. Fvan-cmp shares two of the three conserved motifs but lacks the third motif. The residues are identified by their sequence for clarity.

These motifs were previously determined to play a prominent role in substrate orientation and catalysis in β-lactamases[24]. The first motif contains the nucleophile used during enzyme catalysis. Fvan-cmp shares the same nucleophilic serine found within this motif. The second and third motif are involved in substrate positioning in penicillin-binding proteins[74]. Within the third motif, the glycine in the third position is important in preventing steric interference within substrate binding[74]. Fvan-cmp lacks this glycine and instead contains a phenylalanine in the same position. Figure 10B shows a comparison of the active sites of Fvan-cmp to a reference penicillin-binding protein (PDB ID: 2QMI) and a class C β-lactamase (PDB ID: 4GZB). Despite the structural similarities between the chitinase-modifying proteins and penicillin-binding proteins/β-lactamases there are critical differences that have a large effect on enzymatic function.

2.3.6 Beta-lactamase activity

As discussed previously, Fvan-cmp contains two of the three conserved sequence motifs found within penicillin-binding proteins and beta-lactamases. Noting this, previous work tested for β-lactam binding and β-lactamase activity with two different polyglycine hydrolases, Bz-cmp

and Es-cmp[20]. The potential β-lactamase activity was tested on nitrocefin, a colorimetric substrate, but neither showed activity. Also, the β-lactamase inhibitor clavulanic acid was added to protease reactions containing Bz-cmp or Es-cmp but inhibition of proteolysis on ChitA was not observed.

As Es-cmp did not exhibit β-lactam binding nor β-lactamase activity and in view of its structural similarity to Fvan-cmp, we attempted to introduce β-lactamase activity through site-directed mutagenesis. Specifically, we reduced the proposed steric hindrance to the active site of Fvan-cmp as a single mutant (F534G) and restored the third conserved sequence motif as a double mutant (R563K/D564T). A triple mutant was also constructed (F534G/R563K/D564T). Expression of the mutants was greatly reduced, compared to the wild-type Fvan-cmp, as noted by SDS-PAGE analysis of cell-free media after induction (Fig. 19, Appendix I). Despite the low level of protein that accumulated, purification of the single and double mutants was attempted, but resulted in loss of protein, indicating that they are likely misfolded. Utilizing a nitrocefin assay, we did not observe β-lactamase activity from the cell-free media of either Fvan-cmp or the single, double, or triple mutants. Purified Fvan-cmp also lacked β-lactamase activity as reported for Bz-cmp and Es-cmp.

2.3.7 Identifying the catalytic dyad and oxyanion hole

As discussed within Chapter 1 and earlier in Chapter 2, polyglycine hydrolases are serine proteases that belong to the same family as D-alanyl-D-alanine carboxypeptidases and β-lactamases[20,23]. Upon analysis of the atomic structure of Fvan-cmp, we identified (i) the catalytic dyad for Fvan-cmp, inferring the catalytic dyads in other PGHs and (ii) propose the oxyanion hole forming residues.

We identified all lysines and histidines within the C-domain of the atomic structure to visualize our possible catalytic base residue. We identified three residues visually proximal to the nucleophilic Ser343 within the catalytic site: Lys346, His350, and His392. Closer inspection identified Lys346 to be the only residue capable of activating Ser343 during catalysis as the histidine residues were at a distance incapable of hydrogen bonding. To confirm our catalytic dyad suspicions, we did search for appropriately distanced aspartate and glutamate residues that would be capable of interacting with Lys346. We were unable to identify such a residue to make the Fvan-cmp catalytic triad, strengthening our hypothesis of a Ser-Lys catalytic dyad.

Table 6. Important residues in PGH catalysis

Protein	Catalytic Dyad	Oxyanion coordinating residues
Bz-cmp	Ser369	Gly591
	Lys372	Thr592
Es-cmp	Ser49	Gly571
	Lys352	Thr572
Fvan-cmp	Ser343	Gly565
	Lys346	Ser566

The catalytic dyad residues and oxyanion hole coordinating residue for Fvan-cmp were identified from its atomic structure. The identified residues for Bz-cmp and Es-cmp were inferred from their models and similarity to the Fvan-cmp structure.

The oxyanion hole is important for stabilizing the transition state[75–77]. This feature is found in several types of enzymes and is extremely important for serine proteases[77,78]. We identified our coordinating oxyanion hole residues using our unbound atomic structure and comparing a β-lactamase in a tetrahedral transition state (PDB ID: 1BLH)[79]. Based on the structure similarities PGHs share with β-lactamases, we anticipated their help identifying the oxyanion hole residues. By comparing the positioning of the Ser-Lys dyad position and analyzing the surrounding residue distances we anticipate the coordinating residues that form the oxyanion hole are Gly565 and Ser566 (Figure 11). These residues are within hydrogen bonding

distance of Ser343 oxygen and are characteristically located above the nucleophilic Ser in the active site.

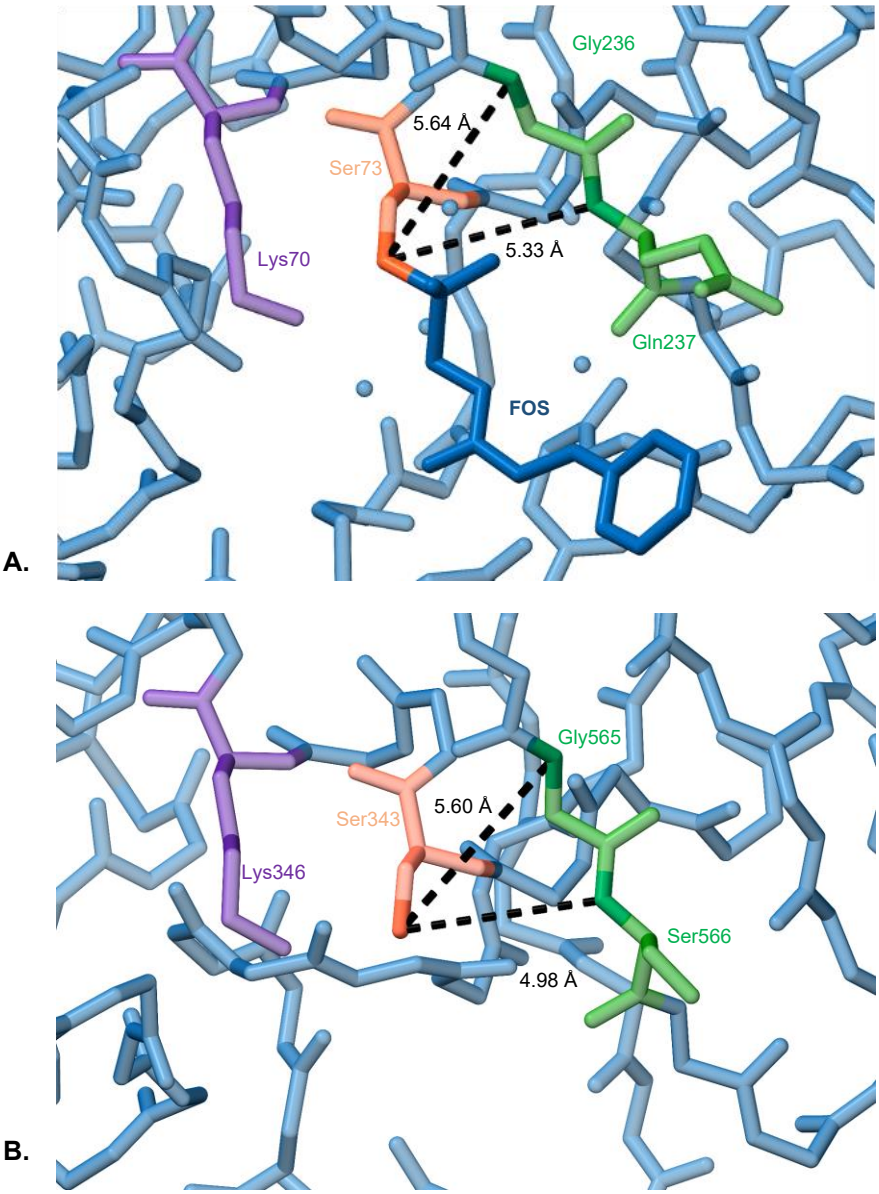


Figure 11. Oxyanion hole observation in Fvan-cmp and a β-lactamase

(A) The atomic structure for a β-lactamase (PDB ID: 1BLH)[79]. The catalytic dyad is labeled, Ser70 (orange) and Lys73 (purple). The coordinating residues for the oxyanion hole are Gly236 and Glu237 (green). The calculated distances are from the oxygen of the nucleophilic serine to the respective nitrogens on the coordinating residues.
(B) The atomic structure of Fvan-cmp (PDB ID: 7TPU). The catalytic dyad is labeled, Ser343 (orange) and Lys346 (purple). In the same orientation as above, we anticipate the coordinating

residues for the oxyanion hole to be Gly565 and Ser566 (green). The calculated distances are from the oxygen of the nucleophilic serine to the respective nitrogens on the coordinating residues. This figure was prepared in ChimeraX-1.4.

2.4 Discussion

2.4.1 Novel N-domain tertiary fold

In our FoldSeek search we came across several predicted proteins within the AlphaFold Structure Database that shared the tertiary fold of the N-terminal domain. All these proteins exhibit the lack of sequence conservation between the individual structural repeats that we observed in Fvan-cmp. The proteins (an abbreviated list in Table 7) are diverse in origin, spanning across all kingdoms, with the majority found in bacteria. These proteins vary in the level of functional characterization however, they share a lack of functional descriptors for the tertiary fold described. In the literature, there is speculation that this N-domain might play a role in substrate positioning and/or exo-site binding of ChitA and ChitB[20,25]. The level of conservation of this domain in all kingdoms suggests a more general function that is not specific to polyglycine hydrolases. The potential for this domain to be involved in protein-protein interactions (PPI) is possible due to its similarities with WD domains. WD domains are the most abundant protein interaction domain in current literature[80]. They are defined by their conserved WD motif and distinct β-propellor 'doughnut' shape[80]. Examples might include a chaperone activity, involved in the folding or stability of the rest of the protein, or a role in transporting or anchoring to ensure localization of the protein to a specific target. Our result opens an area of future work, which will focus on determining the biological function of this tertiary fold and its importance across the kingdoms.

Table 7. Top 50 FoldSeek hits against the AlphaFold Uniprot Database

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-A0A166M196	59.1	1480	1.83E-32	*Colletotrichum incanum*	Penicillin-binding protein
AF-A0A135SQZ1	54.6	1462	4.87E-32	*Colletotrichum salicis*	Pectate lyase
AF-A0A1B7YR40	55.3	1438	1.80E-31	*Colletotrichum higginsianum IMI 349063*	Penicillin-binding protein
AF-N4V2J0	56.1	1400	1.42E-30	*Colletotrichum orbiculare MAFF 240422*	Beta-lactamase containing domain
AF-A0A2V1DDC1	53.7	1361	1.18E-29	*Periconia macrospinosa*	Beta-lactamase/transpeptidase-like protein
AF-A0A4R8RCR4	48.6	1230	1.45E-26	*Colletotrichum trifolii*	flp gene product
AF-A0A6A5RWU7	44.9	1227	1.71E-26	*Didymella exigua CBS 183.55*	Beta-lactamase/transpeptidase-like protein
AF-A0A3M7MH02	43.3	1197	8.73E-26	*Pyrenophora seminiperda CCB06*	Penicillin-binding protein
AF-A0A4Q6A4M9	29.8	1092	2.62E-23	*Sphingobacteriales bacterium*	Class A beta-lactamase related serine hydrolase
AF-A0A067T494	36.7	1086	3.63E-23	*Galerina marginata CBS 339.88*	Uncharacterized
AF-A0A1J9QVD6	37.7	1081	4.77E-23	*Diplodia corticola*	Beta-lactamase containing domain
AF-A0A321LBL7	33.4	1067	1.02E-22	*Blastocatellia bacterium AA13*	Beta-lactamase containing domain
AF-A0A6A6B650	35.6	1067	1.02E-22	*Aplosporella prunicola CBS 121167*	Beta-lactamase containing domain

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-K3V3V4	33.6	1065	1.14E-22	*Fusarium pseudograminearum CS3096*	Beta-lactamase containing domain
AF-A0A409Y2P4	37.7	1044	3.56E-22	*Gymnopilus dilepis*	Beta-lactamase containing domain
AF-A0A067T7P9	30.9	1026	9.46E-22	*Galerina marginata CBS 339.88*	Uncharacterized
AF-A0A1L7TAS3	41.9	972	1.78E-20	*Fusarium mangiferae*	Beta-lactamase containing domain
AF-A0A3M9ZDX4	27.1	965	2.60E-20	*Leptolyngbya sp. IPPAS B-1204*	Class A beta-lactamase related serine hydrolase
AF-A0A849TPW4	23.7	926	2.17E-19	*Nitrospira sp.*	Uncharacterized
AF-A0A7W7CQE2	29.8	925	2.29E-19	*Actinoplanes abujensis*	Uncharacterized
AF-A0A532CUY9	24.9	912	4.64E-19	*Nitrospira sp.*	Uncharacterized
AF-A0A7W1TEK5	29.3	910	5.17E-19	*Planctomycetes bacterium*	Serine hydrolase
AF-A0A7W0KYF8	29.4	905	6.78E-19	*Acidimicrobiia bacterium*	Serine hydrolase
AF-A0A838DV87	25.6	877	3.10E-18	*Ktedonobacteraceae bacterium*	Serine hydrolase
AF-A0A6H9YRX4	24.6	873	3.86E-18	*Actinomadura rudentiformis*	Uncharacterized
AF-A0A7G5IK55	22.1	872	4.07E-18	*Sandaracinobacter sp. M6*	Serine hydrolase
AF-A0A5J6P455	25.8	869	4.80E-18	*Cellvibrio sp. KY-GH-1*	Class A beta-lactamase related serine hydrolase
AF-A0A3N1JPB1	23.1	843	1.97E-17	*Granulicella sp. GAS466*	Beta-lactamase
AF-A0A2V8HSC7	25.6	823	5.84E-17	*Acidobacteria bacterium*	Beta-lactamase containing domain
AF-A0A3A4B505	25.6	794	2.82E-16	*Bailinhaonella thermotolerans*	Non-specific serine/threonine protein kinase
AF-A0A2J6QYV7	18.9	784	4.86E-16	*Hyaloscypha variabilis F*	Beta-lactamase containing domain
AF-A0A2L2U0F8	29	752	2.76E-15	*Fusarium venenatum*	Beta-lactamase containing domain

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-D2B7Z4	19.7	746	3.83E-15	*Streptosporangium roseum DSM 43021*	Beta-lactamase
AF-A0A3N7JU83	21.1	743	4.51E-15	*Albitalea terrae*	Class A beta-lactamase related serine hydrolase
AF-A0A7Y6IQ54	22.1	733	7.76E-15	*Nonomuraea rhodomycinica*	Uncharacterized
AF-A0A1H1BZ32	22.5	724	1.27E-14	*Thermostaphylospora chromogena*	Uncharacterized
AF-A0A7X0U1N1	17.8	723	1.34E-14	*Nonomuraea rubra*	Uncharacterized
AF-A0A2J6SIY4	33.3	719	1.66E-14	*Hyaloscypha bicolor E*	Beta-lactamase/transpeptidase-like protein
AF-A0A367FK14	20.8	716	1.96E-14	*Sphaerisporangium album*	Uncharacterized
AF-A0A7W8ECQ3	22.5	696	5.80E-14	*Nonomuraea endophytica*	Non-specific serine/threonine protein kinase
AF-A0A5R8MSZ6	18.8	695	6.12E-14	*Nonomuraea sp. KC401*	Uncharacterized
AF-A0A5S4FIC2	20.1	694	6.46E-14	*Nonomuraea turkmeniaca*	Uncharacterized
AF-A0A848DM22	15	691	7.60E-14	*Pseudonocardia bannensis*	Beta-lactamase containing domain
AF-A0A0J9ECQ5	19.3	683	1.17E-13	*Candidatus Rhodobacter lobularis*	Beta-lactamase containing domain
AF-A0A553Y292	16.5	676	1.72E-13	*Streptomyces benahoarensis*	Beta-lactamase containing domain
AF-A0A2H3RIQ8	35.4	672	2.14E-13	*Fusarium fujikuroi*	Beta-lactamase containing domain
AF-A0A124DZT4	22.9	659	4.33E-13	*Mycolicibacterium brisbanense*	Beta-lactamase containing domain
AF-A0A239BKE8	20.3	646	8.77E-13	*Streptosporangium subroseum*	Uncharacterized
AF-A0A4Q5NQF3	19.2	626	2.60E-12	*bacterium*	Beta-lactamase containing domain

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-A0A022VXE6	14	624	2.90E-12	*Trichophyton rubrum CBS 288.86*	Uncharacterized

The matches are sorted based on their FoldSeek score metric. All proteins that matched the Fvan-cmp N-domain, none characterized the purpose of the tertiary fold. Within the top 50 results, 20 proteins were from the fungal domain and 30 proteins were from the bacterial domain.

2.4.2 Weak binding of PEG contributes to confusing electron density inconsistencies

The mystery of the electron density surrounding the catalytic serine remains largely speculative owing to constraints of the project. The three data sets collected on Fvan-cmp crystals all show electron density surrounding the serine residue. However, the overall shape of the density is inconsistent - potentially amplified by the varying resolution cut-offs. We propose that the electron density present across all data arise from crystallization conditions. In the deposited data set (Figure 7A), the electron density could be PEG non-covalently bound whereas in the other two data sets (Fig. 7B/7C) it is reasonable to suggest highly ordered water molecules. We speculate associated water molecules due to the individual globular nature of the density compared to the other electron density.

2.4.3 Polyglycine hydrolases & their relationship with lactamases

This chapter highlighted the similarities between the representative polyglycine hydrolase (Fvan-cmp), penicillin-binding proteins and Class C β-lactamases. We showed that Fvan-cmp retains two of the three conserved β-lactamase motifs and the core active site αβα fold but lacks the associated activity. It is reasonable to suggest that polyglycine hydrolases share a common ancestor protein with β-lactamases, as do the β-lactamases and penicillin-binding proteins. Fungal lactamases have already been previously described in the literature but lack the extensive characterization afforded to bacterial β-lactamases[81].

Focusing on the residue similarities between β-lactamases and polyglycine hydrolases, we observed two important features. First, in addition to the retained catalytic motifs, PGHs contain an analog of the AmpC β-lactamase Y150 (Y447 in Fvan-cmp) residue. This residue is an important distinction between the different classes of β-lactamases and integral to the kinetic functioning of β-lactamases[82]. Second, polyglycine hydrolases shared conserved residues with other classes of β-lactamases. A recent study on class A β-lactamases categorized conserved residues into 'shells'. These shells can be defined by proximity to the active site and function[83]. The conserved residues are implicated in the folding, stability, and function of the protein. We found that the polyglycine hydrolases retained several of these residues shown in Table 13 (Appendix I).

The point mutagenesis and structural studies demonstrate that, if the protein is properly folded in the cell-free media, the absence of β-lactamase activity could be due to regions outside the catalytic center (refer to Fig. 19, Appendix I). The Fvan-cmp active site and surface map (data not shown) depicts a region that is sterically limited. It may be that the flexibility of the polyglycine peptides requires that they be constricted into a narrow binding region in these hydrolases, a region that is incompatible with a bulkier lactam ring.

2.4.4 Application of new tools in structural science

RoseTTAFold and AlphaFold have changed the field of structural biology. Before these methods, sequence-based structure predictions were not accurate without having experimental templates. The accuracy of predictions has much improved owing to RoseTTAFold and/or AlphaFold. Accompanying searches for structurally similar proteins using the DALI server or FoldSeek has expanded available resources to learn about a given protein.

The work described herein demonstrates both the power and limitations of these new tools. While the pipeline was critical to the structure determination of Fvan-cmp, there are still questions about the differing specificity and activity of the fungal polyglycine hydrolases that can only be addressed through experimentation. Nevertheless, the insights gained, and hypotheses formed by these results are an exciting advance for this family of proteins.

Chapter 3 : Modelling the polyglycine hydrolase and ChitA interaction and analysis of current prediction methods

3.1 Introduction

While computational approaches have always played a role in macromolecular structure analysis, until recently, the field has relied mostly on traditional biophysical techniques, such as X-ray crystallography, nuclear magnetic resonance spectroscopy and, more recently, cryoelectron microscopy, for definitive results. The goals of this chapter are: (i) to probe the effectiveness of new prediction algorithms in forming realistic hypotheses of protein complexes and (ii) to generate hypotheses for the interaction of the polyglycine hydrolases and *Zea mays* chitinase alloform, ChitA that are consistent with experimental data. Such hypotheses further our understanding of this resistance mechanism and lead to potential, testable interventions applicable to the agriculture industry.

In silico structure modelling has made massive strides in the past five years. Prior to the release of AlphaFold (now AlphaFold2) and RoseTTAFold, structure prediction without a model was inaccurate and computationally intensive[15,49]. As single structure prediction has progressed, improvements in protein-protein docking and/or modelling are following. There is a myriad of bioinformatic methods available - each with their own benefits and drawbacks. We used two methods within this project: HADDOCK v2.4 and AlphaFold Multimer[84,85].

High Ambiguity Driven protein-protein DOCKing (HADDOCK) can be run locally on Linux or through its webserver, https://wenmr.science.uu.nl/haddock2.4. HADDOCK provides varying levels of docking restraints dependent on the user knowledge-base and project[84]. It boasts improved docking models and better docking accuracy with a data-driven approach compared to *de novo* docking. At the simplest input, it requires PDB coordinate files for each

protein and specified interacting residues to narrow the scope of the interacting interface. The output provides ranked models of the predicted protein-protein interaction with analysis statistics further described in the methods section. The ranking is organized by *best* HADDOCK score which can be described as a weighted sum of the outputted analysis statistics[84].

AlphaFold Multimer can be run locally on Linux or through Google's CoLab, https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=kOblAo-xetgx. Multimer can be a sequence-based or templated-based modelling approach that relies on the co-evolution of the interacting proteins[85]. The sequence-based approach will generate an AlphaFold model prior to the complex prediction whereas the template-based approach makes use of atomic coordinate files.

Here, we describe the predicted models generated through HADDOCK and AlphaFold Multimer of complexes that each of the polyglycine hydrolases makes with *Zea mays* ChitA. Despite substantial structural similarity between polyglycine hydrolases and *known* interacting residues in the complex - there are a few potential models generated for this interaction. However, as shown below, the complex modelled between Es-cmp and ChitA aligns well with the experimental data and illustrates a realistic model for this protein-protein interaction. We discuss the merits of each method and the challenges that arise from bioinformatic modelling for novel protein-protein interactions.

3.2 Material & Methods

3.2.1 AlphaFold2 Model Generation & Output

AlphaFold2 model generations were completed through Google Colab v1.4 online server, https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipy

nb#scrollTo=kOblAo-xetgx. The sequence files for Fvan-cmp were supplied by the recombinant protein sequences from previous expression work. To prevent advantages during the model comparison, the input did not include template coordinates.

The output of AlphaFold2 includes a coordinate file in .pdb format and a predicted aligned error (PAE) plot in .json format. These two files can immediately be imported into ChimeraX-1.4 for visualization of the model and quality inspection[72]. The model can be coloured by the two metrics offered intrinsically by AlphaFold2: predicted aligned error (PAE) domains or predicted local distance difference test (pLDDT) score[49]. The domains colouring scheme aids visualization of the different domains of a particular protein but doesn't represent model quality. The pLDDT colouring scheme is a confidence measure for the model ranging from dark blue (high confidence) to red (low confidence). This scale helps visualize regions of a model that may be inaccurate due to the limited information and/or a region expecting to be highly disordered[49].

3.2.2 Visualizing Electrostatic Potential Surface Maps

PDB2PQR webserver, https://server.poissonboltzmann.org/ provided the calculations and generation of charge-assigned maps in .pqr format from an uploaded PDB coordinate file[86]. Files were visualized through ChimeraX-1.4[72]. The color scale assignment follows convention: blue (positive charge), red (negative charge) and white (neutral).

3.2.3 HADDOCK Docking Preparation

HADDOCK docking simulations were completed through the HADDOCK v2.4-2022.08 online server, https://wenmr.science.uu.nl/haddock2.4/. The runs were conducted using simple default parameters with minimal modification to the simulation set-up.

3.2.4 HADDOCK Docking Input Parameters

The docking simulations described rely on previous experimental work that identified the catalytic serine and specific glycine residues within the ChitA polyglycine linker that are cleaved by each polyglycine hydrolase[20,25]. Each polyglycine hydrolase cleaves between different glycine-glycine bonds with varying specificity. For example, Es-cmp produces ChitA cleavage products that correspond to cleavage after G3, G4, G5, and G6 within the polyglycine linker[20]. Each of those glycine residues were specified on input. Additionally, researchers identified key residues outside of the cleavage site on ChitA that were speculated to aid in alignment of the polyglycine hydrolase cleavage site with the polyglycine linker[25].

Bz-cmp + ChitA (1) simulation:

Bz-cmp coordinate file was a RoseTTAFold model, trimmed to 4 Ångstroms error cut-off. It contained 591 of the 644 residues in the protein sequence. ChitA coordinate file was an AlphaFold2 model containing 254 residues. Seven residues were assigned as part of the *known* binding interface: Bz-cmp S369, Bz-cmp K372, Bz-cmp Y472, ChitA G45, ChitA G46, ChitA F62, and ChitA F63 to help guide the simulation from experimental work.

Bz-cmp + ChitA (2) simulation:

Bz-cmp coordinate file was a RoseTTAFold model, trimmed to 4 Ångstroms error cut-off. It contained 591 of the 644 residues in the protein sequence. ChitA coordinate file was an AlphaFold2 model containing 254 residues. Seven residues were assigned as part of the *known* binding interface: Bz-cmp S369, Bz-cmp K372, Bz-cmp Y472, Bz-cmp N474, ChitA G40-G46, ChitA F62, and ChitA F63 to help guide the simulation from experimental work.

Es-cmp + ChitA simulation:

Es-cmp coordinate file was a RoseTTAFold model, trimmed to 4 Ångstroms error cut-off. ChitA coordinate file was an AlphaFold2 model containing 254 residues. Ten residues were assigned as part of the *known* binding interface: Es-cmp S349, Es-cmp K352, Es-cmp Y452, Es-cmp N454, ChitA G40, ChitA G41, ChitA G42, ChitA G43, ChitA F62, and ChitA F63 to help guide the simulation from experimental work.

Fvan-cmp + ChitA (1) simulation:

Fvan-cmp coordinate file was modified atomic structure, eliminating alternate residue conformations and water molecules. ChitA coordinate file was an AlphaFold2 model containing 254 residues. Eight residues were assigned as part of the *known* binding interface: Fvan-cmp S343, Fvan-cmp K346, Fvan-cmp Y447, Fvan-cmp N449, ChitA G38, ChitA G39, ChitA F62, and ChitA F63 to help guide the simulation from experimental work.

Fvan-cmp + ChitA (2) simulation:

Fvan-cmp coordinate file was modified atomic structure, eliminating alternate residue conformations and water molecules. ChitA coordinate file was an AlphaFold2 model containing 254 residues. Seven residues were assigned as part of the *known* binding interface: Fvan-cmp Y447, Fvan-cmp N449, Fvan-cmp F534, ChitA G39, ChitA F63, ChitA K67, and ChitA G71 based on the peptide studies with polyglycine hydrolases.

3.2.5 HADDOCK Docking Visualization

The figures were generated through ChimeraX-1.4 and Inkscape[72]. Each figure maintained a standard colour scheme for ease of comparison. Colour schemes were designed at a chain and residue level. The enzymes, PGHs were custom-coloured “ocean” dark blue and the substrate proteins, ChitA were coloured “sky blue”. The catalytic active site serine was coloured

"tomato" red and the ChitA polyglycine linker was coloured "gold" yellow. Any analysis-identified interacting residues were coloured "moss" green.

3.2.6 HADDOCK Docking Analysis

HADDOCK runs output a cluster-ranked analysis, producing 5-10 top cluster models. Each cluster is ranked by their HADDOCK score which is a weighted sum of the displayed analysis statistics. The top models for the top 4 clusters were visualized in ChimeraX-1.4[72]. Utilizing the Matchmaker tool, the models were structurally aligned by their αCarbons. The alignment was specified to align to Chain A (enzyme) rather than Chain B (substrate). Utilizing the interface tool, each complex model identified interacting residues between the enzyme and the substrate. The four models were compared to find conserved residues within each model complex; these were reported.

3.2.7 AlphaFold2 Multimer Simulations

AlphaFold2 Multimer simulations were completed through Google Colab v1.4 online server, https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=kOblAo-xetgx.

Bz-cmp + ChitA simulation:

The sequence files for Bz-cmp and ChitA were supplied by the recombinant protein sequences from previous expression work. The run completed under the paired + unpaired MSA setting.

Fvan-cmp + ChitA (1) simulation:

The sequence files for Fvan-cmp and ChitA were supplied by the recombinant protein sequences from previous expression work. The run completed under the paired + unpaired MSA setting.

Fvan-cmp + ChitA (2) simulation:

The sequence files for Fvan-cmp and ChitA were supplied by the recombinant protein sequences from previous expression work. The coordinate files were uploaded to be used as a template for this simulation, the Fvan-cmp atomic structure and the AlphaFold2 ChitA model. The run completed under the paired + unpaired MSA setting.

Fvan-cmp + G_6(22) simulation:

The sequence files for Fvan-cmp and G_6(22) peptide were supplied by the recombinant protein sequences from previous expression work. The G6(22) peptide is 22 residues in length with the sequence, GGGGGGSGGANVANVVSDAFFN. The run completed under the paired + unpaired MSA setting.

3.3 Results

3.3.1 Fvan-cmp B-factor examination

Prior to using computational methods to model the PGH - ChitA complex, we attempted to ascertain information about the interaction by looking at the Fvan-cmp structure. There are many refinement parameters that need be taken into consideration when solving a structure and some can be exploited during the post-structure solution analysis. The temperature or commonly referred to as the B-factor is the measure of oscillation for any atom in space with the units, squared Angstroms[87]. This parameter can be used to anticipate flexible or disordered regions within a structure if the reported B-factors are high relative to the average structure B-factor. We predicted that Fvan-cmp would have flexibility within the looped inter-domain region to accommodate the substrate within its active site. The structure has an average B-factor of 36.0 with a range from 13.1 to 117. Using ChimeraX-1.4, we assigned the default colouring scheme

for B-factors: blue (low) to red (high) and applied it to the Fvan-cmp structure as seen in Figure 12. However, where we expected to see high B-factors on the loops within the inter-domain space or neighbouring the active site - we observed no evidence for this.

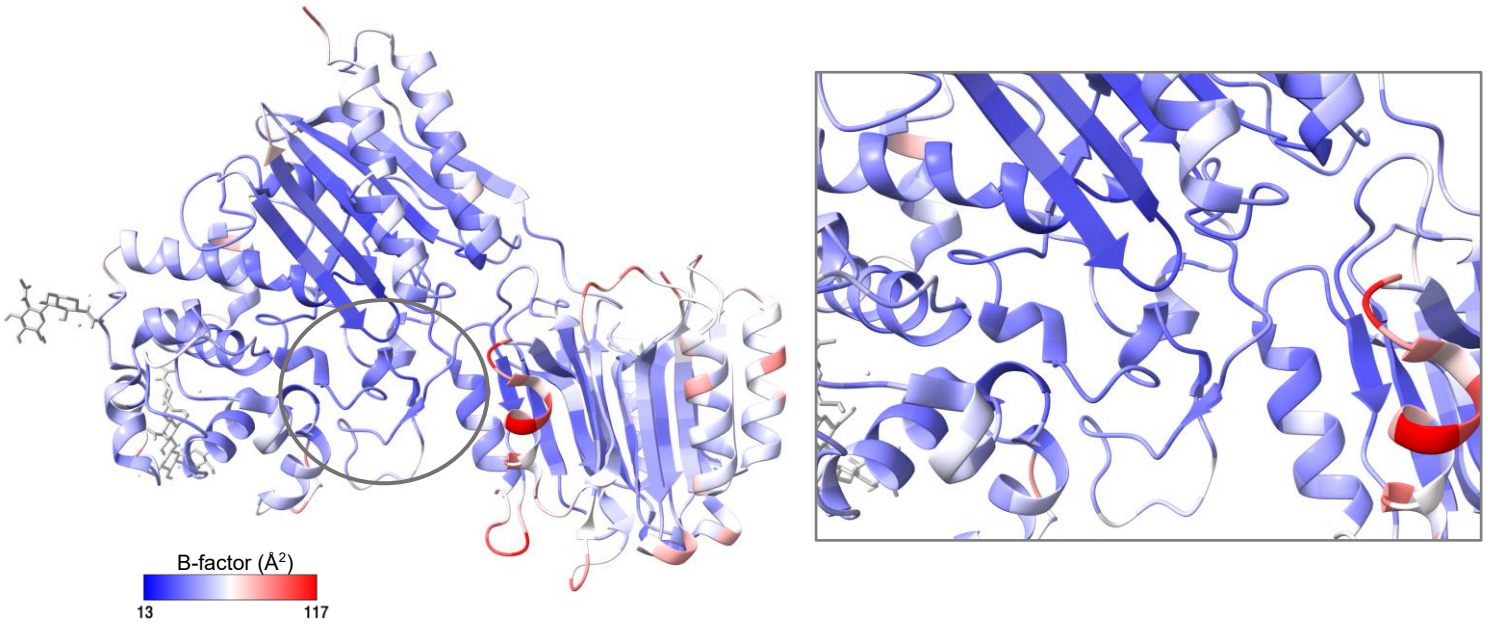


Figure 12. B-factor representation of Fvan-cmp

Fvan-cmp ribbon structure coloured by per-residue B-factors prepared in ChimeraX-1.4. The blue-white-red gradient is the default scale: blue represents residues with reported low B-factors and red represents residues with reported high B-factors.

There are several factors that can affect B-factors within a structure such as data resolution, crystal packing, water content[88]. These factors need to be considered when analyzing B-factors within a structure, compared between structures, or trying to draw conclusions about flexible or disordered regions within protein macromolecular structure. From what has been illustrated, we don't expect large conformational changes from Fvan-cmp based on its B-factors. However, this could largely be influenced by the crystal packing. The packing could impose more rigidity on the structure than what is observed in solution[88]. Owing to this, we chose to

pursue computational modelling to illustrate the interaction between polyglycine hydrolases and its substrate, ChitA.

3.3.2 A brief comparative analysis between AlphaFold2 and RoseTTAFold models to the atomic structure of Fvan-cmp

An overall theme within this thesis focuses on the capacities of computational methods and their potential and limitations in the field of structural biology in combination with *in vitro* and/or *in vivo* methods. In Chapter 2, we discussed the difficulties of the Fvan-cmp structure solution and how the RoseTTAFold model played a crucial role in the success of the solution. Here, we compare the models: AlphaFold2 and RoseTTAFold to the atomic structure.

We aligned the two models (individually represented in Appendix II) and atomic structure by their αCarbons, seen in Figure 13. Quantifying the similarities between the models and structure, we found that the AlphaFold2 model - atomic structure had a Cα-rmsd of 0.555 Å and the RoseTTAFold model - atomic structure had a Cα-rmsd of 1.072 Å. Visually, we found that the global shapes of Fvan-cmp are comparable. The relative orientations of the two domains and the secondary structures within the domains remain consistent between the models and the atomic structure. This is a promising outcome when looking at the predictive power of these methods to get some general outlooks of what the structure may look like for a given protein. The only notable difference between the models and the atomic structure are within the N-domain. As discussed within Chapter 2, this domain is abundant in nature but lacks experimental characterization and annotation which contributed to the issues modelling prior to the release of AlphaFold2 and RoseTTAFold. The local positioning of the helices within the N-domain appears messy since the three models don't align seamlessly as compared to the rest of the structure. For ease of viewing, we omitted one of the models to view the alignments against the

atomic structure, Fig. 13B-1 and Fig. 13B-2. In Figure 13B-1, the RoseTTAFold model has difficulty predicting the spatial relationship of the motifs relative to each other whereas the AlphaFold2 model (Fig. 13B-2) consistently aligns with the atomic structure.

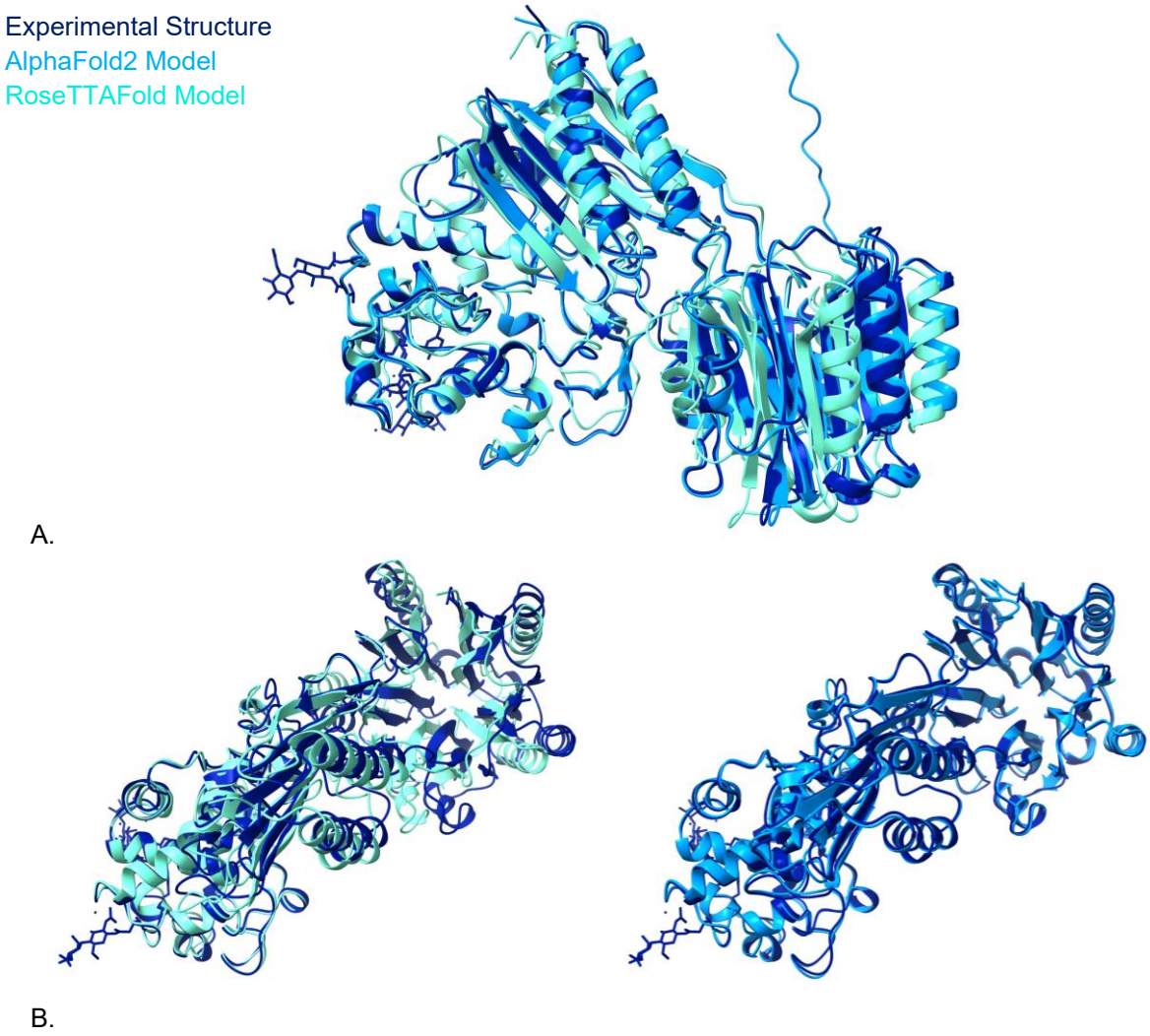


Figure 13. Fvan-cmp atomic structure vs. AlphaFold2 and RoseTTAFold models

Utilizing MatchMaker within the ChimeraX-1.4 program, the three structures were aligned with the default settings by their αCarbons.
(A) The AlphaFold2 model (light blue) and the RoseTTAFold model (cyan) were aligned against the atomic structure (dark blue).
(B) The individual models aligned to the atomic structure in a top-down view.

3.3.3 Predicting the protein-protein interface between polyglycine hydrolases and *Zea mays* ChitA

Bz-cmp & ChitA

The automatic assessment of the docking prediction data for the Bz-cmp and ChitA interaction is illustrated in Table 8. The data are arranged in a cluster ranking with the top cluster having the lowest HADDOCK score. The top-ranking cluster (1) was used as a comparative reference point to examine the other cluster models to identify potential interacting residues. The interacting residues were visualized through the program, Chimera X-1.4, utilizing the interface tool. The interface tool attenuated the eligible clusters for analysis by eliminating any model that did not include appropriate positioning of the nucleophilic serine responsible for the protease activity in polyglycine hydrolases. The cluster model (3) was used for visualization of residues as it best modelled the predicted Bz-cmp + ChitA complex.

Table 8. Bz-cmp + ChitA (2) HADDOCK docking cluster analysis statistics

	HADDOCK	Cluster Size	RMSD ($Å^2$)	Van der Waals ($kcal•mol^{-1}$)	Electrostatic ($kcal•mol^{-1}$)	Desolvation ($kcal•mol^{-1}$)	Restraints Violation	Buried Surface Area ($Å^2$)	Z-score
1	-110.2 ± 2.1	53	27.4 ± 0.1	-63.4 ± 4.0	-270.0 ± 17.5	5.9 ± 1.4	13.0 ± 19.2	2103.7 ± 17.7	-1.2
2	-104.8 ± 5.6	36	22.9 ± 0.3	-65.4 ± 4.5	-154.6 ± 16.1	-8.6 ± 1.8	1.4 ± 1.3	1764.9 ± 88.4	-1.0
3	-104.7 ±7.4	26	0.8 ± 0.5	-64.4 ± 5.7	-256.6 ± 31.2	10.0 ± 4.1	10.9 ± 15.3	2155.3 ± 160.2	-1.0
7	-94.0 ± 10.0	5	15.8 ± 0.3	-53.6 ± 7.6	-170.8 ± 8.3	-6.7 ±2.4	4.7 ± 3.7	1991.9 ± 146.6	-0.5
5	-75.5 ± 5.3	9	22.1 ± 0.1	-38.1 ± 4.3	-163.9 ± 17.2	-5.4 ± 3.6	6.7 ± 4.3	1563.3 ± 112.8	0.3
6	-66.3 ± 4.6	7	24.9 ± 1.2	-38.2 ± 3.9	-128.2 ± 33.2	-6.1 ± 2.0	37.0 ± 19.2	1223.6 ± 119.8	0.7
4	-65.8 ± 2.4	14	25.2 ± 0.1	-42.6 ± 4.0	-125.9 ± 42.0	-0.1 ± 5.7	19.9 ± 19.0	1357.1 ± 157.9	0.7
8	-41.4 ± 5.4	4	24.9 ± 0.7	-23.6 ± 3.6	-95.5 ± 21.4	-1.1 ± 0.2	24.7 ± 23.0	994.2 ± 63.6	1.8

The cluster identities are in the first column and represent the top clusters for this docking run. They are ranked by their HADDOCK score which is a weighted summation of the criteria in the table.

21 residues were identified across the different cluster models for the predicted interaction of Bz-cmp and ChitA, seen in Table 11. The residues were highlighted in Figure 14, with specific attention to the catalytic residue, Ser369 and the polyglycine linker residues. Bz-cmp has two predicted binding interfaces on ChitA, one on each of its domains.

The first interface occurs between Bz-cmp and the N-domain of ChitA. The conserved residues responsible for this interaction are Asn621 (Bz-cmp) and Gln8, Asp28, Pro34, Arg36 (ChitA). The second interface occurs between Bz-cmp and C-domain of ChitA. The conserved residues found across the top complex models are Ser49, Thr81, Ser83, Ser87, Arg181 (ChitA) and Thr391, Gln414, Asp418, Ser470, Tyr472, Tyr578, Gln584, Leu585, Phe586, Tyr588 (Bz-cmp). These residues could potentially play a role in anchoring of the substrate protein to the enzyme and aid in orienting the polyglycine linker for cleavage.

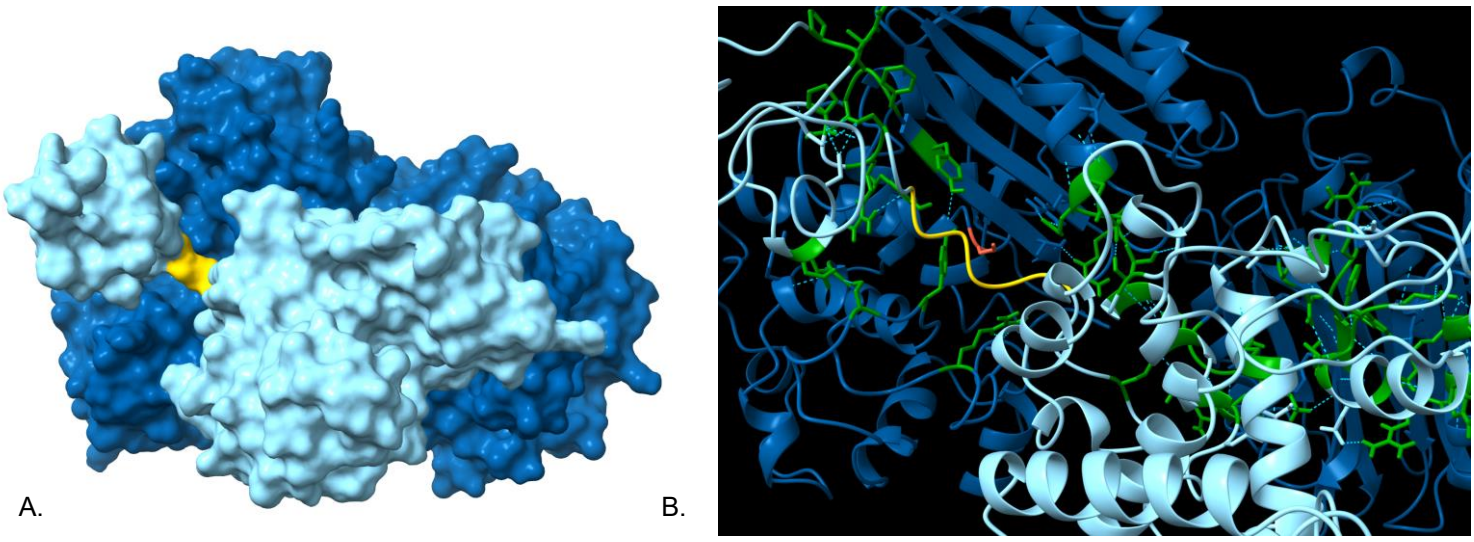


Figure 14. Bz-cmp + ChitA (2) model

(A) Bz-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with the catalytic Ser369 (orange) and polyglycine linker (yellow) highlighted. This model illustrates binding of the ChitA substrate to the C-domain but fails to properly align the polyglycine linker into the known active site.
(B) Bz-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with the catalytic Ser369 (orange) and polyglycine linker (yellow) highlighted. Conserved binding residues from the top models of this interaction are in green.

Es-cmp & ChitA

The automatic assessment of the docking data for the Es-cmp and ChitA predicted interaction is illustrated in Table 9. The data is arranged in a cluster ranking with the top cluster having the lowest HADDOCK score. The top-ranking cluster (7) was used as a comparative reference point to examine the other cluster models to identify potential interacting residues. Identification of the interacting residues were visualized through the program, Chimera X-1.4, utilizing the interface tool. The interface tool attenuated the eligible clusters for analysis by

eliminating any model that did not include the nucleophilic serine responsible for the protease activity in polyglycine hydrolases.

Table 9. Es-cmp + ChitA HADDOCK docking cluster analysis statistics

	HADDOCK	Cluster Size	RMSD ($Å^2$)	Van der Waals (kcal•mol^{-1})	Electrostatic (kcal•mol^{-1})	Desolvation (kcal•mol^{-1})	Restraints Violation	Buried Surface Area ($Å^2$)	Z-score
7	-87.5 ± 4.2	7	1.6 ± 1.7	-63.1 ± 1.3	-137.8 ± 12.8	3.1 ± 1.6	1.1 ± 0.44	1564.5 ± 98.3	-1.9
2	-69.5 ± 3.5	27	4.0 ± 2.6	-44.7 ± 7.0	-141.1 ± 37.5	3.8 ± 4.1	0.7 ± 0.68	1337.7 ± 94.4	-0.7
1	-68.7 ± 1.9	72	14.1 ± 0.2	-47.2 ± 2.0	-95.5 ± 3.2	-2.4 ± 1.9	0.1 ± 0.1	1308.0 ± 22.4	-0.6
3	-63.5 ± 4.0	15	7.9 ± 0.9	-36.2 ± 1.7	-145.5 ± 19.1	1.8 ± 1.8	0.5 ± 0.49	1241.5 ± 58.8	-0.3
8	-61.2 ± 9.0	5	12.8 ± 0.4	-34.5 ± 4.9	-117.1 ± 40.4	-5.7 ± 1.9	24.2 ± 22.24	1250.7 ± 102.1	-0.1
6	-57.6 ± 2.7	7	12.1 ± 1.1	-42.0 ± 2.9	-91.2 ± 15.8	2.6 ± 2.5	0 ± 0	1367.1 ± 47.5	0.1
5	-47.2 ± 3.3	11	12.2 ± 0.2	-14.4 ± 2.1	-174.8 ± 8.9	2.1 ± 3.1	0.3 ± 0.31	987.8 ± 91.1	0.8
4	-43.2 ± 5.5	11	12.7 ± 0.1	-21.8 ± 5.0	-108.6 ± 29.9	-0.3 ± 0.3	2.7 ± 2.78	828.1 ± 85.4	1.0
9	-32.0 ± 6.8	4	14.4 ± 0.4	-24.6 ± 6.1	-36.5 ± 5.7	-0.4 ± 1.8	2.3 ± 1.41	722.0 ± 94.5	1.7

The cluster identities are in the first column and represent the top clusters for this docking run. They are ranked by their HADDOCK score which is a weighted summation of the criteria in the table.

14 residues were identified across the different cluster models for the predicted interaction of Es-cmp and ChitA, seen in Table 11. The residues were highlighted in Figure 15, with specific attention to the catalytic residue, Ser349 and the polyglycine linker residues. There are two distinct networks of interactions between Es-cmp and ChitA. The first network includes the catalytic motif; Ser349 and Tyr452 and Phe413, Arg569, Asp570, Gly571, Thr572 and Tyr568. These residues are bound along the polyglycine linker between Gly42 to Gly46. The second network binds the N-terminal region of ChitA and a potential exo-site on Es-cmp. Es-cmp residues: Arg405, Asp406, Ile407, Pro409, Asp410 and ChitA residues, Gln8 form a hydrogen-bonded network that could be a potential anchor for the binding of the active site to the polyglycine substrate.

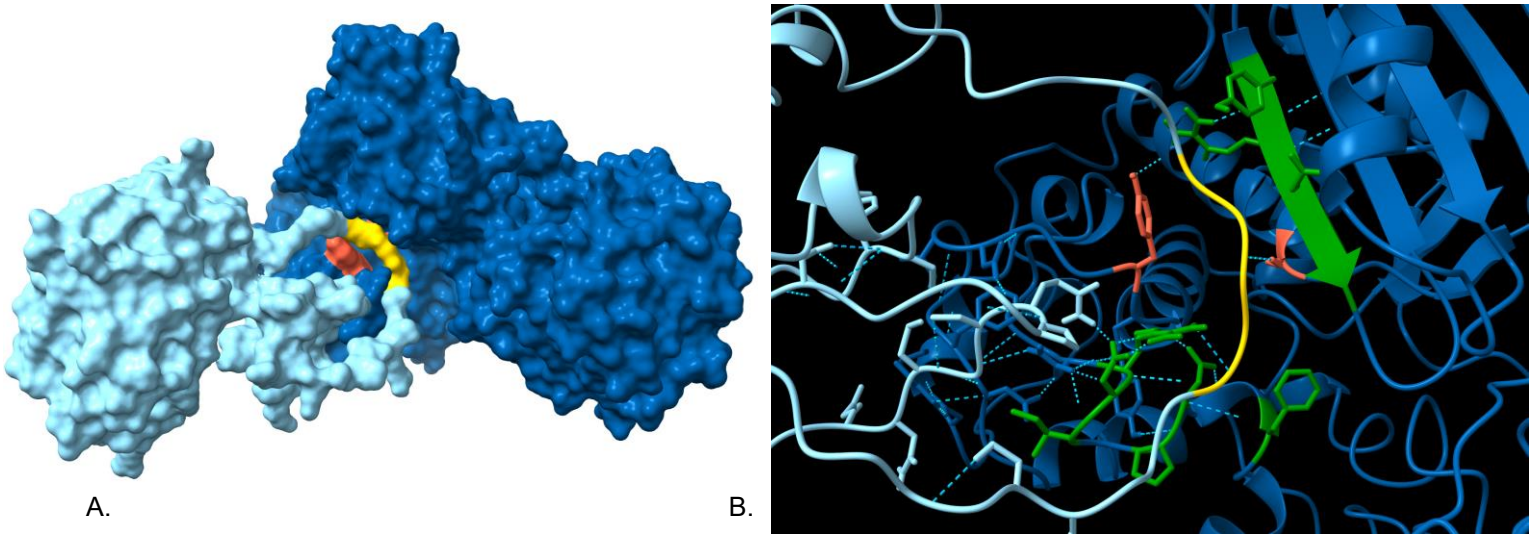


Figure 15. Es-cmp + ChitA model

(A) Es-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with the catalytic Ser349 (orange) and polyglycine linker (yellow) highlighted in a surface representation.
(B) Es-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with the catalytic Ser349 (orange) and polyglycine linker (yellow) highlighted in a ribbon representation. Conserved binding residues from the top models of this interaction are in green. The two Class C β-lactamase motifs are included in these conserved interacting residues for this model.

Fvan-cmp & ChitA

The automatic assessment of the docking data for the Fvan-cmp and ChitA predicted interaction is illustrated in Table 10. The data is arranged in a cluster ranking with the top cluster having the lowest HADDOCK score. The top-ranking cluster (7) was used as a comparative reference point to examine the other cluster models to identify potential interacting residues. Identification of the interacting residues were visualized through the program, Chimera X-1.4, utilizing the interface tool. The interface tool attenuated the eligible clusters for analysis by eliminating any model that did not include the nucleophilic serine responsible for the protease activity in polyglycine hydrolases.

Table 10. Fvan-cmp + ChitA (1) HADDOCK docking cluster analysis statistics

	HADDOCK	Cluster Size	RMSD ($Å^2$)	Van der Waals ($kcal•mol^{-1}$)	Electrostatic ($kcal•mol^{-1}$)	Desolvation ($kcal•mol^{-1}$)	Restraints Violation	Buried Surface Area ($Å^2$)	Z-score
7	-91.0 ± 7.0	9	1.1 ± 0.7	-55.3 ± 6.8	-165.4 ± 10.9	-2.9±0.7	2.8 ± 0.59	1958.2 ± 94.0	-2.4
1	-58.1 ± 2.1	20	19.2 ± 0.1	-33.4 ± 2.6	-117.7 ± 10.8	-1.2±1.9	0.5 ± 0.78	1180.9 ± 12.3	-0.3
2	-57.3 ± 9.2	20	21.2 ± 1.0	-37.5 ± 7.3	-99.3 ± 5.7	-0.1±1.7	0.7 ± 1.00	1024.6 ± 215.7	-0.2
9	-57.2 ± 3.2	8	18.0 ± 0.2	-35.4 ± 5.3	-92.8 ± 28.8	-3.7±2.5	3.6 ± 6.32	911.8 ± 51.8	-0.2
4	-52.5 ± 1.7	16	24.4 ± 0.2	-33.4 ± 1.9	-77.2 ± 3.4	-3.7±0.7	0.2 ± 0.2	887.0 ± 63.6	0.1
3	-50.8 ± 3.3	19	18.3 ± 0.4	-28.1 ± 2.1	-98.4 ± 11.8	-2.9±3.0	0.2 ± 0.28	841.0 ± 118.5	0.2
6	-47.4 ± 1.7	13	17.2 ± 3.2	-27.9 ± 1.8	-64.0 ± 17.8	-6.8±2.0	0.8 ± 0.88	774.7 ± 63.5	0.4
8	-40.3 ± 8.2	8	18.5 ± 0.6	-22.3 ± 1.8	-62.0 ± 22.6	-5.7±3.5	0.3 ± 0.36	810.4 ± 154.1	0.9
5	-33.4 ± 1.6	13	24.1 ± 0.4	-16.7 ± 1.9	-71.9 ± 8.3	-2.4±1.0	0 ± 0	664.6 ± 45.9	1.4

The cluster identities are in the first column and represent the top clusters for this docking run. They are ranked by their HADDOCK score which is a weighted summation of the criteria in the table.

36 residues were identified across the different cluster models for the predicted interaction of Fvan-cmp and ChitA, seen in Table 11. The residues were highlighted in Figure 16, with specific attention to the catalytic residue, Ser343 and the polyglycine linker residues. These residues can be categorized into three different binding regions on Fvan-cmp: N-domain, β-lactamase fold, C-domain α-cluster.

The first binding interaction between Fvan-cmp and ChitA occurs between two residues on Fvan-cmp's N-domain: Asn135 and Asp137 and the end of the ChitA polyglycine linker: Gly46-Gly48. It is speculated that this interaction maintains the orientation of the linker through the active site cleft.

The second binding interaction between Fvan-cmp and ChitA occurs between several residues within the β-lactamase fold in Fvan-cmp's C-domain. The residues involved in the interaction are: Glu290, Arg297, Lys301, Ser561, Asn562, Ile592, Ser593, Glu594, Thr595, Asp598, Glu599, Trp602 and Tyr603. These residues form a heavily hydrogen-bonded network with ChitA residues: Gln3, Asn4, Cys5, Gln6, Gln8, Phe16, Gly17, Tyr18, Cys19, Gly99, Thr100, Glu103, Pro206, and Gln207. Interestingly, this binding interface is only predicted in the Fvan-cmp model. However, the calculated Fvan-cmp electrostatic potential map (Fig. 22, Appendix II) illustrates a substantial localized surface charge along the top of Fvan-cmp's C-domain. This region of localized charge could contribute to a strong electrostatic interaction with the substrate further supporting this binding model theory.

The third binding interaction between Fvan-cmp and ChitA occurs between several residues in Fvan-cmp's C-domain α-cluster. The Fvan-cmp residues involved in this interface are: Asp397, Arg398, Ser399, Asp443, Tyr444, Ser445, and Tyr447. The ChitA residues

involved are Arg36 and Ser37. It is speculated that this interface aids in the positioning of the ChitA polyglycine linker into Fvan-cmp's active site.

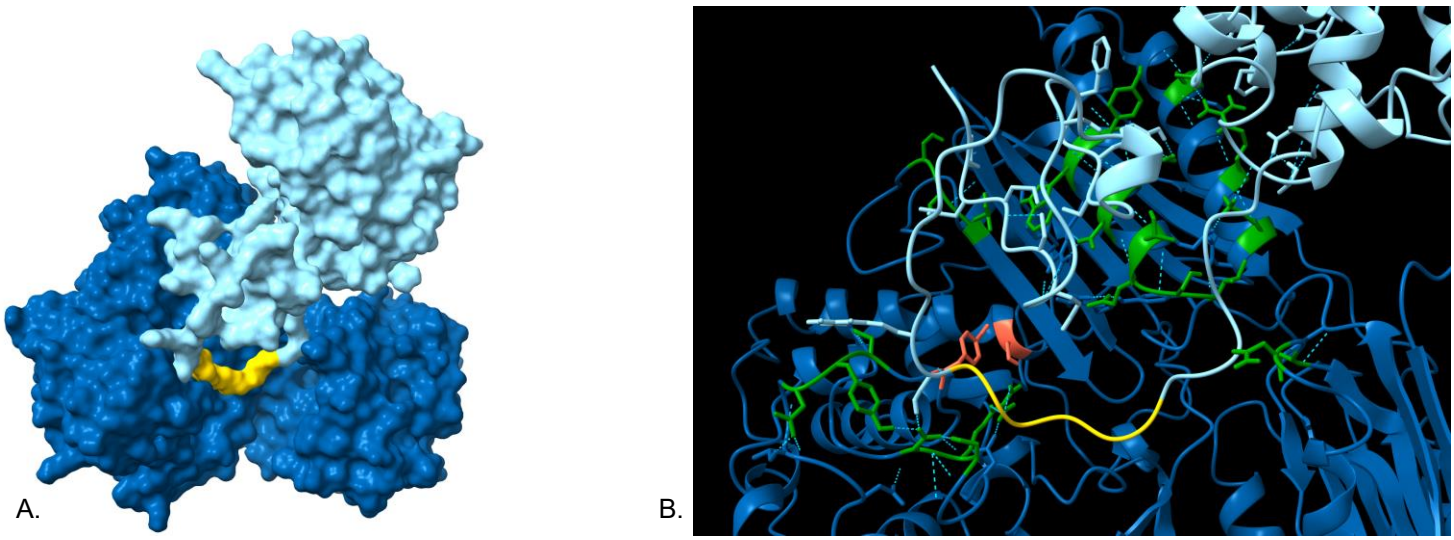


Figure 16. Fvan-cmp + ChitA model

(A) Fvan-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with ChitA polyglycine linker (yellow) highlighted in a surface representation. This complex model depicts binding along the peak of Fvan-cmp C-domain.
(B) Fvan-cmp ("ocean" dark blue) and ChitA ("sky" light blue) are bound with the catalytic Ser343 (orange) and polyglycine linker (yellow) highlighted in a ribbon representation. Conserved binding residues from the top models of this interaction are in green.

Table 11. Implicated predicted polyglycine hydrolase interacting residues

Protein	Residue		
ChitA	*All interactions*		
	Gln8		
	Bz-cmp interaction		
	Asp28	Pro34	Arg36
	Gly40	Ser49	Thr81
	Ser83	Ser87	Arg181
	Fvan-cmp interaction		
	Gln3	Asn4	Gly6
	Phe16	Cys5	Tyr18
	Cys19	Gly17	Ser37
	Thr100	Arg36	Gly99
	Gln207	Glu103	Pro206
Bz-cmp	Thr391	Gln414	Asp418
	Ser470	Tyr472	Tyr578
	Gln584	Leu585	Phe586
	Tyr588	Asn621	
Es-cmp	Asp142*	Arg405	Asp406
	Ile407	Pro409*	Asp410
	Phe413	Tyr568	Arg569
	Asp570	Gly571	Thr572
Fvan-cmp	Asn135	Asp137	Glu290
	Arg297	Lys301	Asp397
	Arg398	Ser399	Asp443
	Tyr444	Ser445	Tyr447
	Ile592	Ser561	Asn562
	Thr595	Ser593	Glu594
	Tyr603	Asp598	Glu599
			Trp602

Identified interface residues from predicted models of the interaction between different polyglycine hydrolases and *Zea mays* ChitA.

Table 12. AlphaFold Multimer identified residues in Fvan-cmp + ChitA interface

Multimer Run	Fvan-cmp residues	G_6(22) residues
Fvan-cmp + G_6(22)	Ser343 Arg398 Tyr447 Trp553 Asp564 Ser566 Arg589 Arg601 Trp602	Gly2* Gly3* Gly4* Gly5* Val15 Val16 Ser17* Asp18 Ala19 Phe21* Phe22*

*These residues correspond to those previously identified in the literature

The identified interface residues from the AlphaFold Multimer simulation using ChimeraX-1.4 'interfaces' tool.

3.3.4 AlphaFold Multimer success modelling protein-peptide interaction

There were attempts to generate predictions for complexes between each polyglycine hydrolase and two substrates: the full-length ChitA and a proposed inhibitory peptide from literature, G_6(22)[25]. The polyglycine hydrolase and ChitA complex predictions were unsuccessful using AlphaFold Multimer. In all generated models, the ChitA polyglycine linker failed to interact with any of the polyglycine hydrolases. However, the protein-peptide interaction was successfully modelled between Fvan-cmp and G_6(22). The interaction is illustrated in Figure 17 in both a surface and ribbon representation. Utilizing the ChimeraX-1.4 interface tool, we identified the predicted interacting residues, summarized in Table 12. The Fvan-cmp residues were identified as Ser343, Arg398, Tyr447, Trp553, Asp564, Ser566, Arg589, Arg601, Trp602 with the majority of the G_5(22) peptide residues also being identified in the protein-peptide interaction.

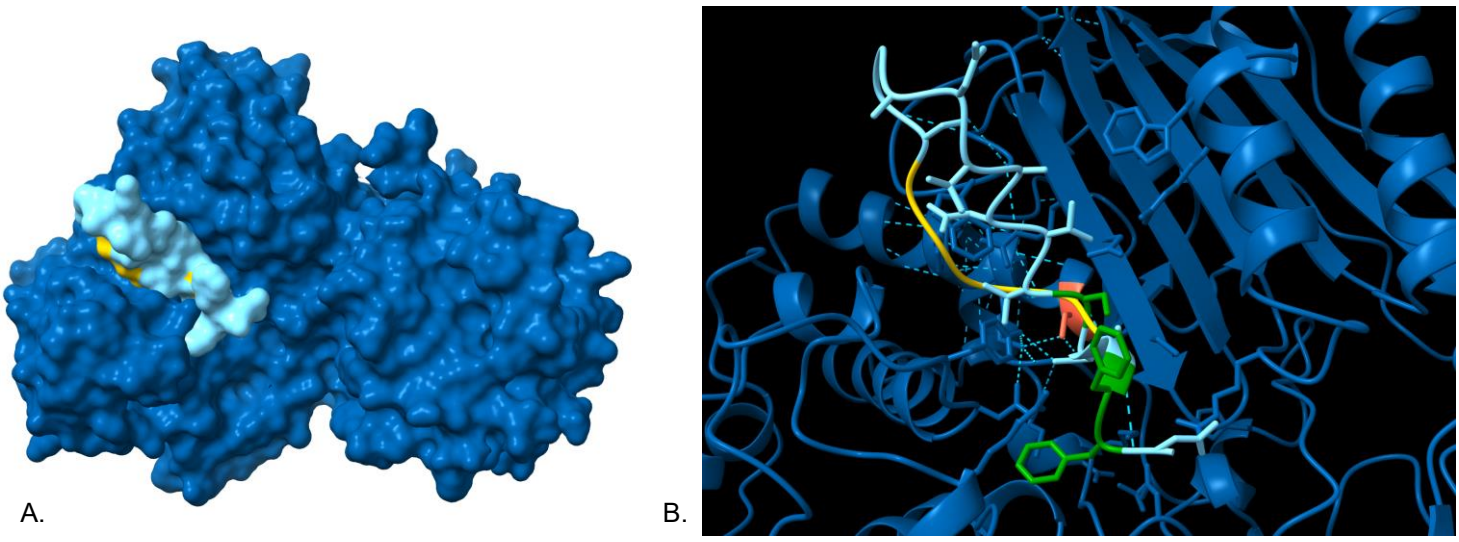


Figure 17. Fvan-cmp + G_6(22) model

(A) Fvan-cmp ("ocean" dark blue) and G6(22) peptide ("sky" light blue) are bound with the catalytic Ser343 (orange) and polyglycine linker (yellow) highlighted in a surface representation. The catalytic serine is hidden in the binding interface in this figure.
(B) Fvan-cmp ("ocean" dark blue) and G6(22) peptide ("sky" light blue) are bound with the catalytic Ser343 (orange) and polyglycine linker (yellow) highlighted in a ribbon representation. Literature defined binding residues outside of the polyglycine linker are in green.

3.4 Discussion

3.4.1 Evaluating the successes of complex modelling

Despite the high degree of similarity shared between the polyglycine hydrolases, the predicted binding interface is less clear. Bz-cmp and Fvan-cmp models neglect the expected interaction of the catalytic serine along the polyglycine linker but contribute potential sources of alternative binding, contributing to their weaker level of activity against ChitA. In contrast, Es-cmp successfully modelled an interaction consistent with catalytic activity. This model is proposed as the template for the binding interaction of PGHs and ChitA. Surprisingly, the respective complexes fail to implicate a significant number of corresponding residues in the

proposed binding interfaces. Importantly, however, the various predicted PGH-ChitA complexes are largely consistent with the variable strength of the enzymes and shed light on the biological implications of the enzymatic reactions.

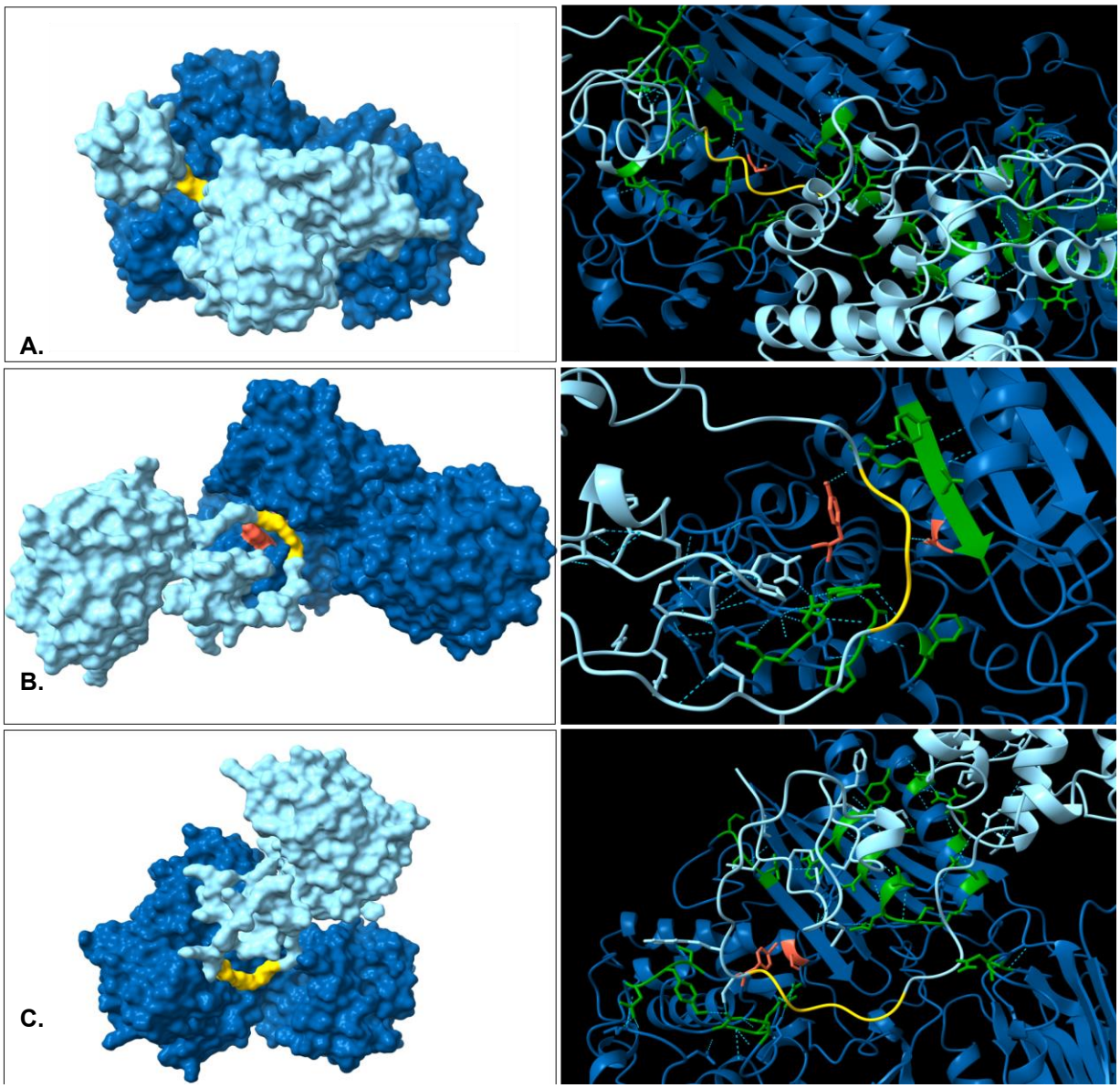


Figure 18. Summative comparison of different PGHs and their complex coordination

(A) Bz-cmp + ChitA, (B) Es-cmp + ChitA, and (C) Fvan-cmp + ChitA. The different PGH complex models are summarized. The PGHs (dark blue) are complexed with ChitA (light blue) and the polyglycine linker target (yellow) is highlighted along with the nucleophilic serine (orange). The surface representation in the left panel illustrates the overall complex and the ribbon representation in the right panel illustrate a closer view of the predicted interactions.

3.4.2 Es-cmp + ChitA complex models the template PGH-ChitA interaction

We hypothesize the Es-cmp + ChitA model represents the proposed interaction between polyglycine hydrolases and ChitA. The Es-cmp + ChitA model reasonably explains the enzyme-substrate interaction and supports the predictions made in literature. The model implicates two regions of binding between Es-cmp and ChitA: (i) the active site [E] and the polyglycine linker [S] and (ii) the C-domain α-cluster [E] and the N-domain [S]. The second binding region has been predicted to aid in proper positioning of the substrate and promote effective binding of Es-cmp to ChitA[12]. When Es-cmp was incubated with a truncated form of ChitA (missing the first 29 residues) it was unable to cleave the substrate[12].

3.4.3 Polyglycine hydrolases have different levels of activity against ChitA

The activity of polyglycine hydrolases against ChitA has been studied in diseased corn ears and *in vitro* reactions[20,50]. The activities of the polyglycine hydrolases have been previously quantified by their ability to convert half of ChitA to its truncated form under standard conditions (E½) and the respective ChitA products have been characterized[20]. As discussed in Chapter 2, the strengths vary as much as 276-fold difference between the weakest, Fvan-cmp, and the strongest, Es-cmp[20]. PGHs exhibit low-level promiscuous activity as they can cleave different glycine-glycine bonds within the ChitA polyglycine linker. For example, Es-cmp cleaves after Gly3, Gly4, Gly5, and Gly6 producing four potential products from the proteolytic reaction[20]. Conversely, Bz-cmp and Fvan-cmp produce six potential products from the reaction, illustrated in Figure 5. The promiscuous nature of these enzymes can be analyzed structurally using the predicted complex structures.

Specifically, there is an inverse correlation between the width of the active site cleft and the strength of the PGHs (based on E½ values). Notably, the differences in cleft width are more

substantial between Fvan-cmp and Bz-cmp/Es-cmp than the differences between Bz-cmp and Es-cmp aligning with the differences in varied strengths of the PGHs. Ranked by strength of PGH: Fvan-cmp << Bz-cmp < Es-cmp[20]. Es-cmp is the strongest polyglycine hydrolase against ChitA and has the narrowest active site cleft measuring 22.2 and 18.7 Å at either end in contrast to 23.1 and 19.6 Å in Fvan-cmp. The width of the active site was quantified by measuring the αC distance between the widest residues on either of the cleft. The region was identified by visualizing the surface map and identifying the residues in ribbon view. Visually, Es-cmp has a narrow, deep active site cleft whereas Fvan-cmp has a broad shallow cleft. Similar observations have been reported in previous studies that discuss enzymes that have a narrowing to their active site cavity to promote an improved fit for substrates[89]. Polyglycine linkers are a flexible, non-sterically hindered substrate that could benefit from high complementarity between the substrate and active site.

3.4.4 Biological implications of polyglycine hydrolase binding

The organisms, *B. zeicola* and *E. sorghi* are parasitic plant pathogens. Their polyglycine hydrolases are proposed to function as fungal defensive enzymes. They evade plant defenses by cleaving the polyglycine linker on chitinases within the plant's apoplastic space[20,90]. Within this environment, there are ample proteins, small molecules, and hormones accessible by polyglycine hydrolases[91]. These enzymes need to be specific for their substrate, accommodated by their narrow active region and high activity to thwart plant antifungal defense[12]. Indications from research to date hint at relatively weak interactions between PGHs and ChitA. Specifically, high E½ values, experimental difficulties co-purifying the bound enzyme-substrate complex, and inconsistencies in modelling the PGH-ChitA complexes. However, biologically there may be advantages to low affinities, as there would be saturating levels of defensive chitinases when

plants are in a stress-induced defensive state. It would therefore be functionally favourable to have weak substrate binding ensuring the activity is rate-limited by the catalysis. This enables constant truncation of ChitA and prevention of fungal infection defense.

3.4.5 Predictions are limited by science's current understanding

In this chapter, we describe two different methods used for modelling a novel protein-protein interaction. HADDOCK boasts a data-driven approach and AlphaFold Multimer relies on evolutionary relationships to model complexes[84,85,92]. Both methods are on the forefront of computational predictive modelling, but neither could be undisputed in its superiority of modelling the PGH-ChitA interaction.

Both methods experienced challenges in modelling the protein-protein and protein-peptide interactions for polyglycine hydrolases. Analyzing the data, we know that HADDOCK provided reasonable models of the protein-protein interaction. The documented promiscuous activity of the PGHs contributed to the challenge of consistently modelling this interaction. However, we were successful in providing a reasonable model for the PGH-ChitA interaction that is congruent with the biochemical data. Despite the success of this, the protein-peptide interaction was unsuccessfully modelled due to lack of structure for the peptide. Conversely, AlphaFold Multimer is an incredibly sophisticated method that relies on Multiple Sequence Alignments (MSA) to infer protein-protein interactions. The evident issue with this method is that if the relationship has not yet been described then this method won't be productive. We saw this in our attempts to model the PGH-ChitA interaction. No model successfully bound ChitA's polyglycine linker within the active site cleft on a polyglycine hydrolase. AlphaFold Multimer was successful in modelling the protein-peptide interaction. The Fvan-cmp + G_6(22) peptide interaction modelled the peptide occupying the length of the active site with the glycine residues

proximal to the catalytic serine residue. We anticipate that this success owes to the minimal complementarity of the peptide with another surface on Fvan-cmp.

Often researchers turn to modelling and prediction algorithms when there's a stall in their experimental work. This chapter has shown that the effectiveness of these methods regrettably hinges on experimental data. The first method, HADDOCK reasonably models protein-protein interactions when the core interacting residues are defined. Without these residues, the method is less sophisticated in its predictions which we anticipate would be further magnified when modelling weak and/or transient protein-protein interactions. The second method, AlphaFold Multimer has had success modelling protein-protein interactions outside the scope of this project[85]. Within this project, it was only successful in modelling the protein-peptide interaction. The sophistication of this method is limited by the current known protein-protein interactions using the MSA to chaperone the prediction. This is problematic when attempting to model a protein-protein interaction that either has not been described before nor is it evolutionarily related. We saw this by its inability to predict a reasonable model for the PGH-ChitA interaction which includes a novel class of enzymes. Despite the strides in new prediction algorithms, they are not at the caliber to substitute experimental work.

Chapter 4: Concluding Remarks & Future Directions

3.4.6 Fvan-cmp structure

Through the combination of X-ray crystallography and deep-learning protein modelling, I was able to determine the atomic structure of Fvan-cmp. Although the structure provided insight about polyglycine hydrolases, it also raised more unknowns about this novel class of enzymes. The structure contained two distinct domains: N- and C- domain. The amino domain consisted of the five structural repeats that made up a novel tertiary fold with a current unknown function. This fold does not appear to be organism specific as it is predicted in proteins across the kingdoms. The carboxyl domain consisted of a β-lactamase fold but did not exhibit β-lactamase activity nor was it inhibited by common β-lactamase inhibitors. Manipulation of the active site residues to recover the missing β-lactamase motif was unsuccessful in the gain-of-function attempt.

Future work should focus on determining the function of the tertiary fold in PGHs' N-domain. Interestingly, this fold has been predicted in proteins found across the kingdoms but has never been functionally annotated. It would be interesting to see if there is a conserved general function to this tertiary fold within all proteins containing it.

3.4.7 PGH structures

Utilizing the Fvan-cmp atomic structure and *de novo* modelling methods, I was able to generate models for Bz-cmp and Es-cmp. The global structures of polyglycine hydrolases are strikingly similar however I identified differences within their catalytic region that could contribute to the known differences in substrate affinity and specificity. The active site between the PGHs varied by their depth and width which I anticipate plays a role in their substrate

affinity and product differences. Es-cmp had the narrowest active site when compared to Bz-cmp and Fvan-cmp and I anticipate this was advantageous for the enzyme's activity.

Future work should address complexed structures of PGHs. As the structures of members of this family of enzymes are clearly closely related, enzyme-substrate or other catalytic complexes will be required to reveal the reported biochemical and activity differences.

3.4.8 Proposal of a catalytic dyad and its oxyanion hole

Using the classification of polyglycine hydrolases and the solved atomic structure of Fvan-cmp I proposed the presence of a Ser-Lys dyad responsible for the catalytic activity of these enzymes. The S12 family of serine proteases maintain a conserved Ser-Lys dyad that supports the acyl-enzyme intermediate step in its catalytic mechanism. This hypothesis was further strengthened by the conserved β-lactamase SVSK motif found in each PGH's active site and the hypothesized oxyanion hole that surrounded this motif.

The oxyanion coordinating residues were identified using a reference β-lactamase in a bound-state to focus the search for eligible residues. I identified the proposed oxyanion coordinating residues: Gly591/Thr592 (Bz-cmp), Gly571/Thr572 (Es-cmp) and Gly565/Ser566 (Fvan-cmp) based on their proximity to the nucleophilic serine and their relative positioning above the catalytic residues.

Future work should focus on confirming these residues by solving an inhibitor-bound structure. The inhibitor would prevent the enzyme from cleavage and increase the probability of solving a structure resembling the transition state. This work could be completed in tandem with site-directed mutagenesis to confirm the dyad and oxyanion residues. The combination of these two methods and results would confirm the hypothesis we've put forth in this work.

3.4.9 PGH-ChitA model interaction

Unable to crystallize a PGH-ChitA or PGH-inhibitor complex, I turned to complex modelling via HADDOCK. I found variation between polyglycine hydrolases and even within a docking run. Analyzing the generated predictions in conjunction with previous work allowed me to hypothesize the model interaction between ChitA and polyglycine hydrolases. I formed this model interaction around the generated model complex for Es-cmp and ChitA. This model depicted the correct orientation of enzyme to substrate for the cleavage to be possible. Despite this, I discussed the merits and limitations of predictive software and/or servers in-length.

Future directions should focus on gaining an atomic structure of a polyglycine hydrolase complexed with an inhibitor or substrate. There are already catalytic mutants for Es-cmp and Fvan-cmp successfully recombinantly expressed and purified. The atomic structure would make the complex and binding residues more than hypothetical and provide key details about the interaction between these two proteins.

3.4.10 Current *in silico* methods: protein modelling and protein-protein docking

Within this thesis, I discussed at length the process, the results and the conclusions that can be drawn from the most prominent *de novo* methods: AlphaFold2, RoseTTAFold and *in silico* docking methods: HADDOCK and AlphaFold2 Multimer.

I found that for this project, the current *de novo* methods did an exceptional job at modelling a protein family that had previously failed to be modelled in entirety. RoseTTAFold was able to predict the novel N-domain tertiary fold that was confirmed with the Fvan-cmp crystal structure. Both modelling methods predicted a largely accurate global structure of Fvan-cmp and its relatives.

The protein-protein docking methods were informative but were not as concrete in their predictions nor consistent. I found HADDOCK was transparent in its process and results, leaving the interpretation with the user. This allowed the method to fine-tune the docking strategy by identifying interacting residues and using those to guide the process. AlphaFold2 Multimer was largely black box in its process and built from protein sequences instead of input coordinates. I found that because of the novelty of the PGH-ChitA interaction, Multimer didn't have enough knowledge to accurately model an interaction whereas HADDOCK input allowed it to represent the complex more-accurately. During this work, it was apparent to take the approach of 'try many and see what works'. Protein modelling for example, has three major types of methodologies - it would be important to have a sampling of each type in the process to find the method that will work best for that particular protein project. This allows a comparison of methods and ensures a more-informed view that will hopefully prevent user-bias and hypothesis-bias in *in silico* work.

3.4.11 Agricultural impacts

This work began and continues with an interest in agricultural impact. Polyglycine hydrolases have been identified in pathogenic fungi involved in corn spoilage. The results reported here will contribute to fundamental understanding of this family of enzymes and set the stage for a sophisticated approach to their intervention. Food spoilage, and food security in general, is a major destabilizing issue in many parts of the world. Any ameliorating contribution will be of benefit.

I was able to solve the structure of a PGH and model the structure of other PGHs. I proposed a catalytic dyad and the residues responsible for stabilizing the enzyme during catalysis. This will inform the search and development of suitable inhibitors to prevent future

PGH action on chitinases. The combination of this and previous work will permit future analysis on these enzymes to find their biochemical and structural vulnerabilities, with a view towards addressing the larger biological and societal issues.

Letter of copyright permission

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jan 04, 2023

This Agreement between Nicole Dowling ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5453191043009
License date	Dec 20, 2022
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Protein Science
Licensed Content Title	Polyglycine hydrolases: Fungal β lactamase like endoproteases that cleave polyglycine regions within plant class IV chitinases
Licensed Content Author	Todd A. Naumann, Michael J. Naldrett, Todd J. Ward, et al
Licensed Content Date	May 22, 2015
Licensed Content Volume	24
Licensed Content Issue	7
Licensed Content Pages	11
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Electronic
Portion	Figure/table
Number of figures/tables	1
Will you be translating?	No
Title	A structural investigation of novel fungal polyglycine hydrolases
Institution name	University of Waterloo
Expected	Feb 2023

presentation date

Portions	Figure 1, whole image
Requestor Location	Nicole Dowling [redacted]
Publisher Tax ID	EU826007151
Total	0.00 CAD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.
- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.
- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you

References

1. Hamid, R. *et al.* Chitinases: An update. *J Pharm Bioallied Sci* 5, 21–29 (2013).

2. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 49, D498–D508 (2020).

3. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* 50, D571–D577 (2021).

4. Patil, R. S., Ghormade, V. & Deshpande, M. V. Chitinolytic enzymes: an exploration. *Enzyme Microb Tech* 26, 473–483 (2000).

5. Oyeleye, A. & Normi, Y. M. Chitinase: diversity, limitations, and trends in engineering for suitable applications. *Bioscience Rep* 38, BSR2018032300 (2018).

6. Rathore, A. S. & Gupta, R. D. Chitinases from Bacteria to Human: Properties, Applications, and Future Perspectives. *Enzym Res* 2015, 791907 (2015).

7. Flach, J., Pilet, P.-E. & Jollès, P. What's new in chitinase research? *Experientia* 48, 701–716 (1992).

8. Taira, T. *et al.* A plant class V chitinase from a cycad (Cycas revoluta): Biochemical characterization, cDNA isolation, and posttranslational modification. *Glycobiology* 19, 1452–1461 (2009).

9. Collinge, D. B. *et al.* Plant chitinases. *The Plant Journal* (1993).

10. Huynh, Q. K. *et al.* Antifungal proteins from plants. Purification, molecular cloning, and antifungal properties of chitinases from maize seed. *J Biol Chem* 267, 6635–6640 (1992).

11. Punja, Z. K. & Zhang, Y. Y. Plant chitinases and their roles in resistance to fungal diseases. *J Nematol* 25, 526–40 (1993).

12. Naumann, T. A., Wicklow, D. T. & Price, N. P. J. Polyglycine hydrolases secreted by Pleosporineae fungi that target the linker region of plant class IV chitinases. *Biochem J* 460, 187–198 (2014).

13. Naumann, T. A. & Wicklow, D. T. Allozyme-Specific Modification of a Maize Seed Chitinase by a Protein Secreted by the Fungal Pathogen Stenocarpella maydis. *Phytopathology* 100, 645–654 (2010).

14. Chaudet, M. M., Naumann, T. A., Price, N. P. J. & Rose, D. R. Crystallographic structure of ChitA, a glycoside hydrolase family 19, plant class IV chitinase from Zea mays. *Protein Sci* 23, 586–593 (2014).

15. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* eabj8754 (2021) doi:10.1126/science.abj8754.

16. Iseli, B., Armand, S., Boller, T., Neuhaus, J.-M. & Henrissat, B. Plant chitinases use two different hydrolytic mechanisms. *Febs Lett* 382, 186–188 (1996).

17. Hoell, I. A., Dalhus, B., Heggset, E. B., Aspmo, S. I. & Eijsink, V. G. H. Crystal structure and enzymatic properties of a bacterial family 19 chitinase reveal differences from plant enzymes. *Febs J* 273, 4889–4900 (2006).

18. Price, N. P. J. & Naumann, T. A. A high-throughput matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry-based assay of chitinase activity. *Anal Biochem* 411, 94–99 (2011).

19. Consortium, T. Caz. Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology* 28, 3–8 (2017).

20. Naumann, T. A., Naldrett, M. J., Ward, T. J. & Price, N. P. J. Polyglycine hydrolases: Fungal β-lactamase-like endoproteases that cleave polyglycine regions within plant class IV chitinases. *Protein Science* 24, 1147–1157 (2015).

21. Naumann, T. A., Wicklow, D. T. & Price, N. P. J. Identification of a Chitinase-modifying Protein from Fusarium verticillioides TRUNCATION OF A HOST RESISTANCE PROTEIN BY A FUNGALYSIN METALLOPROTEASE. *J Biol Chem* 286, 35358–35366 (2011).

22. Naumann, T. A., Naldrett, M. J. & Price, N. P. J. Kilbournase, a protease-associated domain subtilase secreted by the fungal corn pathogen Stenocarpella maydis. *Fungal Genet Biol* 141, 103399 (2020).

23. Cera, E. D. Serine proteases. *Iubmb Life* 61, 510–515 (2009).

24. Goldberg, S. D., Iannuccilli, W., Nguyen, T., Ju, J. & Cornish, V. W. Identification of residues critical for catalysis in a class C β-lactamase by combinatorial scanning mutagenesis. *Protein Sci* 12, 1633–1645 (2003).

25. Naumann, T. A., Bakota, E. L. & Price, N. P. J. Recognition of corn defense chitinases by fungal polyglycine hydrolases. *Protein Science* 26, 1214–1223 (2017).

26. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol* 234, 779–815 (1993).

27. Martí-Renom, M. A. *et al.* COMPARATIVE PROTEIN STRUCTURE MODELING OF GENES AND GENOMES. *Annu Rev Bioph Biom* 29, 291–325 (2000).

28. Hameduh, T., Haddad, Y., Adam, V. & Heger, Z. Homology Modeling in the Time of Collective and Artificial Intelligence. *Comput Struct Biotechnology J* 18, 3494–3506 (2020).

29. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33, W244–W248 (2005).

30. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10, 845–858 (2015).

31. Bienert, S. *et al.* The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res* 45, D313–D319 (2017).

32. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46, W296–W303 (2018).

33. Werner, T., Morris, M. B., Dastmalchi, S. & Church, W. B. Structural modelling and dynamics of proteins for insights into drug interactions. *Adv Drug Deliver Rev* 64, 323–343 (2012).

34. Muhammed, M. T. & Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des* 93, 12–20 (2019).

35. Jones, D. T., Taylort, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* 358, 86–89 (1992).

36. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5, 725–738 (2010).

37. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 7, 1511–1522 (2012).

38. Rost, B., Schneider, R. & Sander, C. Protein fold recognition by prediction-based threading11Edited by F. E. Cohen. *J Mol Biol* 270, 471–480 (1997).

39. McGuffin, L. J. *et al.* IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res* 47, W408–W413 (2019).

40. Zhang, Y., Kolinski, A. & Skolnick, J. TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction. *Biophys J* 85, 1145–1164 (2003).

41. Robson, B. De novo protein folding on computers. Benefits and challenges. *Comput Biol Med* 143, 105292 (2022).

42. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins Struct Funct Bioinform* 37, 194–198 (1999).

43. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *Bmc Biol* 5, 17 (2007).

44. Walsh, I. *et al.* Ab initio and homology based prediction of protein domains by recursive neural networks. *Bmc Bioinformatics* 10, 195 (2009).

45. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct Funct Bioinform* 80, 1715–1735 (2012).

46. Mortuza, S. M. *et al.* Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nat Commun* 12, 5011 (2021).

47. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions1 1Edited by F. E. Cohen. *J Mol Biol* 268, 209–225 (1997).

48. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 1–9 (2021) doi:10.1038/s41586-021-03828-1.

49. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

50. Naumann, T. A., Wicklow, D. T. & Kendra, D. F. Maize seed chitinase is modified by a protein secreted by Bipolaris zeicola. *Physiol Mol Plant P* 74, 134–141 (2009).

51. Naumann, T. A. Modification of recombinant maize ChitA chitinase by fungal chitinase-modifying proteins. *Mol Plant Pathol* 12, 365–372 (2011).

52. Coleman, J. J. *et al.* The Genome of Nectria haematococca: Contribution of Supernumerary Chromosomes to Gene Expansion. *Plos Genet* 5 (2009).

53. Riley, R. *et al.* Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proc National Acad Sci* 111, 9923–9928 (2014).

54. Vagin, A. & Teplyakov, A. MOLREP: an Automated Program for Molecular Replacement. *J Appl Crystallogr* 30, 1022–1025 (1997).

55. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr Sect D Biological Crystallogr* 62, 1002–1011 (2006).

56. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr Sect D Biological Crystallogr* 53, 240–255 (1997).

57. Naumann, T. A., Sollenberger, K. G. & Hao, G. Production of selenomethionine labeled polyglycine hydrolases in Pichia pastoris. *Protein Expres Purif* 194 (2022).

58. O'Callaghan, C. H., Morris, A., Kirby, S. M. & Shingler, A. H. Novel Method for Detection of β-Lactamases by Using a Chromogenic Cephalosporin Substrate. *Antimicrob Agents Ch* 1, 283–288 (1972).

59. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276, 307–326 (1997).

60. 4, C. C. P., Number. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr Sect D Biological Crystallogr* 50, 760–763 (1994).

61. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr Sect D* 60, 2126–2132 (2004).

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

63. Agirre, J. *et al.* Privateer: software for the conformational validation of carbohydrate structures. *Nat Struct Mol Biol* 22, 833–834 (2015).

64. Gupta, V. *et al.* Protein PEGylation for cancer therapy: bench to bedside. *J Cell Commun Signal* 13, 319–330 (2019).

65. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Biorxiv* (2022) doi:10.1101/2022.06.02.494367.

66. Kempen, M. van *et al.* Foldseek: fast and accurate protein structure search. *Biorxiv* (2022) doi:10.1101/2022.02.07.479398.

67. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res* 38, W545–W549 (2010).

68. Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24, 2780–2781 (2008).

69. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* 28, 235–242 (2000).

70. Delfosse, V. *et al.* Structure of the Archaeal Pab87 Peptidase Reveals a Novel Self-Compartmentalizing Protease Family. *Plos One* 4 (2009).

71. Lahiri, S. D. *et al.* Structural Insight into Potent Broad-Spectrum Inhibition with Reversible Recyclization Mechanism: Avibactam in Complex with CTX-M-15 and Pseudomonas aeruginosa AmpC β-Lactamases. *Antimicrob Agents Ch* 57, 2496–2505 (2013).

72. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* 30, 70–82 (2021).

73. Page, M. G. P. A Unified Numbering Scheme for Class C β-Lactamases. *Antimicrob Agents Ch* 64, (2020).

74. Sauvage, E., Kerff, F., Terrak, M., Ayala, J. A. & Charlier, P. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *Fems Microbiol Rev* 32, 234–258 (2008).

75. Kraut, J. Serine Proteases: Structure and Mechanism of Catalysis. *Annu Rev Biochem* 46, 331–358 (1977).

76. Bobofchak, K. M., Pineda, A. O., Mathews, F. S. & Cera, E. D. Energetic and Structural Consequences of Perturbing Gly-193 in theOxyanion Hole of SerineProteases*. *J Biol Chem* 280, 25644–25650 (2005).

77. Yang, H. & Wong, M. W. Oxyanion Hole Stabilization by C–H···O Interaction in a Transition State□A Three-Point Interaction Model for Cinchona Alkaloid-Catalyzed Asymmetric Methanolysis of meso-Cyclic Anhydrides. *J Am Chem Soc* 135, 5808–5818 (2013).

78. Voet, D., Voet, J. G. & Pratt, C. W. *Fundamentals of Biochemistry: Life at the Molecular Level*. (John Wiley & Sons, Inc., 2015).

79. Chen, C. C. H., Rahil, J., Pratt, R. F. & Herzberg, O. Structure of a Phosphonate-inhibited β-Lactamase An Analog of the Tetrahedral Transition State/Intermediate of β-Lactam Hydrolysis. *J Mol Biol* 234, 165–178 (1993).

80. Schapira, M., Tyers, M., Torrent, M. & Arrowsmith, C. H. WD40 repeat domain proteins: a novel target class? *Nat Rev Drug Discov* 16, 773–786 (2017).

81. Gao, M., Glenn, A. E., Blacutt, A. A. & Gold, S. E. Fungal Lactamases: Their Occurrence and Function. *Front Microbiol* 08, 1775 (2017).

82. Dubus, A., Normark, S., Kania, M. & Page, M. G. P. The Role of Tyrosine 150 in Catalysis of .beta.-Lactam Hydrolysis by AmpC .beta.-Lactamase from Escherichia coli Investigated by Site-Directed Mutagenesis. *Biochemistry-us* 33, 8577–8586 (1994).

83. Chikunova, A. & Ubbink, M. The roles of highly conserved, non-catalytic residues in class A β-lactamases. *Protein Sci Publ Protein Soc* 31 (2022).

84. Zundert, G. C. P. van *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* 428, 720–725 (2016).

85. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *Biorxiv* (2022). doi:10.1101/2021.10.04.463034.

86. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci* 27, 112–128 (2018).

87. Rhodes, G. *Crystallography Made Crystal Clear*. (Academic Press, 2006).

88. Sun, Z., Liu, Q., Qu, G., Feng, Y. & Reetz, M. T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem Rev* 119, 1626–1665 (2019).

89. Yang, G., Miton, C. M. & Tokuriki, N. A mechanistic view of enzyme evolution. *Protein Sci* 29, 1724–1747 (2020).

90. Ökmen, B. & Doehlemann, G. Inside plant: biotrophic strategies to modulate host immunity and metabolism. *Curr Opin Plant Biol* 20, 19–25 (2014).

91. Farvardin, A. *et al.* The Apoplast: A Key Player in Plant Survival. *Antioxidants* 9, 604 (2020).

92. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc* 125, 1731–1737 (2003).

Appendix I

These are the supplementary materials for Chapter 2.

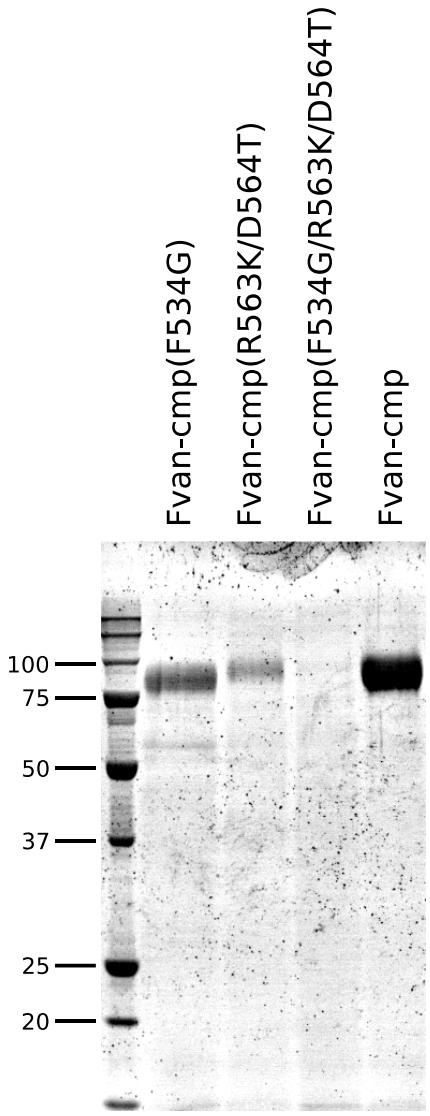


Figure 19. Expression levels of recombinant Fvan-cmp mutants

SDS-PAGE and Coomassie staining analysis of cell-free media. After two days of expression, cell-free media from *K. phaffiii* cultures expressing either Fvan-cmp or the single, double, or triple mutants was analyzed. The site-directed mutants all showed reduced amounts of protein, likely due to decreased stability of the proteins.

Table 13. Conserved β-lactamase shell residues

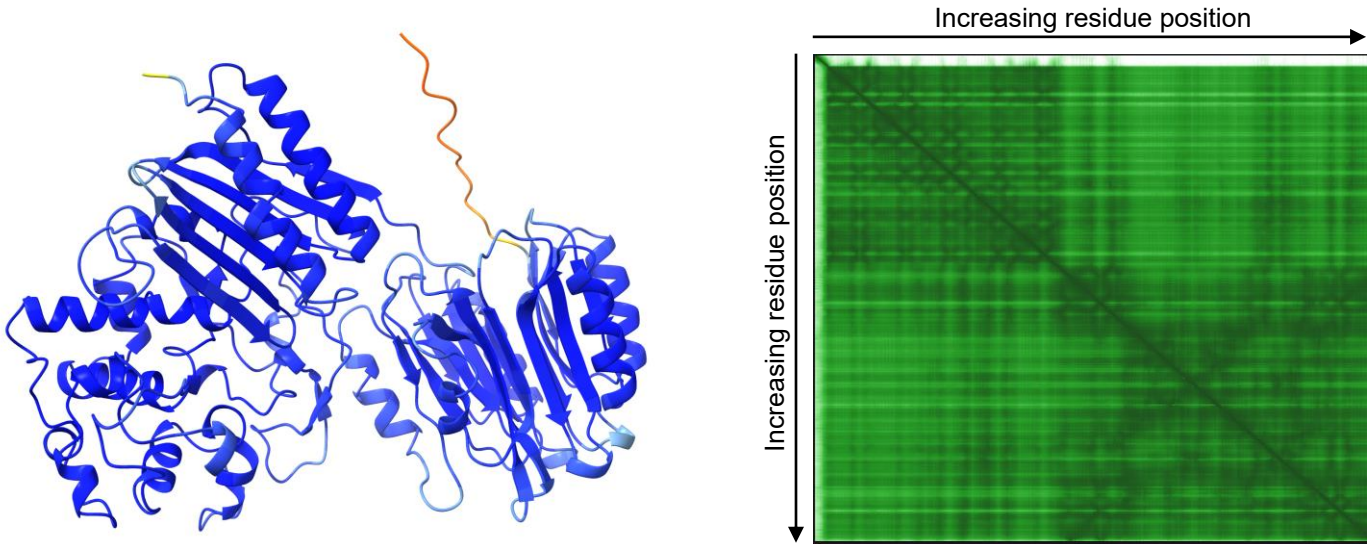
Residue	Fvan-cmp residue*	Shell	Function
E37	Q308	III	Proper folding
(A, D, Q, S)			
R65	P335	III	Loop interactions
(A, T, P, L, H, K, C)			
T71	L341	II	Reduces mobility of active site serine
(S, A, V, L)			
D131	D403	II	Stability and global positioning
A185	S456	III	
(S, T, V, E, Q, R, N, G)			
W229	S501	III	Hydrophobic and stacking interactions
(S, A, Y, C, F)			

*The Fvan-cmp C-domain begins at residue 271, so the corresponding residue was scaled up to fit the full sequence length.

The residues conserved across Class A β-lactamases and are designated to shells based on proximity to the active site and their related function. The bracketed residues are the conserved alternates found in nature. Conserved residues shared between polyglycine hydrolases and β-lactamases are listed.

Appendix II

These are the supplementary materials for Chapter 3.



A. B.

Figure 20. AlphaFold2 model of Fvan-cmp

(A) The AlphaFold2 model of Fvan-cmp in the default predicted local distance difference test (pLDDT) colouring scheme: blue (high confidence) - red (low confidence).
(B) The associated predicted aligned error (PAE) plot for the model generated: dark green (0) - white (30).

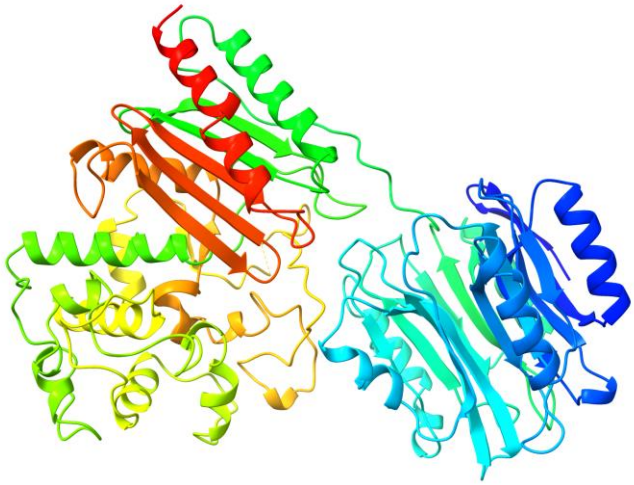


Figure 21. RoseTTAFold model of Fvan-cmp

The RoseTTAFold model of Fvan-cmp uses the rainbow spectrum colour scheme to aid in visualization of the three-dimensional structure in 2D space. The generated model was trimmed to a position-error cut-off of 3 Angstroms. This figure was generated in ChimeraX-1.4.

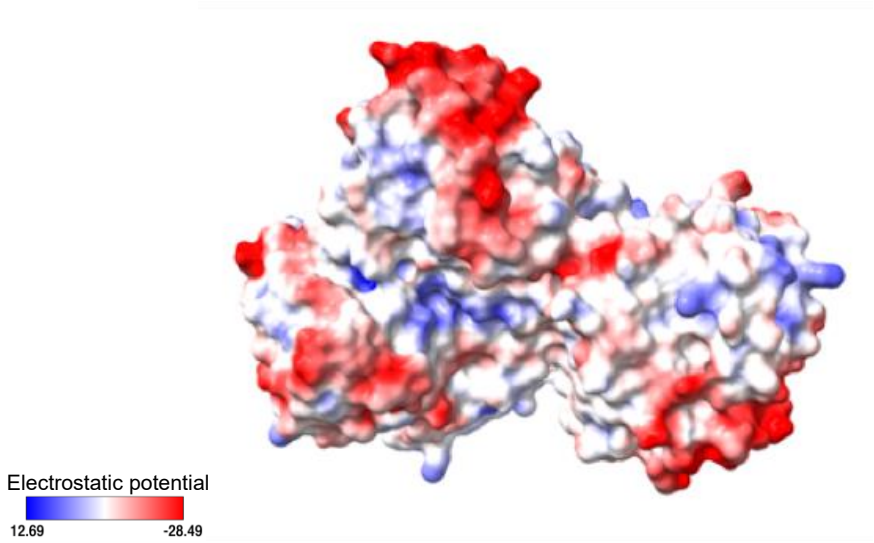


Figure 22. Fvan-cmp electrostatic potential surface map at pH 5.0

Fvan-cmp in the back orientation, the N-domain [right] and the C-domain [left]. The surface map generated in ChimeraX-1.4 follows the charge colour scale blue (positive), red (negative) and white (neutral).

Table 14. Fvan-cmp + ChitA (2) HADDOCK docking cluster analysis statistics

	HADDOCK	Cluster Size	RMSD (Å)	Van der Waals (kcal•mol^{-1})	Electrostatic (kcal•mol^{-1})	Desolvation (kcal•mol^{-1})	Restraints Violation	Buried Surface Area (Å^2)	Z-score
11	-98.0 ± 3.5	6	18.7 ± 0.1	-53.7 ± 0.8	-162.0 ± 11.3	-13.7 ± 1.6	17.8 ± 10.93	1522.5 ± 39.0	-1.9
6	-91.0 ± 3.4	8	12.2 ± 0.1	-51.3 ± 3.8	-200.3 ± 43.3	-0.6 ± 2.2	9.8 ± 13.25	1970.5 ± 32.8	-0.9
3	-89.2 ± 1.9	12	18.8 ± 0.2	-39.7 ± 2.3	-205.3 ± 11.7	-9.1 ± 2.8	6.9 ± 11.77	1260.0 ± 26.6	-0.6
18	-86.8 ± 13.4	4	19.2 ± 0.1	-56.2 ± 10.2	-111.1 ± 19.9	-8.8 ± 2.1	4.5 ± 7.37	1534.1 ± 198.7	-0.3
9	-85.5 ± 5.1	6	16.8 ± 0.5	-53.1 ± 3.0	-137.7 ± 25.8	-10.6 ± 1.1	58.8 ± 28.02	1443.8 ± 124.3	-0.1
4	-85.3 ± 10.1	12	19.0 ± 0.2	-37.2 ± 6.8	-228.3 ± 29.4	-2.5 ± 2.3	0.5 ± 0.36	1442.2 ± 116.5	-0.1
12	-82.5 ± 13.3	5	18.0 ± 0.6	-44.9 ± 6.5	-162.3 ± 65.1	-7.5 ± 3.3	23.5 ± 37.68	1342.0 ± 107.2	0.3
2	-80.3 ± 4.4	14	20.2 ± 0.1	-49.0 ± 5.0	-106.6 ± 23.6	-10.4 ± 3.6	4.2 ± 2.57	1555.4 ± 31.3	0.6
17	-76.5 ± 9.8	4	5.2 ± 0.3	-54.1 ± 7.6	-78.6 ± 11.8	-8.4 ± 3.2	17.1 ± 12.00	1762.4 ± 145.7	1.2
5	-72.3 ± 9.0	10	18.2 ± 0.1	-51.1 ± 6.0	-72.7 ± 15.8	-6.9 ± 1.2	1.8 ± 1.15	1271.7 ± 89.0	1.8

The cluster identities are in the first column and represent the top clusters for this docking run. They are ranked by their HADDOCK score which is a weighted summation of the criteria in the table.